Decomposition of representations into basis representations for the classical groups[a)]

E. D'Hoker
Center for Theoretical Physics, Laboratory for Nuclear Science and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 24 February 1983; accepted for publication 24 June 1983)

We prove decomposition formulae for an arbitrary representation in terms of basis representations for the classical compact Lie groups. Using these decomposition formulae, simple rules are obtained for the product of two arbitrary representations and for the restriction of a representation to a classical subgroup.

PACS numbers: 02.20.Qs

I. INTRODUCTION

E. Cartan[1] has classified all simple compact Lie groups into four infinite sequences of classical groups SU($n+1$), SO($2n+1$), Sp(n), and SO($2n$) of rank n and in addition five exceptional groups, E_6, E_7, E_8, F_4, and G_2. Weyl[2] has shown that every finite-dimensional irreducible representation of a classical group is in one-to-one correspondence with a complex-valued function on the group, called the group character—or simply character—of the representation. If $\lambda(g)$ is a representation, then the character χ_λ is the trace of $\lambda(g)$:

$$\chi_\lambda(g) = \operatorname{tr} \lambda(g). \tag{1.1}$$

It has the following properties:

$$\chi_\lambda(hgh^{-1}) = \chi_\lambda(g), \tag{1.2}$$

$$\chi_{\lambda\oplus\mu}(g) = \chi_\lambda(g) + \chi_\mu(g), \tag{1.3a}$$

$$\chi_{\lambda\otimes\mu}(g) = \chi_\lambda(g)\chi_\mu(g). \tag{1.3b}$$

The set of all characters forms a basis for the regular class functions on the group. Weyl[2] has also shown that the character functions of a classical group of rank n are classified by n nonnegative integers. In addition, for the orthogonal groups, there are the so-called double-valued or spinor representations, which are specified by n half-odd integers. Since the character functions are invariant under conjugation—property (1.2)—they may be completely reconstructed from their value on a Cartan subgroup. Weyl's *first* formula gives the character in terms of n angles $\phi_1, \phi_2, \ldots, \phi_n$, which parametrize the Cartan subgroup in the standard fashion.[2,3] We record Weyl's first formulae[2] here for later reference. We shall henceforth suppress the argument of the character function.

(A) SU(n):

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1},\ldots,\epsilon^{l_n}|}{|\epsilon^{l_1^0},\ldots,\epsilon^{l_n^0}|}. \tag{1.4a}$$

Here we use the definition $|\epsilon^{l_1},\ldots,\epsilon^{l_n}| = \det(E)$, with $E_{ij} = \epsilon_i^{l_j}$ where $\epsilon_i = e^{i\phi_i}$, $l_i^0 = n - i$, $l_i = f_i + l_i^0$ and the integers f_i obey $f_1 \geqslant f_2 \geqslant f_3 \geqslant \cdots \geqslant f_n$ and $(f_1+1, f_2+1, \cdots, f_{n+1}) \equiv (f_1, f_2, \cdots, f_n)$.

(B) SO($2n+1$):

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1}-\epsilon^{-l_1},\ldots,\epsilon^{l_n}-\epsilon^{-l_n}|}{|\epsilon^{l_1^0}-\epsilon^{-l_1^0},\ldots,\epsilon^{l_n^0}-\epsilon^{-l_n^0}|}. \tag{1.4b}$$

Here the l_i^0 are half-integers given by $l_i^0 = n - i + \frac{1}{2}$ and $l_i = f_i + l_i^0$ with f_i either all integers or all half-odd-integers and $f_1 \geqslant f_2 \geqslant \cdots \geqslant f_n \geqslant 0$.

(C) Sp(n):

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1}-\epsilon^{-l_1},\ldots,\epsilon^{l_n}-\epsilon^{-l_n}|}{|\epsilon^{l_1^0}-\epsilon^{-l_1^0},\ldots,\epsilon^{l_n^0}-\epsilon^{-l_n^0}|}. \tag{1.4c}$$

Here the l_i^0 are integers and are given by $l_i^0 = n - i + 1$ and $l_i = f_i + l_i$ with f_i integer and $f_1 \geqslant f_2 \geqslant \cdots \geqslant f_n \geqslant 0$.

(D) SO($2n$):

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1}+\epsilon^{-l_1},\ldots,\epsilon^{l_n}+\epsilon^{-l_n}|}{|\epsilon^{l_1^0}+\epsilon^{-l_1^0},\ldots,\epsilon^{l_n^0}+\epsilon^{-l_n^0}|}. \tag{1.4d}$$

Here the l_i^0 are integers defined by $l_i^0 = n - i$ and $l_i = f_i + l_i^0$ with f_i either all integers or all half-integers and $f_1 \geqslant f_2 \geqslant \cdots \geqslant |f_n| > 0$. When $f_n = 0$, the right-hand side of (1.4d) is divided by a factor of 2.

The ordered set $(f_1, f_2, \ldots, f_n)$ is called the signature[1] and is also equal to the highest weight vector. To every irreducible representation corresponds one and only one signature $(f_1, f_2, \ldots, f_n)$ such that $f_1 \geqslant f_2 \geqslant \ldots \geqslant f_n$, and this signature is called dominant.

Using algebraic manipulations, one can rewrite expressions (1.4a–d) in terms of a set of characters of generating representations instead of in terms of the exponential functions ϵ_i. Weyl's *second* formula[2,3] gives the characters in terms of the so-called symmetric representations. For SU(n), Weyl's second formula reads

$$\chi_{(f_1,f_2,\ldots,f_n)} = \det\Sigma, \tag{1.5a}$$

$$\Sigma_{ij} = \chi_{d^{i-j+f_i}}. \tag{1.5b}$$

d^k has signature $(k, 0, 0, \ldots, 0)$ when $k \geqslant 0$, and is defined to vanish when $k < 0$.

Similar formulae exist for the other three series of classical groups.[2] Weyl's second formula is quite useful: It provides a practical algorithm for the decomposition of the ten-

[a)] This work is supported in part through funds provided by the U. S. Department of Energy (DOE) under contract DE-AC02-76ERO3069.

 0022-2488/84/010001-12$02.50

sor product of two representations into irreducible representations.[3] Indeed, from (1.5) it is clear that the tensor product of two arbitrary representations λ and μ can be expressed in terms of products of λ with *symmetric* representations. The latter products may be evaluated using a set of rules deduced from Weyl's first formula.[2,3]

In the present paper, we shall show that, for the classical groups the character of an arbitrary representation may also be expressed as a determinant only involving the so-called *basis* representations in an elementary way. For a group of rank n, there are precisely n basis representations whose signatures are listed below.[3] (Henceforth, we identify a representation with its dominant signature.)

(A) $SU(n+1)$: $d_p = (\underbrace{1,1,...,1}_{p},0,...,0)$ and $p = 1,...,n$; (1.6a)

(B) $SO(2n+1)$: d_p, $p = 1,...,n-1$ and the spinor representation $s = (\frac{1}{2},...,\frac{1}{2})$; (1.6b)

(C) $Sp(n)$: d_p, $p = 1,...,n$; (1.6c)

(D) $SO(2n)$: d_p, $p = 1,n-2$, and the two spinor representations $s^+ = (\frac{1}{2},...,\frac{1}{2},\frac{1}{2})$ (1.6d)
$s^- = (\frac{1}{2},...,\frac{1}{2},-\frac{1}{2})$.

Our formulae give all characters in terms of only a finite number of generators[4]: $\{\chi_{d_p}\}_{p=1,n}$ for $SU(n+1)$ and $Sp(n)$, $\{\chi_{d_p},\chi_s\}_{p=1,n-1}$ for $SO(2n+1)$ and $\{\chi_{d_p},\chi_{s^+},\chi_{s^-}\}_{p=1,n-2}$ for $SO(2n)$. The proof of these relations, henceforth called *decomposition formulae*, is the main objective of the present paper, and is given in Secs. II, III, IV, and V, respectively for $SU(n)$, $SO(2n+1)$, $Sp(n)$, and $SO(2n)$. For each of these groups, we shall first determine rules for the product of a basis representation with an arbitrary representation and then prove the decomposition formulae, essentially by explicit calculation of the determinant.

The case of $SU(n)$ is simplest, and will be developed in much detail; the case of $SO(2n+1)$ requires several important modifications, which we shall fully describe. For $Sp(n)$, only the final results will be given, and, for $SO(2n)$, special attention will be devoted to subtleties like double characters. Finally, in the last section we shall discuss three applications. First, we show that our decomposition formulae provide rules for the tensor multiplication of two arbitrary representations of any of the classical groups, just as Weyl's second formula did.[2] These rules are only slightly more complicated for the groups $Sp(n)$ or $SO(n)$ than for the group $SU(n)$, and may present an interesting alternative to the rather involved rules discussed in standard references.[5] Second, we prove a relation between the dimensions of the representations of $Sp(n)$ and these of the spinor representations of $SO(2n+1)$. Finally, we show that our decomposition formulae yield a simple algorithm for the calculation of the restriction of a representation to a classical subgroup of the original group. Thus branching rules for nonmaximal subgroups can be obtained. Let us also remark that the simple rules for products and branching of representations could be easily implemented in a computer program.

The extension of our formulae to the case of exceptional groups is presently under investigation.

II. THE SPECIAL UNITARY GROUPS SU(n)

Multiplication of a basis representation with an arbitrary representation

The weight diagram[3] for the basis representations is deduced from Weyl's first formula (1.4a)

$$\chi_{d_p} = \sum_{i_1<i_2<\cdots<i_p} \epsilon_{i_1}\cdots\epsilon_{i_p}. \tag{2.1}$$

The character of the tensor product of d_p with a representation $\lambda = (f_1, f_2, ..., f_n)$ is found using (1.3):

$$\chi_{d_p\otimes\lambda} = \sum_{i_1<i_2<\cdots<i_p} \epsilon_{i_1}\epsilon_{i_2}\cdots\epsilon_{i_p} \frac{|\epsilon^{l_1},...,\epsilon^{l_n}|}{|\epsilon^{l_1^0},...,\epsilon^{l_n^0}|}. \tag{2.2}$$

The integers l_i are defined in terms of the f_i by $l_i = f_i + l_i^0$.

Note that χ_λ and χ_{d_p} are invariant under the action of the Weyl group,[2,3] which permutes the angles ϕ_i. Using this invariance for $\chi_{d_p\otimes\lambda}$, we find

$$\chi_{d_p\otimes\lambda} = \sum_{i_1<i_2<\cdots<i_p} \frac{|\epsilon^{l_1},...,\epsilon^{l_{i_1}},...,\epsilon^{l_{i_p}},...,\epsilon^{l_n}|}{|\epsilon^{l_1^0},...,\epsilon^{l_n^0}|} \tag{2.3}$$

so that

$$\chi_{d_p\otimes\lambda} = \sum_{i_1<i_2<\cdots<i_p} \chi_{(f_1,...,f_{i_1}+1,...,f_{i_p}+1,...,f_n)}. \tag{2.4a}$$

In (2.4a), a character corresponding to a signature which is not dominant must be omitted. Expression (2.4a), together with the one-to-one correspondence between dominant characters and irreducible representations, implies the following formula for the representations:

$$d_p\otimes\lambda = \sum_{i_1<i_2<\cdots<i_p} (f_1,...,f_{i_1}+1,...,f_{i_p}+1,...,f_n). \tag{2.4b}$$

Here again, nondominant signatures are deleted.

The decomposition formula for the symmetric representations

Before attacking the full problem, we shall first prove a decomposition formula for the symmetric representation d^k of $SU(n)$ [defined in (1.5)].

Theorem 1. Let M^k be the following determinant[6]

$$M^k = \begin{vmatrix} d_1 & 1 & 0 & & 0 \\ d_2 & d_1 & 1 & & 0 \\ d_3 & d_2 & d_1 & & 0 \\ \vdots & \vdots & & \ddots & \\ d_k & d_{k-1} & & & d_1 \end{vmatrix}\otimes. \tag{2.5}$$

Then we have $M^k = d^k$.

In formula (2.5) it is understood that $d_k = 0$ if $k > n$ or $k < 0$.

Proof: Upon multiplication by the determinant

$$1 = \begin{vmatrix} 1 & 0 & 0 & & 0 \\ -d^1 & 1 & 0 & & 0 \\ d^2 & -d^1 & 1 & & 0 \\ & & & \ddots & \vdots \\ (-1)^{k-1}d^{k-1} & \cdots & & & 1 \end{vmatrix}\otimes, \tag{2.6}$$

making use of the well-known[3] duality relation

$$\sum_{p=0}^{n-1}(-1)^p d_p \otimes d^{k-p} = \begin{cases} 1 & \text{if } k=0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.7}$$

it is clear that $M^k = d^k$, as announced.

We now give also a different proof, the method of which will generalize to the case of an arbitrary representation of SU(n) as well as to the other classical groups. The expansion of the determinant in (2.5) along the first column yields a sum of products of a basis representation d_j with a minor Δ_j. The crucial remark is that this minor Δ_j is of the same form as the original determinant: $\Delta_j = M^{k-j}$. Thus we have

$$M^k = \sum_{\alpha=1}^{k} d_\alpha \otimes M^{k-\alpha}(-1)^{\alpha-1}. \tag{2.8}$$

We can prove (2.5) by induction. Suppose that $M^k = d^k$ for all $k \leqslant p-1$ and clearly $M^1 = d^1$; then we wish to prove that $M^p = d^p$. The induction hypothesis together with (2.8) yields

$$M^k = \sum_{\alpha=1}^{k} d_\alpha \otimes d^{k-\alpha}(-1)^{\alpha-1}. \tag{2.9}$$

Using (2.4b), we see that

$$d_\alpha \otimes d^{k-\alpha} = B_\alpha + B_{\alpha+1}, \tag{2.10}$$

where

$$B_\alpha = (\underbrace{k-\alpha+1,1,\cdots,1}_{\alpha},0,\ldots,0) \tag{2.11}$$

and $B_{k+1} = 0$ since it corresponds to a nondominant signature. Hence we have

$$M^k = \sum_{\alpha}^{k}(-1)^{\alpha-1}(B_\alpha + B_{\alpha+1}) \tag{2.12}$$

so that $M^k = B_1 = (k, 0, \ldots, 0) = d^k$ as announced. Upon replacing M^j by d^j in (2.8) we get precisely (2.7). This finishes the proof of Theorem 1.

Combination of (2.5) and (1.5) shows that every representation can be written as a function of basis representations d_p alone. We shall now prove a much more convenient formula for the decomposition in terms of basis representations.

The general decomposition formula

Let λ be a representation with dominant signature $(f_1, f_2, \ldots, f_n)$; the nonnegative projection numbers r_i are obtained from the projection of the highest weight vector onto the roots[3]:

$$r_i = f_i - f_{i+1}, \quad i = 1,\ldots,n-1. \tag{2.13}$$

We shall now prove the following general decomposition formula for the unitary groups[6,7]:

Theorem 2: Let

$$\mu = \begin{vmatrix} d_1 & 1 & \cdots & 0 & \cdots & & & & \\ d_2 & d_1 & & \vdots & & & & & \\ \vdots & & \ddots & & & & & & \\ d_{r_1} & \cdots & & d_1 & 1 & 0 & & & \\ d_{r_1+2} & \cdots & & d_3 & d_2 & d_1 & \cdots & & \\ & & & & & & d_2 & & \\ & & & & & & & \ddots & \\ & & & & & & & & d_{n-1} \\ & & & & & & & \ddots & d_{n-1}\ d_{n-2} \\ & & & & & & 0 & \cdots & 1\ \ d_{n-1} \end{vmatrix} \begin{matrix} \otimes \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} r_1 \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} r_2 \\ \vdots \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} r_{n-1} \end{matrix} \tag{2.14}$$

or

$$\mu = \otimes det(\mathscr{D}) \quad \text{with } \mathscr{D}_{ij} = d_{i-j+k} \tag{2.15}$$

and let k be defined by $r_1 + r_2 + \cdots + r_{k-1} < i \leqslant r_1 + r_2 + \cdots + r_k$. Then we have $\mu = \lambda$.

Observe that in formula (2.14) we have r_i times the representation d_i on the diagonal. The off-diagonal elements of the determinant are obtained by incrementing (resp. decrementing) the index i by one unit when moving to the left (resp. to the right).

Proof: In analogy with Theorem 1, the expansion of the determinant (2.14) along the first column yields a sum of products of a basis representation and a minor, which is of the same form as the original determinant μ. We proceed with a proof by induction on the first coordinate of the signature f_1. Suppose $\mu = \lambda$ for all $f_1 \leqslant p-1$; then we want to prove that $\mu = \lambda$ for *all* representations such that $f_1 = p$. Clearly, we have $\mu = \lambda$ for $f_1 = 1$. As a consequence of the induction hypothesis, we see that every minor corresponds to one and only one irreducible representation. Indeed, the minor associated with d_1 has signature $(f_1 - 1, f_2, \ldots, f_n)$, the minor associated with d_2 has signature $(f_1 - 2, f_2, \ldots, f_n)$; this pattern continues until one encounters the element d_{r_1} in the first column which has minor $(f_2, f_2, f_3, \ldots, f_n)$. If $r_2 \neq 0$, then at least one d_2 is present on the diagonal, and the next element in the first column is d_{r_1+2} with minor $(f_2 - 1, f_2 - 1, f_3, \ldots, f_n)$. Upon increasing the index of the element in the first column by 1, the second entry in the signature of the minor decreases by 1. It is remarkable that each minor in the expansion of determinant (2.14) is again an irreducible representation with a signature such that its first entry is always smaller than f_1. Thus we must prove that the expansion of the determinant, for a representation with dominant signature $(f_1, f_2, \ldots, f_n)$, with all minors replaced by their actual value precisely yields $\mu = \lambda$.

Using the signature notation, the above described expansion becomes

$$\mu = \oplus \sum_{\alpha=1}^{r_1} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}$$
$$\oplus \sum_{\alpha=1}^{r_2} d_{r_1+1+\alpha} \otimes (f_2-1, f_2-\alpha, f_3, \ldots, f_n)$$
$$\times(-1)^{\alpha+r_1-1} \oplus \cdots$$
$$\oplus \sum_{\alpha=1}^{r_{n-1}} d_{r_1+\cdots+r_{n-2}+\alpha+n-2}$$
$$\otimes (f_2-1, f_3-1, \ldots, f_{n-1}-1, f_n-\alpha)$$
$$\times(-1)^{r_1+\cdots+r_{n-2}+\alpha-2}. \tag{2.16}$$

All signatures appearing in (2.16) are dominant by construction.

Formula (2.16) may, however, be simplified through the use of nondominant signatures—henceforth called signatures as opposed to dominant signatures. We shall generalize (1.4a) to signatures $(f_1, f_2, \ldots, f_n)$, which need not be dominant, by defining their character as

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1}, \epsilon^{l_2}, \ldots, \epsilon^{l_n}|}{|\epsilon^{l_1^0}, \epsilon^{l_2^0}, \ldots, \epsilon^{l_n^0}|} \tag{2.17}$$

even when f is not dominant. Of course, every signature is either related to a dominant signature by permutation of columns in (2.16) or must vanish.[8] Thus we have, e.g., $\chi_{(2,4,1)} = -\chi_{(3,3,1)}$ but $\chi_{(2,3,1)} = 0$. Using the definition of (non dominant) signatures, (2.16) may be rewritten

$$\mu = \oplus \sum_{\alpha=1}^{r_1} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}$$
$$\oplus \sum_{\alpha=1}^{r_2} d_{r_1+1+\alpha} \otimes (f_2-\alpha-1, f_2, \ldots, f_n)(-1)^{r_2+\alpha}$$
$$\oplus \sum_{\alpha=1}^{r_3} d_{r_1+r_2+2+\alpha} \otimes (f_3-\alpha-2, f_2, f_3, \ldots, f_n)$$
$$\times(-1)^{r_1+r_2+\alpha+1}$$
$$\oplus \cdots. \tag{2.18}$$

A shift in the summation variable produces

$$\mu = \oplus \sum_{\alpha=1}^{r_1} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}$$
$$\oplus \sum_{\alpha=r_1+2}^{r_1+r_2+1} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}$$
$$\oplus \sum_{\alpha=r_1+r_2+3}^{r_1+r_2+r_3+2} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1} \oplus \cdots$$
$$\oplus \sum_{\alpha=r_1+r_2+\cdots+r_{n-2}+n-1}^{r_1+r_2+\cdots+r_{n-1}+n-2} d_\alpha \otimes (f_1-\alpha, f_2, f_3, \ldots, f_n)$$
$$\times(-1)^{\alpha-1}. \tag{2.19}$$

The signature vanishes at values of α which are missing from the summation, through the use of (2.17). Hence we have, e. g.,

$$(f_1-r_1-1, f_2, f_3, \ldots, f_n) = (f_2-1, f_2, f_3, \ldots, f_n) = 0. \tag{2.20}$$

Using the above property, we obtain our final formula:

$$\mu = \otimes \sum_{\alpha=1}^{f_1+n-2} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}. \tag{2.21}$$

We shall now prove that $\mu = \lambda$, by explicit calculation of μ.

First we need the generalization (2.4) to nondominant signatures:

$$d_p \otimes (f_1, f_2, \ldots, f_n)$$
$$= \oplus \sum_{i_1<i_2<\cdots<i_p} (f_1, \ldots, f_{i_1}+1, \ldots, f_{i_p}+1, \ldots, f_n). \tag{2.22}$$

In (2.21), all terms may be kept since the nondominant signatures automatically vanish when $(f_1, f_2, \ldots, f_n)$ is dominant. By permutation of columns in (2.2) and (2.17), it is also clear that formula (2.22) holds even when f is not dominant.

The explicit calculation of all terms in the tensor products in (2.21) is unnecessary, and we shall introduce the following convenient shorthand

$$(f_1, \{f_2, \ldots, f_n\}_+^p)$$
$$= \oplus \sum_{1<i_1<i_2\cdots<i_p} (f_1, \ldots, f_{i_1}+1, \ldots, f_{i_p}+1, \ldots, f_n) \tag{2.23}$$

with

$$(f_1, \{f_2, \ldots, f_n\}_+^0) = (f_1, f_2, \ldots, f_n) \tag{2.24a}$$
$$(f_1, \{f_2, \ldots, f_n\}_+^p) = 0 \quad \text{when} \quad p<0 \text{ or } p>n-1. \tag{2.24b}$$

Then the tensor products in (2.21) can be computed using (2.23) and (2.24):

$$d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n) = (f_1-\alpha+1, \{f_2, \ldots, f_n\}_+^{\alpha-1}) + (f_1-\alpha, \{f_2, \ldots, f_n\}_+^{\alpha}) \tag{2.25}$$

so that

$$\mu = \oplus \sum_{\alpha=1}^{f_1+n-2} (-1)^{\alpha-1}(f_1-\alpha+1, \{f_2, \ldots, f_n\}_+^{\alpha-1})$$
$$\oplus \sum_{\alpha=1}^{f_1+n-2} (-1)^{\alpha-1}(f_1-\alpha, \{f_2, \ldots, f_n\}_+^{\alpha}). \tag{2.26}$$

A shift in the summation variable of the second sum yields

$$\mu = (f_1, f_2, f_3, \ldots, f_n)$$
$$\oplus (-1)^{f_1+n-1}(-n+2, \{f_2, \ldots, f_n\}_+^{f_1+n-2}). \tag{2.27}$$

Since $f_1 \geqslant 1$, the second bracket vanishes with the use of (2.24b) and since $(f_1, f_2, \ldots, f_n)$ is dominant, we have proven that $\mu = \lambda$.

Example: The representation λ with signature (4,2,1,0,0) of SU(5) is decomposed as follows:

$$(4,2,1,0,0) = \begin{vmatrix} d_1 & 1 & 0 & 0 \\ d_2 & d_1 & 1 & 0 \\ d_4 & d_3 & d_2 & d_1 \\ 0 & 1 & d_4 & d_3 \end{vmatrix}.$$

One can, e. g., check the dimensions by taking the character of both sides of (2.27) and computing the determinant at the identity element. With the use of tables of dimensions,[9] we find

$$700 = \begin{vmatrix} 5 & 1 & 0 & 0 \\ 10 & 5 & 1 & 0 \\ 5 & 10 & 10 & 5 \\ 0 & 1 & 5 & 10 \end{vmatrix}$$

III. THE ORTHOGONAL GROUPS SO(2n + 1)

The decomposition formulae for the spinor representation are different and will be treated separately from the formulae for single valued representations. Furthermore, it will appear natural to express the decomposition formulae in terms of the basis representations of (1.6b) *plus* the representation $d_n = (1,1,...,1)$. Later we shall prove that the latter representation is simply expressed in terms of the former ones.

Multiplication of a basis representation with an arbitrary representation

The weight diagram of the representations d_p is deduced from Weyl's first formula (1.4b):

$$\chi_{d_p} = \frac{|\epsilon^{l_1} - \epsilon^{-l_1},...,\epsilon^{l_n} - \epsilon^{-l_n}|}{|\epsilon^{l_1^0} - \epsilon^{-l_1^0},...,\epsilon^{l_n^0} - \epsilon^{-l_n^0}|}, \tag{3.1}$$

where $l_i^0 = n - i + \frac{1}{2}$, $l_i = l_i^0$ if $i > p$ and $l_i = l_i^0 + 1$ if $i \leqslant p$. We shall obtain a more convenient expression for this weight diagram by introducing the function

$$\mathscr{A}_p = |\epsilon^n + \epsilon^{-n},...,\epsilon^{n-p+1} + \epsilon^{-n+p-1},\epsilon^{n-p-1} + \epsilon^{-n+p+1},...,\epsilon + \epsilon^{-1},1|. \tag{3.2}$$

After division of numerator and denominator in (3.1) by the common factor $\Pi_{i=1}^n(\epsilon_i^{1/2} - \epsilon_i^{-1/2})$, χ_{d_p} can be rewritten as follows.

$$\chi_{d_p} = (\mathscr{A}_p + \mathscr{A}_{p-1})/\mathscr{A}_0. \tag{3.3}$$

The function $\mathscr{A}_p$ can be easily evaluated using the binomial coefficients C_n^k:

$$\mathscr{A}_p = \sum_{\alpha=0}^{[p/2]} C_{n-p+2\alpha}^{\alpha} R(p-2\alpha)\mathscr{A}_0, \tag{3.4}$$

with

$$R(q) = \sum_{i_1<i_2<\cdots<i_q} \sum_{\sigma_{i_j}=\pm 1} \epsilon_{i_1}^{\sigma_{i_1}}\epsilon_{i_2}^{\sigma_{i_2}}\cdots\epsilon_{i_q}^{\sigma_{i_q}}. \tag{3.5}$$

The function $R(q)$ may be thought of as the character of the representation d_q of a unitary group with ϵ_i replaced with $\epsilon_i + \epsilon_i^{-1}$. Combination of formulae (3.3)–(3.5) gives us the weight diagram of d_p:

$$\chi_{d_p} = \sum_{\alpha=0}^{[p/2]} C_{n-p+2\alpha}^{\alpha} R(p-2\alpha) + \sum_{\alpha=0}^{[(p-1)/2]} C_{n-p+1+2\alpha}^{\alpha} R(p-1-2\alpha). \tag{3.6}$$

For the spinor representation, the weight diagram is computed directly from (1.4b)

$$\chi_s = \sum_{\sigma_i=\pm 1} \epsilon_1^{\sigma_1/2}\epsilon_2^{\sigma_2/2}\cdots\epsilon_n^{\sigma_n/2}. \tag{3.7}$$

The functions $R(q)$, χ_{d_p} and χ_s are invariant under the action of the Weyl group which permutes the ϕ_i and changes their sign. Again we generalize the dominant signatures in (1.4b) to (nondominant) or generalized signatures, defined through the same formula (1.4b) but where the signature $(f_1, ..., f_n)$ need not be dominant. Every generalized signature is again either related to a dominant signature or must vanish. An important special case is

$$\chi_{(f_1,f_2,...,f_{n-1},-1)} = -\chi_{(f_1,f_2,...,f_{n-1},0)} \tag{3.8}$$

for $(f_1, f_2, ..., f_{n-1}, 0)$ dominant.

With the help of the weight diagram computed previously, we can evaluate the tensor product of d_p and s with a representation $\lambda = (f_1, f_2, ..., f_n)$. First we need the product of $R(q)$ with χ_λ, obtained using the invariance under the action of the Weyl group:

$$R(q)\chi_\lambda = \sum_{i_1<i_2<\cdots<i_q} \sum_{\sigma_{i_j}=\pm 1} \chi_{(f_1,...,f_{i_1}+\sigma_{i_1},...,f_{i_q}+\sigma_{i_q},...,f_n)}. \tag{3.9}$$

Formula (3.9) also holds for $(f_1, f_2, ..., f_n)$ *not* dominant. The product of d_p and λ is gotten by combining (3.6) and (3.9). The tensor product of the spinor representation with λ is deduced from (3.7) in an analogous fashion.

$$\chi_s\chi_\lambda = \sum_{\sigma_i=\pm 1} \chi_{(f_1+\sigma_1/2, f_2+\sigma_2/2,..., f_n+\sigma_n/2)}. \tag{3.10}$$

As an example of these multiplication rules, we compute the following product for SO(9):

$$\begin{aligned}(1,1,0,0)\otimes(3,1,0,0) &= (4,\{1,0,0\}^1) + (3,\{1,0,0\}^2) + (2,\{1,0,0\}^1) + \cdots \\ &= (4,2,0,0) + (4,1,1,0) + (4,0,0,0) + (3,2,1,0) \\ &\quad + (3,1,1,1) + (2,2,0,0) + 2(3,1,0,0) \\ &\quad + (2,1,1,0) + (2,0,0,0).\end{aligned}$$

The decomposition formula for nonspinor representations

It is not hard to generalize formula (2.14) to the case of the group SO(2n + 1). The explicit form of the weight diagram in (3.6) suggests that we should take the linear combinations

$$D_k = d_k \ominus d_{k-1} \oplus d_{k-2} \oplus \cdots \oplus (-1)^k d_0 \tag{3.11}$$

as elementary building blocks in a determinantal expression like (2.14). Examination of a few simple special cases shows that this is basically correct, provided one modifies (2.14) in a way which we shall now specify. Let λ be a representation with dominant signature $(f_1, ..., f_n)$ and define the integers

$$r_i = f_i - f_{i+1}, \quad i = 1,2,...,n-1, \quad r_n = f_n, \tag{3.12}$$

as well as the sequence of direct sums and differences of basis representations

$$\begin{aligned} &0 \leqslant k \leqslant n, && D_k = d_k \ominus d_{k-1} \oplus d_{k-2} + \cdots \oplus (-1)^k d_0, \\ &n \leqslant k \leqslant 2n, && D_k = D_{2n-k}, \\ &k > 2n \text{ or } k < 0, && D_k = 0. \end{aligned} \tag{3.13}$$

Theorem 3: Let

$$\mu = \begin{vmatrix} D_1 & D_0 & & \cdots & & & 0 & \Big\} \otimes \\ D_2 & D_1 & & & & & 0 & \Big\} r_1 \\ \vdots & & \ddots & & & & \vdots & \\ D_{r_1} & \cdots & D_1 & & & & 0 & \\ D_{r_1+2} & \cdots & D_3 & D_2 & & & & \Big\} r_2 \\ & & & & \ddots & & & \\ & & & & D_2 & & & \\ & & & & & \ddots & & \\ & & & & & D_n \oplus & D_{n-r_n} & \\ & & & & & & \ddots & \\ & & & \cdots & & D_n \oplus D_{n-3} & D_{n-1} \oplus D_{n-2} & \\ & & & \cdots & & D_{n+1} \oplus D_{n-2} & D_n \oplus D_{n-1} & \Big\} r_n \end{vmatrix} \tag{3.14}$$

or in components

$$\mu = \otimes \det(\mathscr{D}),$$
$$\mathscr{D}_{ij} = D_{i-j+k} \oplus D_{i+j+k-1-2f_i},$$
$$r_1 + r_2 + \cdots + r_{k-1} < i \leqslant r_1 + r_2 + \cdots + r_k. \tag{3.15}$$

Then $\mu = \lambda$.

Proof: The proof again proceeds by induction of f_1. Suppose $\mu = \lambda$ for *all* λ such that $f_1 \leqslant p-1$; then we wish to prove that $\mu = \lambda$ for *all* λ such that $f_1 = p$. The expansion of the determinant along the first column yields tensor products of representations $D_{i-1+k} \oplus D_{i+k-2f_i}$ with minors. These minors are of the same form as the original determinant, and by the recurrence hypothesis equal to irreducible representation of which the first entry in the signature never exceeds $f_1 - 1$. The resulting formula for μ is the same as (2.16) but with d_α replaced with $D_\alpha \oplus D_{\alpha+1-2f_i}$. The definition of (generalized) signatures for the SO($2n+1$) again leads to a drastic simplification, analogous to the one that leads to (2.21) and we finally get

$$\mu = \sum_{\alpha=1}^{f_1+n-1} (D_\alpha \oplus D_{\alpha+1-2f_i}) \otimes (f_1 - \alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}. \tag{3.16}$$

We prove that $\mu = \lambda$ by explicit calculation of μ. As for the unitaries, the tensor products in (3.16) need not be worked out explicitly and we introduce the following shorthand:

$$(f_1, \{f_2, \ldots, f_n\}_\pm^p) = \oplus \sum_{1 < i_1 < i_2 \cdots < i_p} \sum_{\sigma_{i_j} = \pm 1} (f_1, \ldots, f_{i_1} + \sigma_{i_1}, \ldots, f_{i_p} + \sigma_{i_p}, \ldots, f_n), \tag{3.17a}$$

$$(f_1, \{f_2, \ldots, f_n\}_\pm^0) = (f_1, f_2, \ldots, f_n), \tag{3.17b}$$

$$(f_1, \{f_2, \ldots, f_n\}_\pm^p) = 0 \quad \text{if} \quad p < 0 \text{ or } p > n-1. \tag{3.17c}$$

We first evaluate the product of the functions $R(p)$ with an arbitrary character using (3.9):

$$R(p)\chi_{(f_1, f_2, \ldots, f_n)} = \chi_{(f_1+1, \{f_2, \ldots, f_n\}_\pm^{p-1})} + \chi_{(f_1, \{f_2, \ldots, f_n\}_\pm^p)} + \chi_{(f_1-1, \{f_2, \ldots, f_n\}_\pm^{p-1})}. \tag{3.18}$$

The weight diagram of D_p is computed using (3.5) and is given by

$$\chi_{D_p} = \sum_{\beta=0}^{[p/2]} C_{n-p+2\beta}^{\beta} R(p-2\beta) \tag{3.19}$$

for all $p \geqslant 0$. Next, we compute the tensor products occurring in (3.16) and make use of the shorthand introduced in (3.17)–(3.18)

$$D_\alpha \otimes (f_1 - \alpha, f_2, \ldots, f_n) = \sum_{\beta=0}^{[\alpha/2]} C_{n-\alpha+2\beta}^{\beta} (B_{\alpha-1}^{\beta} \oplus B_\alpha^\beta \oplus B_{\alpha+1}^{\beta+1}), \tag{3.20a}$$

$$D_{\alpha+1-2f_i} \otimes (f_1 - \alpha, f_2, \ldots, f_n) = \sum_{\beta=0}^{[(\alpha+1-2f_1)/2]} C_{n-\alpha-1+2f_1+2\beta}^{\beta} \times (\mathscr{K}_{\alpha-1}^{\beta} \oplus \mathscr{K}_\alpha^\beta \oplus \mathscr{K}_{\alpha+1}^{\beta+1}) \tag{3.20b}$$

with

$$B_\alpha^\beta = (f_1 - \alpha, \{f_2, \ldots, f_n\}_\pm^{\alpha-2\beta}), \tag{3.21a}$$
$$\mathscr{K}_\alpha^\beta = (f_1 - \alpha, \{f_2, \ldots, f_n\}_\pm^{\alpha-2\beta-2f_1+1}). \tag{3.21b}$$

We compute μ in two steps:

$$\mu_1 = \sum_{\alpha=1}^{f_1+n-1} D_\alpha \otimes (f_1 - \alpha, f_2, \ldots, f_n)(-1)^{\alpha-1} = \sum_{\alpha=1}^{f_1+n-1} \sum_{\beta=0}^{[\alpha/2]} C_{n-\alpha+2\beta}^{\beta} \times (B_{\alpha-1}^{\beta} \oplus B_\alpha^\beta \oplus B_{\alpha+1}^{\beta+1})(-1)^{\alpha-1}. \tag{3.22}$$

Upon performing the appropriate shifts in the summation variables, we find

$$\mu_1 = \sum_{\alpha=0}^{f_1+n-2} \sum_{\beta=0}^{[(\alpha+1)/2]} C_{n-\alpha-1+2\beta}^{\beta} B_\alpha^\beta (-1)^\alpha \ominus \sum_{\alpha=1}^{f_1+n-1} \sum_{\beta=0}^{[\alpha/2]} C_{n-\alpha+2\beta}^{\beta} B_\alpha^\beta (-1)^\alpha \oplus \sum_{\alpha=2}^{f_1+n} \sum_{\beta=1}^{[(\alpha+1)/2]} C_{n-\alpha-1+2\beta}^{\beta-1} B_\alpha^\beta (-1)^\alpha. \tag{3.23}$$

For α odd we have

$$B_\alpha^{[(\alpha+1)/2]} = (f_1 - \alpha, \{f_2, \ldots, f_n\}_\pm^{-1}) = 0 \tag{3.24}$$

so that the summation over β in the second term may be

extended from $[\alpha/2]$ to $[(\alpha+1)/2]$. Then we rearrange expression (3.23) as follows:

$$\begin{aligned}\mu_1 = {} & B_0^0 \ominus B_1^0 \ominus n B_1^1 \oplus B_1^0 \oplus (n+1) B_1^1 \\ & \oplus \sum_{\beta=0}^{[(f_1+n)/2]} C^{\beta}_{1-f_1+2\beta} B^{\beta}_{f_1+n-1} (-1)^{f_1+n} \\ & \oplus \sum_{\beta=1}^{[(f_1+n)/2]} C^{\beta-1}_{-f_1+2\beta} B^{\beta}_{f_1+n-1} (-1)^{f_1+n-1} \\ & \oplus \sum_{\beta=1}^{[(f_1+n+1)/2]} C^{\beta-1}_{-f_1-1+2\beta} B^{\beta}_{f_1+n} (-1)^{f_1+n} \\ & \oplus \sum_{\alpha=2}^{f_1+n-2} \sum_{\beta=0}^{[(\alpha+1)/2]} (C^{\beta}_{n-\alpha-1+2\beta} - C^{\beta}_{n-\alpha+2\beta} \\ & + C^{\beta-1}_{n-\alpha+2\beta}) B^{\beta}_{\alpha} (-1)^{\alpha}. \end{aligned} \tag{3.25}$$

The double sum in (3.25) vanishes due to Pascal's equality on binomial coefficients. Using (3.24) again for the term B_1^1 and Pascal's equality,

$$\begin{aligned}\mu_1 = {} & B_0^0 \oplus \sum_{\beta=0}^{[(f_1+n)/2]} C^{\beta}_{-f_1+2\beta} B^{\beta}_{f_1+n-1} (-1)^{f_1+n} \\ & \oplus \sum_{\beta=1}^{[(f_1+n+1)/2]} C^{\beta-1}_{-f_1-1+2\beta} B^{\beta}_{f_1+n} (-1)^{f_1+n}. \end{aligned} \tag{3.26}$$

The same sequence of manipulations may be applied to the expression for μ_2,

$$\mu_2 = \sum_{\alpha=1}^{f_1+n-1} D_{\alpha+1-2f_1} \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}, \tag{3.27a}$$

and it yields

$$\begin{aligned}\mu_2 = {} & \sum_{\beta=f_1}^{[(n+f_1+1)/2]} (-1)^{f_1+n} C^{\beta-f_1}_{-f_1+2\beta-1} \mathscr{K}^{\beta-f_1}_{f_1+n-1} \\ & \oplus \sum_{\beta=f_1}^{[(n+f_1)/2]} (-1)^{f_1+n} C^{\beta-f_1}_{-f_1+2\beta} \mathscr{K}^{\beta-f_1+1}_{f_1+n} \end{aligned} \tag{3.27b}$$

Using the properties of the binomial coefficients, we see that the sums in (3.26) actually only start at $\beta = f_1$ instead of at $\beta = 0$ or $\beta = 1$. Taking this remark into account, we obtain the following result for μ:

$$\begin{aligned}\mu = {} & B_0^0 \oplus \sum_{\beta=f_1}^{[(f_1+n)/2]} C^{\beta}_{-f_1+2\beta} (B^{\beta}_{f_1+n-1} \\ & \oplus \mathscr{K}^{\beta-f_1+1}_{f_1+n})(-1)^{f_1+n} \\ & \oplus \sum_{\beta=f_1}^{[(f_1+n+1)/2]} C^{\beta-1}_{-f_1+2\beta-1} (B^{\beta}_{f_1+n} \\ & \oplus \mathscr{K}^{\beta-f_1}_{f_1+n-1})(-1)^{f_1+n}. \end{aligned} \tag{3.28}$$

From the definition of B and $\mathscr{K}$ in (3.20) and making use of the properties of generalized signatures we see that

$$\begin{aligned}\mathscr{K}^{\beta-f_1+1}_{f_1+n} & = (-n, \{f_2, \ldots, f_n\}_{\pm}^{f_1+n-2\beta-1}) \\ & = (-1)^{n-1}(\{f_2-1, f_3-1, \ldots, f_n-1\}_{\pm}^{f_1+n-2\beta-1}, -1). \end{aligned}$$

With the help of (3.8) this reduces to

$$\begin{aligned}\mathscr{K}^{\beta-f_1+1}_{f_1+n} & = (-1)^n(\{f_2-1, f_3-1, \ldots, f_n-1\}_{\pm}^{f_1+n-2\beta-1}, 0) \\ & = -(1-n, \{f_2, \ldots, f_n\}_{\pm}^{f_1+n-2\beta+1}). \end{aligned} \tag{3.29a}$$

Comparison with the definition of B yields

$$\mathscr{K}^{\beta-f_1+1}_{f_1+n} = -B^{\beta}_{f_1+n-1},$$

and similarly we have

$$\mathscr{K}^{\beta-f_1}_{f_1+n-1} = -B^{\beta}_{f_1+n}. \tag{3.29b}$$

As a consequence, the two sums in (3.28) cancel exactly, and we get

$$\mu = B_0^0 = (f_1, f_2, \ldots, f_n), \tag{3.30}$$

as announced in Theorem 3.

The decomposition formula for spinor representations

Examination of some simple examples again suggests that the correct building blocks for the decomposition formula are the D_k introduced in (3.13). The correct modification of (3.14) is then easily found, and will now be given. Let λ be a spinor representation of SO($2n+1$), with signature $(f_1, f_2, \ldots, f_n)$, and define the integers

$$r_i = f_i - f_{i+1}, \quad i = 1, \ldots, n-1, \quad r_n = f_n - \tfrac{1}{2}. \tag{3.31}$$

Theorem 4: Let

$$\mu = s \otimes \begin{vmatrix} D_1 & D_0 & 0 & & & & & & \\ D_2 & D_1 & D_0 & & & & & & \\ & D_1 & D_0 & 0 & & & & & \\ & D_3 & D_2 & D_1 & & & & & \\ & & & \ddots & & & & & \\ & & & & D_2 & & & & \\ & & & & & \ddots & & & \\ & & & & & & D_n \ominus D_{n-r_1} & & \\ & & & & & & & \ddots & \\ & & & & & & & & D_n \ominus D_{n-4} \\ & & & & & & & & D_{n+1} \ominus D_{n-3} \end{vmatrix} \begin{matrix} \otimes \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} r_1 \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} r_2 \\ \vdots \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} r_n \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ D_{n-1} \ominus D_{n-3} \\ D_n \ominus D_{n-2} \end{matrix} \tag{3.32}$$

or in components

$$\mu = s \otimes \det \mathscr{D}, \tag{3.33}$$

$$\mathscr{D}_{ij} = D_{i-j+k} \ominus D_{i-j+k-2-2f_i}$$

and k is defined by $r_1 + r_2 + \cdots + r_{k-1} < i \leqslant r_1 + r_2 + \cdots + r_k$. Then $\mu = \lambda$. Please note the difference in *sign* between (3.15) and (3.33) as well as the difference in index in the second term.

Outline of the proof: The proof proceeds by induction on f_1 as before. The definitions of generalized signatures and braces { } of (3.17) are extended to spinor representations and μ is again computed directly using the expansion of determinant (3.32) along the first column. Proceeding along the lines of the proof of Theorem 3, we find that we must calculate

$$\mu = \sum_{\alpha=1}^{f_1+n-1} (D_\alpha \ominus D_{-2f_1+\alpha}) \otimes (f_1-\alpha, f_2, ..., f_n)(-1)^{\alpha-1}. \tag{3.34}$$

The tensor products are evaluated with the help of the weight diagram of d_p in (3.5) and collected with the brace notation (3.17). After simplifications analogous to those made in the proof of Theorem 3, we find

$$\mu = (f_1, f_2, ..., f_n), \tag{3.35}$$

as announced.

In both Theorems 3 and 4, we have decomposed all representations of SO($2n+1$) in terms of $d_1, d_2, ..., d_n$ and s, even though d_n is not on the list of basis representations in (1.6b). We have done so because $d_1, d_2, ..., d_n$ and s form the natural set in terms of which the decomposition formulae are simplest. In addition, this presents no loss of generality since d_n itself is expressed in terms of the set of basis representations (1.6b) by a simple formula, which we shall now derive. From (3.10), we deduce

$$s \otimes s = \oplus \sum_{\sigma_i=0,1} (\sigma_1, \sigma_2, ..., \sigma_n). \tag{3.36}$$

Cancelling nondominant signatures leaves us with

$$s \otimes s = \oplus \sum_{p=0}^{n} d_p \tag{3.37}$$

so that

$$d_n = s \otimes s \ominus \sum_{p=0}^{n-1} d_p. \tag{3.38}$$

IV. THE SYMPLECTIC GROUP Sp(n)

The representation theory for the symplectic group is much simpler than that for SO($2n+1$), since there are no spinor representations. Moreover, the decomposition formuale as well as their proof are very similar to the case of SO($2n+1$). For this reason, we just quote the results for the decomposition formula; the reader should have no problem reconstructing the proof.

Multiplication of a basis representation with an arbitrary representation

The weight diagram of the representation d_p with signature $(\underbrace{1,1,...,1}_{p},0,...,0)$ (for $p = 1, ..., n$) is deduced from (1.4c) and can be conveniently expressed as

$$\chi_{d_p} = (\mathscr{A}_p - \mathscr{A}_{p-2})/\mathscr{A}_0. \tag{4.1}$$

The function $\mathscr{A}_p$ has been defined in (3.2), and the resulting weight diagram of d_p is found with the help of (3.4)

$$\chi_{d_p} = \sum_{\alpha=0}^{[p/2]} C^{\alpha}_{n-p+2\alpha} R(p-2\alpha) - \sum_{\alpha=0}^{[(p-2)/2]} C^{\alpha}_{n-p+1+2\alpha} R(p-2-2\alpha). \tag{4.2}$$

Here R is the function defined in (3.4b), and the product of R with the character of a representation with signature $(f_1, f_2, ..., f_n)$ is given by

$$R(q)\chi_\lambda = \sum_{i_1<i_2<\cdots<i_q} \sum_{\sigma_{i_j}=\pm 1} \chi_{(f_1,...,f_{i_1}+\sigma_{i_1},...,f_{i_q}+\sigma_{i_q},...,f_n)}. \tag{4.3}$$

The tensor product of d_p with the representation λ is then simply obtained combining (4.2) and (4.3).

The decomposition formula

Let λ be a representation with signature $(f_1, f_2, ..., f_n)$. Define the integers r_i by

$$r_i = f_i - f_{i+1}, \quad i = 1,...,n-1, \quad r_n = f_n \tag{4.4}$$

as well as the sequence of reducible representations

$$\begin{aligned} &0 \leqslant k \leqslant n, && \hat{D}_k = d_k \oplus d_{k-2} \oplus d_{k-4} \oplus \cdots \oplus \begin{cases} d_1 & \text{if } k \text{ odd}, \\ d_0 & \text{if } k \text{ even}, \end{cases} \\ &n \leqslant k \leqslant 2n, && \hat{D}_k = \hat{D}_{2n-k}, \\ &k > 2n \text{ or} \\ &k < 0, && \hat{D}_k = 0. \end{aligned} \tag{4.5}$$

Theorem 5: Define

$$\mu = \begin{vmatrix} \hat{D}_1\hat{D}_0 0 & & & & & & \\ \hat{D}_2\hat{D}_1\hat{D}_0 & & & & & & \\ \vdots \quad \ddots & & & & & & \\ \hat{D}_{r_1} \cdots \hat{D}_1 & & & & & & \\ & \hat{D}_3\hat{D}_2\hat{D}_1 \cdots & & & & & \\ & & \ddots & & & & \\ & & & \hat{D}_2 & & & \\ & & & & \ddots & & \\ & & & & \hat{D}_n \ominus \hat{D}_{n-4} & \hat{D}_{n-1} \ominus \hat{D}_{n-3} & \\ & & & & \hat{D}_{n+1} \ominus \hat{D}_{n-1} & \hat{D}_n \ominus \hat{D}_{n-2} & \end{vmatrix}^{\otimes} \tag{4.6}$$

(braces: r_1, r_2, r_n)

or in components

$$\begin{aligned} &\mu = \otimes \det \mathscr{D} \\ &\mathscr{D}_{ij} = \hat{D}_{i-j+k} \ominus \hat{D}_{i-j+k-2-2f_1} \end{aligned} \tag{4.7}$$

and k is defined by $r_1 + r_2 + \cdots + r_{k-1} < i \leqslant r_1 + r_2 + \cdots + r_l$. Then $\mu = \lambda$.

V. THE ORTHOGONAL GROUP SO($2n$)

According to whether $f_n = 0$ or $\neq 0$, the characters of the group SO(2n) defined in (1.4d) correspond to irreducible or reducible representations. When $f_n = 0$, the representations is non-self-associate, and the character is simple. When $f_n \neq 0$, the representation is self-associate, reducible into two associate irreducible representations of the same dimension, and the character is said to be a double character. We shall

show how to construct the tensor product of an arbitrary representation λ with a basis representation (whether self-associate or not) and if the basis representation is reducible, we shall also show how to find the product of its irreducible associate components with λ. A decomposition formula will be proven for both self-associate and non-self-associate representations. We have not found a decomposition formula for the irreducible components of a self-associate representation.

Multiplication of a generating representation with an arbitrary representation

We shall need the following generating representations

$$\begin{aligned} &d_p = (\underbrace{1,1,\ldots,1}_{p},0,\ldots,0), \qquad p = 1,\ldots,n-1, \\ &d_n^{\pm} = (1,1,\ldots,\pm 1), \quad s^{\pm} = (\tfrac{1}{2},\tfrac{1}{2},\ldots,\pm\tfrac{1}{2}), \\ &d_n = d_n^{+} \oplus d_n^{-}, \quad s = s^{+} \oplus s^{-}. \end{aligned} \tag{5.1}$$

The representations $\pm$ are associate to each other, whereas s and d_n are self-associate. When n is odd, $\pm$ are actually complex conjugates, whereas, for n even, both $+$ and $-$ are real. We first determine the weight diagrams of d_p and s and then indicate how those of $s^{\pm}$ and $d_n^{\pm}$ may be gotten.

From Weyl's *first* formula (1.4d) and using the definition of the function $\mathscr{A}_p$ in (3.2), we see that

$$\chi_{d_p} = \mathscr{A}_p / \mathscr{A}_0. \tag{5.2}$$

With the help of (3.4b), χ_{d_p} may be expressed in terms of R:

$$\chi_{d_p} = \sum_{\alpha=0}^{[p/2]} C^{\alpha}_{n-p+2\alpha} R(p-2\alpha). \tag{5.3}$$

The double character of the spinor representation is given by

$$\chi_s = \sum_{\sigma_i = \pm 1} \epsilon_1^{\sigma_1/2} \epsilon_2^{\sigma_2/2} \cdots \epsilon_n^{\sigma_n/2}. \tag{5.4}$$

The quantity σ defined as

$$\sigma = \prod_{i=1}^{n} \sigma_i \tag{5.5}$$

may take the values ± 1 in (5.4). The characters of s^{+} (resp. s^{-}) are also defined by (5.4), but now σ must be restricted to be 1 (resp. -1). The expression for the character of $d_n^{\pm}$ is more complicated, and we shall not give it here. It may be deduced from the relation

$$d_n^{\pm} = s^{\pm} \otimes s^{\pm} - d_{n-2} - d_{n-4} - \cdots. \tag{5.6}$$

For representations with $f_n = 0$ and self-associate representations, a generalized signature may be defined:

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1} + \epsilon^{-l_1}, \ldots, \epsilon^{l_n} + \epsilon^{-l_n}|}{|\epsilon^{l_1^0} + \epsilon^{-l_1^0}, \ldots, 1|} \tag{5.7}$$

even if $(f_1,f_2,\ldots,f_n)$ is not dominant. For the two irreducible associate representations into which a self-associate representation decomposes, similar generalized signatures may be defined, but we shall not need these here.

Tensor multiplication is effected using formulae (5.3)–(5.4); the product of R with the character of an arbitrary representation λ with signature $(f_1,f_2,\ldots f_n)$ is given by

$$R(q)\chi_\lambda = \sum_{i_1 < i_2 \cdots < i_q} \sum_{\sigma_{i_j} = \pm 1} G(\sigma) \chi_{(f_1,f_2,\ldots,f_{i_1}+\sigma_{i_1},\ldots,f_{i_q}+\sigma_{i_q},\ldots,f_n)} \tag{5.8}$$

and the product of χ_s and χ_λ is

$$\chi_s \chi_\lambda = \sum_{\sigma_i = \pm 1} F(\sigma) \chi_{(f_1+\sigma_1/2,\ldots,f_n+\sigma_n/2)}. \tag{5.9}$$

The integers $G(\sigma)$ and $F(\sigma)$ are present to obtain the correct counting of self-associate and non-self-associate representations. They are determined from (3.4b) and (5.7) using the invariance under the action of the Weyl group: first we make $(f_1,f_2,\ldots,f_n)$ *dominant*. For the integer G we have

$$\begin{aligned} &G(\sigma) = 1 \quad \text{if } i_q \neq n \text{ or } \sigma_n \neq -1, \\ &G(\sigma) = 1 \quad \text{if } \sigma_n = -1 \text{ and } f_n > 1, \\ &G(\sigma) = 2 \quad \text{if } \sigma_n = -1 \text{ and } f_n = 1, \\ &G(\sigma) = 0 \quad \text{if } \sigma_n = -1 \text{ and } f_n = 0. \end{aligned} \tag{5.10}$$

For the integer $F(\sigma)$ we have

$$\begin{aligned} &F(\sigma) = 1 \quad \text{if } \sigma_n = 1, \\ &F(\sigma) = 1 \quad \text{if } \sigma_n = -1 \text{ and } f_n \neq \tfrac{1}{2}, \\ &F(\sigma) = 2 \quad \text{if } \sigma_n = -1 \text{ and } f_n = \tfrac{1}{2}. \end{aligned} \tag{5.11}$$

Products with the representations s^{+} or s^{-} are obtained by making the appropriate restrictions on σ given in (5.9).

As an example of these multiplication rules, one may compute the following product for SO(10).[10] [We use the definition $R(q) = \operatorname{tr}\rho(q)$.]

$$\begin{aligned} &(1,1,1,1,0) \otimes (3,2,2,2,1) = [\rho(4) + 3\rho(2) + 10\rho(0)] \otimes (3,2,2,2,1), \\ &\rho(4) \otimes (3,2,2,2,1) = (4,\{2,2,2,1\}^3_{\pm}) \oplus (3,\{2,2,2,1\}^4_{\pm}) \oplus (2,\{2,2,2,1\}^3_{\pm}) \\ &\quad = (4,3,3,3,1) + (4,3,3,2,2) + 2(4,3,3,2,0) + 2(4,3,2,1,0) \\ &\qquad + (4,3,3,1,1) - (4,3,2,2,1) + (4,3,1,1,1) - 2(4,2,2,2,2) - 4(4,2,2,2,0) + 2(4,2,1,1,0) \\ &\qquad - (4,2,2,1,1) + (4,1,1,1,1) + (3,3,3,3,2) + 2(3,3,3,3,0) + 2(3,3,3,1,0) - (3,3,2,2,2) \\ &\qquad - 2(3,3,2,2,0) + 2(3,3,1,1,0) - 2(3,2,2,1,0) + 2(3,1,1,1,0) - 2(2,2,2,2,2) - 4(2,2,2,2,0) \\ &\qquad - (2,2,2,1,1) + (2,1,1,1,1) + 2(2,2,1,1,0), \\ &\rho(2) \otimes (3,2,2,2,1) = (4,\{2,2,2,1\}^1) + (3,\{2,2,2,1\}^2) + (2,\{2,2,2,1\}^1) \\ &\quad = (4,3,2,2,1) + (4,2,2,1,1) + (4,2,2,2,2) + 2(4,2,2,2,0) \\ &\qquad + (3,3,3,2,1) + (3,3,2,1,1) + (3,3,2,2,2) + 2(3,3,2,2,0) + 2(3,2,2,1,0) \\ &\qquad - (3,2,2,2,1) + (2,2,2,2,2) + (2,2,2,1,1) + 2(2,2,2,2,0) - (3,2,2,2,1) + (3,2,1,1,1). \end{aligned}$$

Putting all together, we obtain

$$(1,1,1,1,0)\otimes(3,2,2,2,1) = (4,3,3,3,1) + (4,3,3,2,2) + 2(4,3,3,2,0) + 2(4,3,2,1,0) + (4,3,3,1,1) + 2(4,3,2,2,1) + (4,3,1,1,1) + (4,2,2,2,2) + 2(4,2,2,2,0) + 2(4,2,1,1,0) + 2(4,2,2,1,1) + (4,1,1,1,1) + (3,3,3,3,2) + 2(3,3,3,3,0) + 2(3,3,3,1,0) + 2(3,3,2,2,2) + 4(3,3,2,2,0) + 2(3,3,1,1,0) + 4(3,2,2,1,0) + 2(3,1,1,1,0) + (2,2,2,2,2) + 2(2,2,2,2,0) + 2(2,2,2,1,1) + (2,1,1,1,1) + 3(3,3,2,1,1) + 4(3,2,2,2,1) + 3(3,3,3,2,1) + 2(2,2,1,1,0) + 3(3,2,1,1,1). \tag{5.12}$$

With the help of the tables of dimensions of representations,[9] we may check that dimensions work out correctly:

$$\begin{aligned}210\times 50\,688 = {} & 945\,945 + 660\,660 + 1\,698\,840 + 1\,048\,576 \\ & + 882\,882 + 2\times 848\,925 + 242\,550 + 90\,090 + 274\,560 + 143\,000 \\ & + 2\times 199\,017 + 17\,325 + 84\,942 + 165\,165 + 210\,210 + 128\,700 \\ & + 2\times 189\,189 + 73\,710 + 2\times 72\,765 + 8085 + 2772 + 8910 \\ & + 2\times 6930 + 1050 + 3\times 128\,700 + 4\times 50\,688 + 3\times 219\,648 + 5940 \\ & + 3\times 23\,040.\end{aligned}$$

The decomposition formula for nonspinor representations simple and double characters

Let λ be a representation with (dominant) signature $(f_1,f_2,\ldots,f_n)$[10] and define the sequence of representations

$$\begin{aligned} &0\leqslant k\leqslant n, && \delta_k = d_k \\ &n\leqslant k\leqslant 2n, && \delta_k = d_{2n-k}, \\ &k\geqslant 2n \text{ or } k<0, && \delta_k = 0.\end{aligned} \tag{5.13}$$

Note that all nonzero representations in this sequence are irreducible and that d_n is self-associate. We shall now prove the following decomposition theorem in the case of nonspinor representations.

Theorem 6: Let

$$\begin{aligned}&\mu = \otimes \det \mathscr{D}, \\ &\mathscr{D}_{ij} = (\delta_{i-j+k}\oplus\delta_{i+j+k-2f_i})/(1+\delta_{j,f_i})\end{aligned} \tag{5.14}$$

and let k be defined by $f_1 - f_k < i \leqslant f_1 - f_{k+1}$. Then $\mu = \lambda$.

Proof: As for the other three classical groups, the proof proceeds by induction on f_1. Expansion of determinant (5.14) along the first column and the use of generalized signatures reduce the calculation of μ to the evaluation of the following expression:

$$\begin{aligned}\mu = {} & \sum_{\alpha=1}^{f_1+n-1} (\delta_\alpha \oplus \delta_{\alpha-2f_1+2}) \\ & \otimes (f_1-\alpha, f_2,\ldots,f_n)(-1)^{\alpha-1}.\end{aligned} \tag{5.15}$$

To work out the products in (5.15), we use (5.8) and rearrange different contributions to the product of $R(q)$ and χ_λ with the help of the brace notation introduced in (3.17).

$$\begin{aligned}&\rho(q)\otimes(f_1-\alpha,f_2,\ldots,f_n) \\ &\quad = (f_1-\alpha+1,\{f_2,\ldots,f_n\}_{\pm}^{q-1}) \\ &\qquad \oplus (f_1-\alpha,\{f_2,\ldots,f_n\}_{\pm}^{q}) \\ &\qquad \oplus G(f_1-\alpha-1,\{f_2,\ldots,f_n\}_{\pm}^{q-1}),\end{aligned} \tag{5.16a}$$

where G is determined by the rules of (5.10):

$$\begin{aligned}&G=0 \quad \text{if } \alpha = f_1+n-1, \\ &G=2 \quad \text{if } \alpha = f_1+n-2, \\ &G=1 \quad \text{if otherwise.}\end{aligned} \tag{5.16b}$$

Then we make use of (5.3) and obtain

$$\begin{aligned}&\delta_\alpha\otimes(f_1-\alpha,f_2,\ldots,f_n) \\ &\quad = \sum_{\beta=0}^{[\alpha/2]} C^{\beta}_{n-\alpha+2\beta}(B^{\beta}_{\alpha-1}\oplus B^{\beta}_{\alpha}\oplus GB^{\beta+1}_{\alpha+1}),\end{aligned} \tag{5.17a}$$

where G is defined in (5.16b):

$$\begin{aligned}&\delta_{\alpha-2f_1+2}\otimes(f_1-\alpha,f_2,\ldots,f_n) \\ &\quad = \sum_{\beta=0}^{[\alpha/2]-f_1+1} C^{\beta}_{n-\alpha+2f_1-2+2\beta} \\ &\qquad \times(B^{\beta+f_1-1}_{\alpha-1}\oplus B^{\beta+f_1-1}_{\alpha}\oplus GB^{\beta+f_1}_{\alpha+1}).\end{aligned} \tag{5.17b}$$

Here we have made use of the quantity

$$B^{\beta}_{\alpha} = (f_1-\alpha,\{f_2,f_3,\ldots,f_n\}_{\pm}^{\alpha-2\beta}). \tag{5.18}$$

Shifts in summation variables, the use of Pascal's equality, and the explicit definition of G lead to

$$\begin{aligned}&\mu = \mu_1\oplus\mu_2, \\ &\mu_1 = B^0_0 \oplus \sum_{\beta=0}^{[(f_1+n)/2]} C^{\beta}_{-f_1+2\beta}B^{\beta}_{f_1+n-1}(-1)^{f_1+n} \\ &\qquad \ominus \sum_{\beta=0}^{[(f_1+n)/2]} C^{\beta-1}_{-f_1+2\beta}B^{\beta}_{f_1+n-1}(-1)^{f_1+n},\end{aligned} \tag{5.19a}$$

$$\begin{aligned}&\mu_2 = \sum_{\beta=0}^{[(-f_1+n)/2]+1} C^{\beta}_{f_1-2+2\beta}B^{\beta+f_1-1}_{f_1+n-1}(-1)^{f_1+n} \\ &\qquad \ominus \sum_{\beta=0}^{[(-f_1+n)/2]+1} C^{\beta-1}_{f_1-2+2\beta}B^{\beta+f_1-1}_{f_1+n-1}(-1)^{f_1+n}.\end{aligned} \tag{5.19b}$$

Making the substitution $\beta\to\beta-f_1+1$ in (5.19a) and using the properties of the binomial coefficients, it can be shown that the four β summations in (5.19) precisely cancel, leaving only $\mu = B^0_0 = (f_1,\ldots,f_n) = \lambda$ as announced.

The decomposition formula for spinor representations

Finally we shall exhibit a decomposition formula in the case of the (always self-associate) spinor representations. The proof is completely analogous to the proofs of Theorems 4 and 6 and will not be given here. We define the same sequence of representations δ_k in (5.13), let λ be a representation with signature $(f_1,\ldots,f_n)$.

Theorem 7: Let

$$\mu = \otimes \det \mathscr{D} \otimes s,$$

$$\mathscr{D}_{ij} = \delta_{i-j+k} \ominus \delta_{i+j+k-2f_1-1}, \tag{5.20}$$

and let k be defined by $f_1 - f_k < i \leqslant f_1 - f_{k+1}$. Then $\mu = \lambda$. Please note the difference in sign and the difference in indices between (5.20) and (5.14).

VI. APPLICATIONS

A. Multiplication of arbitrary representations

Several algorithms exist in the literature for the decomposition into irreducible representations of the tensor product of two irreducible representations.[2,3,5] If the weight diagram of one of the representations is known, Weyl's first formula can be used to obtain the irreducible components.[2,3] However, the determination of the weight diagram is a notoriously difficult problem. Želobenko[3] and Murnaghan[5] use Weyl's second formula, respectively for the unitary and orthogonal groups. The knowledge of the product of a symmetric representation with an arbitrary representation then suffices to perform the product of two arbitrary representations. This method is very attractive for the unitary groups,[3] but appears quite involved for the orthogonal groups.[5]

With the Theorems 2–7, we dispose of decomposition formulae in terms of the *basis* representations. In proving these relations, we have also shown how to perform the tensor product of any of these basis representations with an arbitrary representation. Thus, we dispose of an algorithm that allows us to compute the tensor product of two arbitrary representations, and the rules of this algorithm seem rather convenient, even though the calculations remain lengthy.

To demonstrate the practicality of these rules, we shall work out an example of intermediate difficulty: the tensor product in Sp(4) of the representations α and β with signatures (2,1,1,0) and (3,2,2,1). To do so, we use Theorem 5 for α:

$$(2,1,1,0) \otimes (3,2,2,1) = \begin{vmatrix} \hat{D}_1 & \hat{D}_0 \\ \hat{D}_4 - \hat{D}_0 & \hat{D}_3 - \hat{D}_1 \end{vmatrix}^{\otimes} \otimes (3,2,2,1)$$
$$= \hat{D}_1 \otimes (\hat{D}_3 - \hat{D}_1) \otimes (3,2,2,1) - (\hat{D}_4 - \hat{D}_0) \otimes (3,2,2,1),$$

$$(\hat{D}_3 - \hat{D}_1) \otimes (3,2,2,1) = (4,3,3,1) + (4,3,1,1) + (4,3,2,2) + (4,3,2,0)$$
$$+ (4,2,1,0) + (4,2,2,1) + (4,1,1,1) + (3,3,3,2) + (3,3,3,0) + (3,3,1,0)$$
$$+ (3,2,2,2) + (3,2,2,0) + (3,1,1,0) + (2,2,1,0) + (2,2,2,1) + (2,1,1,1)$$
$$+ 2(3,3,2,1) + 2(3,2,1,1),$$

$$(\hat{D}_4 - \hat{D}_0) \otimes (3,2,2,1) = (4,3,3,2) + (4,3,3,0) + (4,3,1,0) + (4,2,2,2)$$
$$+ (4,2,2,0) + (4,1,1,0) + (2,2,2,2) + (2,2,2,0) + (2,1,1,0) + 2(4,3,2,1)$$
$$+ 2(4,2,1,1) + 2(3,3,3,1) + 2(3,3,1,1) + 2(3,3,2,2) + 2(3,3,2,0) + 2(3,2,1,0)$$
$$+ 2(3,1,1,1) + 3(3,2,2,1) + 2(2,2,1,1).$$

Putting all together, we obtain

$$(2,1,1,0) \otimes (3,2,2,1) = (5,3,3,1) + (4,4,3,1) + (5,3,1,1) + (4,4,1,1)$$
$$+ 5(4,3,2,1) + (5,3,2,2) + (4,4,2,2) + 2(4,3,3,2) + (5,3,2,0)$$
$$+ (4,4,2,0) + 2(4,3,3,0) + 3(4,3,1,0) + (5,2,1,0) + 3(4,2,2,0)$$
$$+ (4,2,0,0) + (5,2,2,1) + 4(4,2,1,1) + 2(4,2,2,2) + (5,1,1,1)$$
$$+ 2(4,1,1,0) + (3,3,3,3) + 3(3,3,3,1) + 4(3,3,2,0) + (3,3,0,0)$$
$$+ (2,2,2,2) + 3(3,3,2,2) + 2(2,2,2,0) + 5(3,2,1,0) + 5(3,2,2,1)$$
$$+ 2(2,1,1,0) + (3,1,0,0) + 3(3,1,1,1) + (2,2,0,0) + (1,1,1,1)$$
$$+ 4(3,3,1,1) + 3(2,2,1,1).$$

It is also useful to check the dimensions using the tables[9]:

$$315 \times 6237 = 213\,444 + 122\,850 + 96\,228 + 41\,250 + 5 \times 65\,536$$
$$+ 142\,155 + 67\,760 + 2 \times 56\,628 + 146\,250 + 66\,528 + 2 \times 42\,042$$
$$+ 3 \times 29\,106 + 36\,864 + 3 \times 16\,848 + 4914 + 63\,063 + 4 \times 14\,300$$
$$+ 2 \times 13\,728 + 9009 + 2 \times 3696 + 4719 + 3 \times 12\,012 + 4 \times 10\,010$$
$$+ 2184 + 594 + 3 \times 9009 + 2 \times 825 + 5 \times 4096 + 5 \times 6237$$
$$+ 2 \times 315 + 594 + 3 \times 1155 + 308 + 42 + 4 \times 7020 + 3 \times 792.$$

B. A relation between the dimensions of the representations of Sp(n) and spinor representations of SO($2n+1$)

Formulae (3.32) and (4.6) have the same formal structure except for the overall tensor product with the spinors in (3.32) and the difference in definition for D_k and $\hat{D}_k$. In particular, the value of the characters of D_k and $\hat{D}_k$ at the identity of the group can be shown to be equal. Indeed, upon using formulae (3.4)–(3.6) and (3.13) on one hand and formulae (4.2) and (4.5) on the other, we find that

$$\chi_{D_k}(e) = \chi_{\hat{D}_k}(e) = \sum_{\alpha=0}^{[k/2]} C^{\alpha}_{n-k+2\alpha} 2^{k-2\alpha} C^{k-2\alpha}_{n} \tag{6.1}$$

Substitution of (6.1) into (3.32) and (4.6) yields the following result. Let λ be an arbitrary representation of Sp(n) with dominant signature $(f_1, f_2, \ldots, f_n)$, let Λ be the representation of SO($2n+1$) with signature $(f_1 + \frac{1}{2}, f_2 + \frac{1}{2}, \ldots, f_n + \frac{1}{2})$, and let s be the fundamental spinor of SO($2n+1$) with signature $(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$. Then we have

$$\dim \Lambda = \dim s \dim \lambda. \tag{6.2}$$

In fact, it is also clear from (3.5) and (3.13) that, in the canonical basis, a more general relation holds

$$\chi_{\Lambda}(h) = \chi_s(h)\chi_{\lambda}(h), \tag{6.3}$$

where h is an element of the Cartan subgroup, parametrized by the angles $\phi_1, \ldots, \phi_n$.

C. Restriction of a representation to a subgroup

Let G be any of the four classical groups, and let G_0 be any of its classical nontrivial subgroups. We wish to determine the irreducible components of the restriction of the representation λ of G to G_0. If, by classical methods, we can derive the restriction of the basis representations of G to the subgroup G_0, then we can calculate the restriction of any representation by Theorems 2–7.

We shall treat the following simple example:

$$G = \mathrm{SU}(2n), \quad G_0 = \mathrm{SO}(2n).$$

The restriction of the basis representations of SU($2n$) to SO($2n$) are *irreducible* and given by[11]

$$d^{A}_{k}|_{\mathrm{SO}(2n)} = \delta^{D}_{k}, \quad k = 1, \ldots, 2n, \tag{6.4}$$

where δ_k is defined by (5.13). The restriction of a representation of SU($2n$) is then given by determinant (2.14) in which d^{A}_{k} is replaced by δ^{D}_{k}. It is usually not necessary to fully work out the products in this new determinant, as often irreducible representations of SO($2n$) may be recognized in it. Consider, e. g., the restriction of the representation (2,2,2,1,1,0,0) of SU(8) to SO(8),

$$(2,2,2,1,1,0,0,0)_A\,|_{\mathrm{SO}(8)} = \begin{vmatrix} d^{A}_{3} & d^{A}_{2} \\ d^{A}_{6} & d^{A}_{5} \end{vmatrix}_{\mathrm{SO}(8)} = \begin{vmatrix} \delta^{D}_{3} & \delta^{D}_{2} \\ \delta^{D}_{2} & \delta^{D}_{3} \end{vmatrix},$$

$$\delta^{D}_{3} \otimes \delta^{D}_{3} = (2,2,2,0) + (2,2,1,1) + (2,2,0,0) + 2(2,1,1,0) + (2,0,0,0) + (1,1,1,1) + 2(1,1,0,0) + (0,0,0,0),$$

$$\delta^{D}_{2} \otimes \delta^{D}_{2} = (2,2,0,0) + (2,1,1,0) + (2,0,0,0) + (1,1,1,1) + (0,0,0,0) + (1,1,0,0),$$

$$(2,2,2,1,1,0,0,0)_A\,|_{\mathrm{SO}(8)} = (2,2,2,0)_D + (2,2,1,1)_D + (2,1,1,0)_D + (1,1,0,0)_D. \tag{6.5}$$

Using the tables,[9] we can easily check that the dimensions work out:

$$2352_A = 840_D + 567_D + 567_D + 350_D + 28_D. \tag{6.6}$$

In a completely analogous fashion, the restrictions of the basis representation of SU($2n+1$) to SO($2n+1$) are also irreducible, and can be used to calculate the restrictions of arbitrary representations to SO($2n+1$).

The restrictions of the basis representations of SU($2n$)to Sp(n) are reducible and can be easily derived using conventional methods:

$$d^{A}_{k}|_{\mathrm{Sp}(n)} = \hat{D}^{C}_{k}, \tag{6.7}$$

where $\hat{D}_k$ has been defined in (4.5). We shall illustrate this restriction with an extremely simple example, the decomposition of the representation α of SU(6) with signature (2,2,1,1,0) to Sp(3):

$$\alpha|_{\mathrm{Sp}(3)} = \begin{vmatrix} d^{A}_{2} & d^{A}_{1} \\ d^{A}_{5} & d^{A}_{4} \end{vmatrix}^{\otimes}_{\mathrm{Sp}(3)} = \begin{vmatrix} \hat{D}^{C}_{2} & \hat{D}^{C}_{1} \\ \hat{D}^{C}_{1} & \hat{D}^{C}_{2} \end{vmatrix}^{\otimes}.$$

Working out these products, one finds

$$(2,2,1,1,0,0)_A\,|_{\mathrm{Sp}(3)} = (2,2,0)_C \oplus (2,1,1)_C \oplus 2(1,1,0)_C \otimes (0,0,0)_C \tag{6.8}$$

with dimensions

$$189_A = 90_C + 70_C + 2\times 14_C + 1_C.$$

The peculiar property of this algorithm is that we only need to know the restrictions of a *finite* number of representations to compute that of all representations. The procedure can be easily generalized to arbitrary classical groups G and G_0.

Note added in manuscript: The problem of decomposing a given representation into a finite set of basis representations has also been discussed by A. J. Feingold, Proc. Am. Math. Soc. **70**, 109 (1978). I thank Professor J. Patera for drawing my attention to this work.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor Bob Sharp, Dr. Yvan Saint-Aubin, and Dr. Baha Balantekin for stimulating discussions and for helpful remarks on the manuscript.

[1] For a review of the theory of compact Lie groups, see Refs. 2 and 3.

[2] H. Weyl, *The Classical Groups* (Princeton U. P., Princeton, NJ, 1973), and references therein.

[3] D. P. Želobenko, *Compact Lie Groups and Their Representations* (American Mathematical Society, Providence, RI, 1973).

[4] Variants of Weyl's second formula have been studied in N. E. Samra and R. C. King, J. Phys. A. Gen. **12**, 2305 (1979); R. C. King, J. Math. Phys. **12**, 1588 (1971); M. J. Newell, Proc. Roy. Irish Acad., 153 (1951).

[5] F. D. Murnaghan, *The Theory of Group Representations* (Baltimore, 1938).

[6] It is understood that, in the expansion of the determinant, all products are tensor products and all sums are direct sums. Whenever a minus sign occurs, it is understood that the representation is multiplied by -1, and will cancel an identical term with a $+$ sign.

[7] A formula essentially the same as (2.14) appears in the following references for the representations of the symmetric group and for that of the unitary group, respectively: D. E. Littlewood, *The Theory of Group Characters and Matrix Representations of Groups* (Oxford U. P., Oxford, 1940); M. J. Newell, Proc. Roy. Irish Acad., 345 (1949). (I am grateful to A. B. Balantekin for bringing the latter reference to my attention.)

[8] By definition, a character vanishes identically on the group if and only if its signature vanishes.

[9] W. G. McKay and J. Patera, *Table of Dimensions, Indices and Branching Rules for Representations of Simple Lie Algebras* (Marcel Dekker, New York, 1981).

[10] It is understood that in the case $f_n \neq 0$, the representation is self-associate and that its character is double.

[11] In this section, we shall indicate with a superscript to which group the representation d_k belongs. The superscripts are A, B, C, and D for respectively SU(n), SO($2n+1$), Sp(n), and SO($2n$).

On classes of integrable systems and the Painlevé property[a)]

John Weiss
La Jolla Institute, 8950 Villa La Jolla Drive, Suite 2150, La Jolla, California 92037 and Institute for Pure and Applied Physical Science, University of California, San Diego, La Jolla, California 92093

(Received 14 March 1983; accepted for publication 9 September 1983)

The Caudrey–Dodd–Gibbon equation is found to possess the Painlevé property. Investigation of the Bäcklund transformations for this equation obtains the Kuperschmidt equation. A certain transformation between the Kuperschmidt and Caudrey–Dodd–Gibbon equation is obtained. This transformation is employed to define a class of p.d.e.'s that identically possesses the Painlevé property. For equations within this class Bäcklund transformations and rational solutions are investigated. In particular, the sequences of higher order KdV, Caudrey–Dobb–Gibbon, and Kuperschmidt equations are shown to possess the Painlevé property.

PACS numbers: 02.30. + g

1. INTRODUCTION

In Ref. 1 the Painlevé property for partial differential equations was defined. Briefly, we say that a partial differential equation has the Painlevé property when the solutions of the p.d.e. are "single-valued" about the movable, singularity manifold and the singularity manifold is "noncharacteristic." To be precise, if the singularity manifold is determined by

$$\varphi(z_1,z_2,\ldots,z_n)=0 \tag{1.1}$$

and $u=u(z_1,\ldots,z_n)$ is a solution of the p.d.e., then we require that

$$u=\varphi^{\alpha}\sum_{j=0}^{\infty}u_j\varphi^{j}, \tag{1.2}$$

where $u_0\neq 0$, $\varphi=\varphi(z_1,\ldots,z_n)$, $u_j=u_j(z_1,\ldots,z_n)$ are analytic functions of (z_j) in a neighborhood of the manifold (1.1), and α is an integer. The requirement that the manifold (1.1) be noncharacteristic insures that the expansion (1.2) will be well defined, in the sense of the Cauchy–Kowalevsky theorem. Substitution of (1.2) into the p.d.e. determines the value(s) of α, and defines the recursion relations for u_j, $j=0,1,2,\ldots$. When the anzatz (1.2) is correct, the p.d.e. is said to possess the Painlevé property and is conjectured to be integrable. The "Painlevé conjecture," as originally formulated by Ablowitz *et al.*,[2] states that when all the ordinary differential equations obtained by exact similarity transforms from a given partial differential equation have the Painlevé property, then the partial differential equation is "integrable." The above definition of the "Painlevé property" allows this conjecture to be stated directly for the partial differential equation.

In Ref. 3 Bäcklund transformations were obtained by truncating the expansion (1.2) at the "constant" level term. That is, we set

$$u=u_0\varphi^{-N}+u_1\varphi^{-N+1}+\cdots+u_N \tag{1.3}$$

and find, from the recursion relations for u_j, an overdetermined system of equations for $(\varphi,u_j,j=0,1,\ldots,N)$, where u_N will satisfy the (original) p.d.e. Upon solving the overdetermined system, it was found, for those equations considered, that φ satisfied an equation formulated in terms of the Schwarzian derivative:

$$\{\varphi;x\}=\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right)-\frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2. \tag{1.4}$$

The invariance of (1.4) under the Moebius group

$$\varphi=\frac{a\psi+b}{c\psi+d},\quad \{\varphi;x\}=\{\psi;x\} \tag{1.5}$$

motivates the substitution

$$\varphi=v_1/v_2, \tag{1.6}$$

by which the Lax pairs may be found.[3]

Investigation of a certain class of equations formulated in terms of the Schwarzian derivatives revealed that these equations have the Painlevé property about movable, singularity manifolds of order -1. However, the occurrence of an additional type of movable singularity prevents this class of equations from identically possessing the Painlevé property. Hence, nonintegrable behavior can arise.[2]

In this paper a restriction (symmetry) is imposed that allows one to conclude that, when an equation is formulated in terms of the Schwarzian derivative and has this "symmetry," the equation identically possesses the Painlevé property. Within this class of equations are found the KdV, Caudrey–Dodd–Gibbon and Kuperschmidt equations. Furthermore, the "symmetry" property and invariance under the Moebius group allow effective Bäcklund transforms to be defined for these equations. In particular, rational or algebraic [in (x,t)] solutions can be generated iteratively.

In the next section, the Painlevé property and Bäcklund transformation for the KdV equation are reviewed for later reference.

In Sec. 3 the Painlevé property and Bäcklund transforms for the Caudrey–Dodd–Gibbon equation are presented. From these considerations the Kuperschmidt equation is found. The transformation between the Caudrey–Dodd–Gibbon and Kuperschmidt equations can be regarded as a

[a)] This work supported by Department of Energy Contract DOE DE-AC03-81ER10923 and AFOSR Grant No. AFOSR 83-0095.

 0022-2488/84/010013-12$02.50

certain "symmetry" under which these equations are "dual."

In Sec. 4, the "symmetry" discovered in Sec. 3 is employed to define a class of p.d.e.'s that possess the Painlevé property. The KdV equation is shown to be contained in this class of equations and self-dual w.r.t. this symmetry. Then, the sequences of higher order KdV, Caudrey–Dodd–Gibbon, and Kuperschmidt equations are found to be within this identically Painlevé class of equations and Bäcklund transformations are obtained for these sequences of equations.

In Sec. 5 rational [in (x,t)] solutions are constructed for several equations. In Appendix A the Lax pair for the Caudrey–Dodd–Gibbon equation is derived. In Appendix B further considerations relating to the seventh-order equations are presented.

2. THE KORTEWEG–DE VRIES EQUATION

The KdV equation

$$u_t + uu_x + u_{xxx} = 0 \tag{2.1}$$

possesses the Painlevé property.[1] The expansion about the singularity manifold has the form

$$u = \varphi^{-2} \sum_{j=0}^{\infty} u_j \varphi^j. \tag{2.2}$$

The "resonances" occur at

$$j = -1,4,6, \tag{2.3}$$

and (φ,u_4,u_6) are arbitrary functions of (x,t) in the expansion (2.2). We now assume the following "Bäcklund" transformation:

$$u = u_0/\varphi^2 + u_1/\varphi + u_2 \tag{2.4}$$

and find the following overdetermined system of equations,

$$\begin{aligned}
&\text{(i)} && u_0 = -12\varphi_x^2,\\
&\text{(ii)} && u_1 = 12\varphi_{xx},\\
&\text{(iii)} && \varphi_x\varphi_t + \varphi_x^2 u_2 + 4\varphi_x\varphi_{xxx} - 3\varphi_{xx}^2 = 0,\\
&\text{(iv)} && \varphi_{xt} + \varphi_{xx}u_2 + \varphi_{xxxx} = 0,\\
&\text{(v)} && u_{2t} + u_2u_{2x} + u_{2xxx} = 0,
\end{aligned} \tag{2.5}$$

$$u = 12\frac{\partial^2}{\partial x^2}\ln\varphi + u_2 \tag{2.6}$$

and, by eliminating u_2 in (2.5 iii,iv),

$$\varphi_t/\varphi_x + \{\varphi;x\} = \lambda, \tag{2.7}$$

where

$$\{\varphi;x\} = \frac{\partial}{\partial_x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\frac{\varphi_{xx}^2}{\varphi_x^2} \tag{2.8}$$

is the Schwarzian derivative of φ. Equation (2.7) is invariant under the Moebius group:

$$\varphi = \frac{a\psi + b}{c\psi + d}, \qquad \{\varphi;x\} = \{\psi;x\}. \tag{2.9}$$

The substitution[2]

$$\varphi = v_1/v_2, \quad \text{where } (v_1,v_2)\text{satisfy} \tag{2.10}$$

$$v_{xx} = av, \quad v_t = bv_x + cv, \tag{2.11}$$

readily obtains the Lax pair:

$$\begin{aligned}
a &= -\tfrac{1}{6}(u_2 + \lambda),\\
b &= -u_2/3 + \tfrac{2}{3}\lambda,\\
c &= u_x/6.
\end{aligned} \tag{2.12}$$

As noted in Ref. 2, Eq. (2.7) has an expansion

$$\varphi = \psi^{-1}\sum_{j=0}^{\infty}\varphi_j\psi^j \tag{2.13}$$

about a singularity manifold

$$\psi(x,t) = 0. \tag{2.14}$$

The resonances occur at

$$j = -1,0,1 \tag{2.15}$$

and the compatability conditions at $j = 0$ and 1 are satisfied identically. Thus, Eq. (2.7) has the Painlevé property about singularities of the form (2.13). However, we note that the vanishing of φ_x in (2.7) introduces the possibility of new, movable, singularities. This point will be resolved in Sec. 4.

The most general form of the Bäcklund transform defined by the expression

$$\varphi = \varphi_0/\psi + \varphi_1 \tag{2.16}$$

can be shown to be equivalent to the Moebius transformation (2.9). Again, an "effective" Bäcklund transformation for equation (2.7) will be defined in Sec. 4.

3. THE CAUDREY–DODD–GIBBON EQUATION

The Caudrey–Dodd–Gibbon equation[4,5]

$$u_t + \frac{\partial}{\partial x}(u_{xxxx} + 30uu_{xx} + 60u^3) = 0 \tag{3.1}$$

possesses the Painlevé property. The expansion about the singularity manifold is of the form

$$u = \varphi^{-2}\sum_{j=0}^{\infty}u_j\varphi^j. \tag{3.2}$$

There are found to be two solution branches.

Branch i: $u_0 = -\varphi_x^2$: The resonances occur at

$$j = -1,2,3,6,10. \tag{3.3}$$

Branch ii: $u_0 = -2\varphi_x^2$: The resonances occur at

$$j = -2,-1,5,6,12. \tag{3.4}$$

Both branches of the solution possess the Painlevé property.

The Bäcklund transformation defined for the "branch i" form of the solution is

$$u = u_0/\varphi^2 + u_1/\varphi + u_2. \tag{3.5}$$

The resulting overdetermined system of equations for (φ,u_0,u_1,u_2) is found to be

$$\begin{aligned}
&\text{(i)} && u_0 = -\varphi_x^2,\\
&\text{(ii)} && u_1 = \varphi_{xx},
\end{aligned} \tag{3.6}$$

$$\text{(iii)}\quad \frac{\varphi_t}{\varphi_x} + 6\frac{\varphi_{xxxxx}}{\varphi_z} - 15\frac{\varphi_{xx}\varphi_{xxxx}}{\varphi_x^2} + 10\frac{\varphi_{xxx}^2}{\varphi_x^2} + 30\left\{u_{2xx} + 4\left(\frac{\varphi_{xxx}}{\varphi_x} - 3\frac{\varphi_{xx}^2}{\varphi_x^2}\right)u_2 + 6u_2^2\right\} = 0, \tag{3.7}$$

(iv) $$\frac{\varphi_{xt}}{\varphi_x}+\frac{\varphi_{xxxxxx}}{\varphi_x}+30\left\{\frac{\varphi_{xx}}{\varphi_x}u_{2xx}+\frac{\varphi_{xxxx}}{\varphi_x}u_2+60\frac{\varphi_{xx}}{\varphi_x}u_2^2\right\}=0, \tag{3.8}$$

(v) $$u_{2t}+\frac{\partial}{\partial x}(u_{2xxxx}+30u_2u_{2xx}+60u_2^3)=0.$$

Using (3.6), Eq. (3.5) is

$$u=\frac{\partial^2}{\partial x^2}\ln\varphi+u_2. \tag{3.9}$$

We note that if

$$\varphi=1/\psi, \tag{3.10}$$

then

$$u_2=\frac{\partial^2}{\partial x^2}\ln\psi+u \tag{3.11}$$

and

$$W=u_2+\frac{1}{4}\frac{\varphi_{xx}^2}{\varphi_x^2}=u+\frac{1}{4}\frac{\psi_{xx}^2}{\psi_x^2}. \tag{3.12}$$

To employ this invariance, we let

$$u_2=W-\frac{1}{4}\frac{\varphi_{xx}^2}{\varphi_x^2} \tag{3.13}$$

and find

(i) $$\frac{\varphi_t}{\varphi_x}+6\frac{\partial^2}{\partial x^2}\{\varphi;x\}+19\{\varphi;x\}^2+30[W_{xx}+6W^2+4\{\varphi;x\}W]=0, \tag{3.14}$$

(ii) $$\frac{\partial}{\partial x}\left(\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+\frac{13}{2}\{\varphi;x\}^2\right)+30W\frac{\partial}{\partial x}\{\varphi;x\}=0, \tag{3.15}$$

where $\{\varphi;x\}$ is the Schwarzian derivative. To simplify these expressions, we let

$$\vartheta=\{\varphi;x\}+6W \tag{3.16}$$

and find

(i) $$\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2+5(\vartheta_{xx}+\vartheta^2+2\{\varphi;x\}\vartheta)=0, \tag{3.17}$$

(ii) $$\frac{\partial}{\partial x}\left(\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2\right)+5\vartheta\frac{\partial}{\partial x}\{\varphi;x\}=0. \tag{3.18}$$

From the consistency of (3.17) and (3.18)

$$\vartheta\vartheta_{xx}-\frac{\vartheta_x^2}{2}+\frac{2}{3}\vartheta^3+\{\varphi;x\}\vartheta^2=C. \tag{3.19}$$

Herein, we shall consider only the trivial solution

$$\vartheta=C=0, \tag{3.20}$$

$$\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2=0, \tag{3.21}$$

$$u_2=-\frac{1}{6}\frac{\varphi_{xxx}}{\varphi_x}. \tag{3.22}$$

It can be shown that Eqs. (3.21) and (3.22) imply that u_2 satisfies the Caudrey–Dodd–Gibbon equation. Actually, as is explained in Appendix A, (3.21) and (3.22) constitute a Lax pair for the Caudrey–Dodd–Gibbon equation.

We now let

$$\varphi=v_1/v_2, \quad \text{where } (v_1,v_2) \text{ satisfy} \tag{3.23}$$

$$v_{xx}=-\tfrac{1}{2}av, \quad v_t=bv_x+cv. \tag{3.24}$$

Equations (3.21), (3.23), and (3.24) imply that

$$a_t+\frac{\partial}{\partial x}\left(a_{xxxx}+\frac{45}{2}a_x^2+30aa_{xx}+60a^3\right)=0. \tag{3.25}$$

Equation (3.25) is known as the Kuperschmidt equation.[6] Analysis reveals that it possesses the Painlevé property. The expansion about the singularity manifold is of the form

$$a=\psi^{-2}\sum_{j=0}^{\infty}a_j\psi^j. \tag{3.26}$$

Again, there are two branches.

Branch i: $a_0=-\psi_x^2/2$: The resonances occur at

$$j=-1,3,5,6,7. \tag{3.27}$$

Branch ii: $a_0=-4\psi_x^2$: The resonances occur at

$$j=-7,-1,6,10,12. \tag{3.28}$$

We define the Bäcklund transformation about branch i:

$$a=a_0/\psi^2+a_1/\psi+a_2 \tag{3.29}$$

and find that

$$a_0=-\frac{\psi_x^2}{2}, \quad a_1=\frac{\psi_{xx}}{2}, \tag{3.30}$$

$$a_2=-\frac{1}{6}\{\psi;x\}-\frac{1}{8}\frac{\psi_{xx}^2}{\psi_x^2}, \tag{3.31}$$

$$\frac{\psi_t}{\psi_x}+\frac{\partial^2}{\partial x^2}\{\psi;x\}+\frac{1}{4}\{\psi;x\}^2=0. \tag{3.32}$$

We note that on account of the resonance structure, (3.27), (3.29)–(3.32) is *not* an overdetermined system.

Letting

$$\psi=W_1/W_2, \quad \text{where } (W_1,W_2) \text{ satisfy} \tag{3.33}$$

$$W_{xx}=-6uW, \quad W_t=bW_x+cW, \tag{3.34}$$

it is found that u satisfies Eq. (3.1).

Furthermore, if

$$v=\varphi_{xx}/\varphi_x=-\tfrac{1}{2}\psi_{xx}/\psi_x, \tag{3.35}$$

where φ satisfies Eq. (3.21) and ψ satisfies Eq. (3.32), then

$$v_t+\frac{\partial}{\partial x}(v_{xxxx}+5v_xv_{xx}-5v^2v_{xx}-5vv_x^2+v^5)=0. \tag{3.36}$$

The above implies the nonlinear transformation found in Ref. 6. For our purposes we note that (3.35) provides the transformation:

$$\psi_x=\varphi_x^{-2}. \tag{3.37}$$

Equation (3.37) indicates that Eqs. (3.21) and (3.32)

identically possess the Painlevé property. Each equation has the Painlevé property about "poles" or order -1,[2] and, about the possible movable singularities (where $\psi_x = 0$ or $\varphi_x = 0$), the transformation (3.37) provides the appropriate representation of the solution. For instance, (3.37) refers the behavior of φ at points where $\varphi_x = 0$ to the expansion of ψ at points where $\psi_x \to \infty$ (the poles of ψ). And, as is explained in the next section, this allows us to conclude that φ is single-valued at these points.

4. AN INTEGRABLE CLASS OF PARTIAL DIFFERENTIAL EQUATIONS

An equation

$$\varphi_t/\varphi_x + B(\{\varphi;x\}) = 0, \tag{4.1}$$

where $B(\{\varphi;x\})$ is a constant coefficient multinomial in $(\partial^j/\partial x^j)\{\varphi;x\}$, will identically possess the Painlevé property when there exists a transformation

$$\varphi_x = \psi_x^m, \tag{4.2}$$

where m is rational and negative and ψ satisfies an equation of the form (4.1). The form of Eq. (4.1) is sufficient to guarantee the existence of "meromorphic" expansions about the "poles" of order -1. That is,

$$\varphi = \vartheta^{-1} \sum_{j=0}^{\infty} \varphi_j \vartheta^j, \tag{4.3}$$

where the resonances occur at $j = -1, 0, 1, \ldots, n+1$ and n is the order of the highest derivative (of the Schwarzian) appearing in B. The transformation (4.2) provides a representation of the solution in a neighborhood of the points where $\varphi_x = 0$ ($\psi_x = 0$) by associating these points with the behavior of solutions of the "dual" equation in a neighborhood of their singularities.

To see the validity of the expansion (4.3), we observe that for singularities of the form (4.3) the expansion for the Schwarzian derivative begins at the constant level (is nonsingular). And, consequently, the (n) derivatives of the Schwarzian merely "shift" the recursion relations to the appropriate higher coefficient, φ_{n+2}, adding one resonance for each derivative. For particular equations of the form (4.1) higher order poles (φ^{-m}) can occur. We shall find that these singularities can be "reduced" to (4.3) through the invariance of (4.1) under the "symmetry" (4.2) and the Moebius group.

Consider forms of $B(\{\varphi;x\})$ that are linear in the highest order derivative of the Schwarzian and order the terms defining $B(\{\varphi;x\})$ into expressions that are homogeneous of the same degree under the change of variable

$$x \to a^{-1}x, \tag{4.4}$$

$$\{\varphi;x\} = \frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\left(\frac{\varphi_{xx}^2}{\varphi_x^2}\right) \Rightarrow a^2\{\varphi;x\}. \tag{4.5}$$

These are

$$\text{(i)}\quad \{\varphi;x\},$$

$$\text{(ii)}\quad \frac{\partial}{\partial x}\{\varphi;x\},$$

$$\text{(iii)}\quad \frac{\partial^2}{\partial x^2}\{\varphi;x\} + \lambda\{\varphi;x\}^2, \tag{4.6}$$

$$\text{(iv)}\quad \frac{\partial^3}{\partial x^3}\{\varphi;x\} + \lambda\{\varphi;x\}\frac{\partial}{\partial x}\{\varphi;x\},$$

$$\text{(v)}\quad \frac{\partial^4}{\partial x^4}\{\varphi;x\} + \alpha\{\varphi;x\}\frac{\partial^2}{\partial x^2}\{\varphi;x\} + \beta\left(\frac{\partial}{\partial x}\{\varphi;x\}\right)^2 + \lambda\{\varphi;x\}^3,$$

etc.

We consider equations (4.6i,ii,iii,v). Therefore, let

$$\frac{\varphi_t}{\varphi_x} + \{\varphi;x\} = \lambda \tag{4.7}$$

and

$$\varphi_x = \psi_x^m. \tag{4.8}$$

Then

$$\{\varphi;x\} = m\frac{\psi_{xxx}}{\psi_x} - \left(\frac{m^2}{2} + m\right)\frac{\psi_{xx}^2}{\psi_x^2} \tag{4.9}$$

and

$$m\psi_x^{m-1}\psi_{xt} + \frac{\partial}{\partial x}\psi_x^m\left(m\frac{\psi_{xxx}}{\psi_x} - \left(\frac{m^2}{2} + m\right)\frac{\psi_{xx}^2}{\psi_x^2}\right) = 0. \tag{4.10}$$

Direct calculation obtains

$$m\frac{\partial}{\partial x}\left(\psi_t + \psi_{xxx} - \frac{3\psi_{xx}^2}{2\psi_x} - \lambda\psi_x\right) + \left(2m - \frac{m^3}{2} - \frac{3m}{2}\right)\frac{\psi_{xx}^3}{\psi_x^2} = 0. \tag{4.11}$$

For Eq. (4.11) to be of the form (4.1),

$$2m - m^3/2 - 3m/2 = 0 \tag{4.12}$$

or

$$m = 0, \pm 1. \tag{4.13}$$

Then, if

$$\varphi_x = \psi_x^{-1}, \tag{4.14}$$

ψ will satisfy

$$\psi_t/\psi_x + \{\psi;x\} = \lambda, \tag{4.15}$$

assuming the constant of integration introduced in expression (4.11) is to vanish. For instance, we can assume that all solutions approach time-independent constants when x approaches $-\infty$.

Thus, Eqs. (4.11), (4.14), and (4.15) define a Bäcklund transformation that will be employed, with the invariance under the Moebius group, in Sec. 5 to generate rational solutions. Equation (4.7) is directly related to the KdV equation (Sec. 2).

Next, it can be readily shown that the equation

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial}{\partial x}\{\varphi;x\} = 0 \tag{4.16}$$

does not have a transformation

$$\varphi_x = \psi_x^m$$

that remains within the class (4.1). This equation, studied in Ref. 3, is transformable to an equation with complex reson-

ances (self-similar natural boundary)[7] and is thought to be nonintegrable.

It is useful to observe that, by Eq. (4.9), a transformation of type (4.2) does not change the degree of homogeneity (4.4) of the expressions in (4.6). Thus, if a transformation exists, it can only effect the value of the coefficients in (4.6).

Equations of the form

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial^2}{\partial x^2}\{\varphi;x\} + \lambda\{\varphi;x\}^2 = 0 \tag{4.17}$$

have a transformation

$$\varphi_x = \psi_x^m \tag{4.18}$$

that preserves the formulation (4.1) when

$$\begin{aligned} &\text{(i)} \quad m = -1, \quad \lambda = \tfrac{3}{2}, \\ &\text{(ii)} \quad m = -2, \quad \lambda = \tfrac{1}{4}, \\ &\text{(iii)} \quad m = -\tfrac{1}{2}, \quad \lambda = 4. \end{aligned} \tag{4.19}$$

Equation (4.19i) is (essentially) the first higher-order (fifth degree) KdV equation.[2] Equation (4.19i,ii) are (obtained from) the Kuperschmidt and Caudrey–Dodd–Gibbon equations, respectively (see Sec. 3). Then, the Kuperschmidt equation and Caudrey–Dodd–Gibbon equation are, in a sense, "dual" under the transformation

$$\psi_x = \varphi_x^{-2}. \tag{4.20}$$

The KdV equation (4.7) and fifth-degree higher-order KdV equation (4.19i) are then "self-dual."

We note that the property of possessing a transformation within the class (4.1) is additive (by construction) for expressions with the same value of exponent m.

Thus, by (4.19i) and (4.14) the equation

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial^2}{\partial x^2}\{\varphi;x\} + \frac{3}{2}\{\varphi;x\}^2 + \lambda\{\varphi;x\} = 0 \tag{4.21}$$

has, for any λ, an (auto) Bäcklund transform

$$\varphi_x = \psi_x^{-1}. \tag{4.22}$$

Finally, the equation

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial^4}{\partial x^4}\{\varphi;x\} + \alpha\{\varphi;x\}\frac{\partial^2}{\partial x^2}\{\varphi;x\} + \beta\left(\frac{\partial}{\partial x}\{\varphi;x\}\right)^2 + \lambda\{\varphi;x\}^3 \tag{4.23}$$

has a transformation

$$\varphi_x = \psi_x^m \tag{4.24}$$

preserving the form of Eq. (4.23) when

$$\begin{aligned} &\text{(i)} \quad m = -1, \quad \alpha = 5, \quad \beta = \tfrac{5}{2}, \quad \lambda = \tfrac{5}{2}, \\ &\text{(ii)} \quad m = -2, \quad \alpha = \tfrac{3}{2}, \quad \beta = \tfrac{3}{4}, \quad \lambda = \tfrac{1}{6}, \\ &\text{(iii)} \quad m = -\tfrac{1}{2}, \quad \alpha = 12, \quad \beta = 6, \quad \lambda = \tfrac{32}{5}. \end{aligned} \tag{4.25}$$

These are higher order KdV, Kuperschmidt, and Caudrey–Dodd–Gibbon equations, respectively. Further information concerning Eq. (4.23) is contained in Appendix B.

We now consider the sequence of higher-order KdV equations determined by the "Lenard recursion relation"[8]

$$\frac{\partial}{\partial x} b^{n+1} = b^n_{xxx} + 2ub^n_x + u_x b^n, \tag{4.26}$$

where

$$u_t + \frac{\partial}{\partial x} b^{n+1}(u) = 0 \tag{4.27}$$

for $n = 1,2,3,\ldots$ are the sequence of higher-order KdV equations and

$$\begin{aligned} b^0 &= 1, \\ b^1 &= u, \\ b^2 &= u_{xx} + \tfrac{3}{2}u^2, \\ b^3 &= u_{xxxx} + 5uu_{xx} + \tfrac{5}{2}u_x^2 + \tfrac{5}{2}u^3. \end{aligned} \tag{4.28}$$

Now inspection of Eqs. (4.7), (4.17), and (4.23) leads us to formulate the following.

Theorem 1: The sequence of higher-order KdV equations

$$u_t + \frac{\partial}{\partial x} b^{n+2}(u) = 0 \tag{4.29}$$

for $n = 0,1,2,\ldots$ has the following Bäcklund transformation:

$$u = 4\frac{\partial^2}{\partial x^2}\ln\varphi + u_2, \tag{4.30}$$

$$u_2 = -\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2, \tag{4.31}$$

$$\frac{\varphi_t}{\varphi_x} + b^{n+1}(\{\varphi;x\}) = 0. \tag{4.32}$$

Furthermore,

$$\omega = \{\varphi;x\} \tag{4.33}$$

(and u_2) satisfies Eqs. (4.29) and (4.32) is invariant under the transformation

$$\varphi_x = \psi_x^{-1}. \tag{4.34}$$

Note: To simplify the statement of the above results, we require the sequence of b^n to be defined by precisely Eq. (4.26). "Scalings" in the argument "u" of Eq. (4.26) is essential for the definition of Eq. (4.32), but not for Eq. (4.29).

Proof: We prove the above by the following observations: For each n, let

$$V = \varphi_{xx}/\varphi_x. \tag{4.35}$$

Then Eq. (4.32) obtains the "higher-order modified KdV equation"

$$V_t + \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x} + V\right) b^{n+1}(V_x - \tfrac{1}{2}V^2) = 0, \tag{4.36}$$

where

$$\omega = \{\varphi;x\} = V_x - \tfrac{1}{2}V^2. \tag{4.37}$$

From Eqs. (4.36) and (4.37) we find that

$$\omega_t + \left(\frac{\partial^3}{\partial x^3} + 2\omega\frac{\partial}{\partial x} + \omega_x\right) b^{n+1}(\omega) = 0, \tag{4.38}$$

or, using Eq. (4.26),

$$\omega_t + \frac{\partial}{\partial x} b^{n+2}(\omega) = 0. \tag{4.39}$$

This equation (4.32) implies that ω is a solution of Eq. (4.29). From Eqs. (4.31) and (4.35)

$$u_2 = -V_x - \tfrac{1}{2}V^2. \tag{4.40}$$

Now, if

$$\varphi_x \to \varphi_x^{-1}, \tag{4.41}$$

then

$$V \to -V, \quad u_2 \to \omega, \tag{4.42}$$

$$w \to u_2. \tag{4.43}$$

Hence, both u_2 and ω will be solutions of Eq. (4.29) if Eq. (4.32) is invariant under (4.41), or equivalently, if Eq. (4.36) is invariant under (4.42).

To see this, we let

$$D = \frac{\partial}{\partial x}, \tag{4.44}$$

$$M_v = D(D+V), \tag{4.45}$$

$$L_v = D^{-1}(D-V)M_v, \tag{4.46}$$

and find that the Lenard relationship (4.26) becomes

$$b^{n+2}(V_x - \tfrac{1}{2}V^2) = L_v b^{n+1}(V_x - \tfrac{1}{2}V^2) \tag{4.47}$$

while Eq. (4.36) is

$$V_t + M_v b^{n+1}(V_x - \tfrac{1}{2}V^2) = 0. \tag{4.48}$$

The condition of invariance of (4.48) under (4.42) reads

$$M_v b^{n+1}(V_x - \tfrac{1}{2}V^2) + M_{-v} b^{n+1}(-V_x - \tfrac{1}{2}V^2) = 0. \tag{4.49}$$

We verify (4.49) by induction. Previous calculations demonstrate (4.49) for $n = 0,1$. We assume (4.49) with $n = 0,1,2,\ldots,m-1$. Then with $n = m$ and, using (4.47), Eq. (4.49) is

$$M_v L_v b^m(V_x - \tfrac{1}{2}V^2) + M_{-v} L_{-v} b^m(-V_x - \tfrac{1}{2}V^2) = 0. \tag{4.50}$$

However, from (4.46),

$$M_v L_v = I_v M_v, \tag{4.51}$$

where

$$I_v = D(D+V)D^{-1}(D-V). \tag{4.52}$$

Using the identity for constants a,b,

$$(D+aV)D^{-1}(D+bV) = (D+bV)D^{-1}(D+aV), \tag{4.53}$$

it is found that

$$I_v = I_{-v} \tag{4.54}$$

and, with (4.51), Eq. (4.50) is

$$I_v\{M_v b^m(V_x - \tfrac{1}{2}V^2) + M_{-v} L_{-v} b^m(-V_x - \tfrac{1}{2}V^2)\} = 0. \tag{4.55}$$

Since the term in brackets vanishes by assumption, (4.49) is verified for $n = m$. We note that (4.52) is a recursion operator for the higher-order modified KdV equations.

Equation (4.32) and the invariance (4.34) obtain that (ω,u_2) are solutions of Eq. (4.29). We now show that Eqs. (4.32), (4.31), and (4.30) imply that u [defined in Eq. (4.30)] will be a solution of Eq. (4.29), completing the proof of the existence of the Bäcklund transform.

To begin, we note that Eq. (4.32) is invariant under the Moebius group.

Letting

$$\varphi = 1/\psi, \tag{4.56}$$

we find that ψ satisfies Eq. (4.32) and that

$$u_2 = 4\frac{\partial^2}{\partial x^2}\ln\psi + u, \tag{4.57}$$

$$u_2 = -\frac{\partial}{\partial x}\left(\frac{\psi_{xx}}{\psi_x}\right) - \frac{1}{2}\left(\frac{\psi_{xx}}{\psi_x}\right)^2 + 4\frac{\partial^2}{\partial x^2}\ln\psi, \tag{4.58}$$

or

$$u = -\frac{\partial}{\partial x}\left(\frac{\psi_{xx}}{\psi_x}\right) - \frac{1}{2}\left(\frac{\psi_{xx}}{\psi_x}\right)^2. \tag{4.59}$$

By the previous calculation Eq. (4.59) implies u satisfies Eq. (4.29), completing the proof.

Remark 1: Equation (4.32) effectively defines three distinct solutions of Eq. (4.29). That is,

$$u_2,\omega = \{\varphi;x\}$$

and

$$u = 4\frac{\partial^2}{\partial x^2}\ln\varphi + u_2. \tag{4.60}$$

Remark 2: If we consider the stationary solutions of a higher-order KdV equation[8] Theorem 1 defines Bäcklund transformations for the associated ordinary differential equations. Furthermore, to construct solutions of the $(n+2)$ equation

$$\frac{\partial}{\partial x} b^{n+2}(u) = 0, \tag{4.61}$$

we integrate the $(n+1)$ equation

$$b^{n+1}(\omega) = 0 \tag{4.62}$$

and set

$$\omega = \{\varphi;x\}. \tag{4.63}$$

Then

$$\varphi = V_1/V_2, \tag{4.64}$$

where V_1 and V_2 satisfy

$$V_{xx} = -\tfrac{1}{2}wV, \tag{4.65}$$

defines the solutions (u,u_2) of (4.61).

Thus the solution of the $(n+1)$ equation is the "potential" in a associated linear, Schrödinger equation, that defines the solutions and Bäcklund transforms for the $(n+2)$ equation. Further consideration of these Bäcklund transforms for (Painlevé) ODE's and the iterative construction of solutions seems warranted.

We now generalize Theorem 1 to allow for the inclusion of a spectral parameter, λ.

Theorem 2: The sequence of higher-order KdV equations

$$u_t + \frac{\partial}{\partial x} b^{n+2}(u) = 0 \tag{4.66}$$

for $n = 0,1,2,\cdots$ has the Bäcklund transformation

$$u = 4\frac{\partial^2}{\partial x^2}\ln\varphi + u_2, \tag{4.67}$$

$$u_2 = -\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2 + \lambda, \tag{4.68}$$

$$\frac{\varphi_t}{\varphi_x} + \alpha_{n+1,j}\, b^j(\{\varphi;x\}) = 0, \tag{4.69}$$

where $\alpha_{n+1,j} = \alpha_{n+1,j}(\lambda)$, with a summation convention over $j = 0,1,\ldots,n+1$.

Furthermore,

$$w = \{\varphi;x\} + \lambda \tag{4.70}$$

satisfies equation (4.66) and equation (4.69) is invariant under the transform

$$\varphi_x = \psi_x^{-1}. \tag{4.71}$$

Proof: By a previous remark invariance (4.71) follows immediately from (4.34). Now, let

$$V = \varphi_{xx}/\varphi_x. \tag{4.72}$$

Then

$$V_t + \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x} + V\right)\alpha_{n+1,j}b^j\left(V_x - \frac{1}{2}V^2\right) \tag{4.73}$$

and

$$\omega_t + \left(\frac{\partial^3}{\partial x^3} + 2(\omega-\lambda)\frac{\partial}{\partial x} + \omega_x\right)\alpha_{n+1,j}b^j(\omega-\lambda) = 0. \tag{4.74}$$

By (4.26)

$$b_x^{j+1}(\omega-\lambda) = \left(\frac{\partial^3}{\partial x^3} + 2(\omega-\lambda)\frac{\partial}{\partial x} + \omega_x\right)b^j(\omega-\lambda). \tag{4.75}$$

Lemma 1:

$$b^j(\omega-\lambda) = \sum_{k=0}^{j} a_{jk}b^k(\omega),$$

where

$$a_{jj} = 1,$$

$$a_{j0} = -\lambda\,((2j-1)/j)a_{j-1,0}$$

and

$$a_{j,k} = a_{j-1,k-1} - 2\lambda a_{j-1,k}, \quad \text{where } k < j.$$

Proof: By induction, using (4.26).

Now using Eqs. (4.70), (4.71), Lemma 1, and requiring that ω satisfy Eq. (4.66) determines, for each n, the $\alpha_{n+1,j}$, $j = 0,1,\ldots,n+1$.

We find the following triangular system of linear equations for

$$\alpha_{n+1} = \begin{pmatrix} \alpha_{n+1,n+1} \\ \alpha_{n+1,n} \\ \alpha_{n+1,n-1} \\ \vdots \\ \alpha_{n+1,0} \end{pmatrix}, \tag{4.76}$$

$$\begin{pmatrix} 1 & 0 & & 0 \\ a_{n+2,m+1} & 1 & & \\ a_{n+2,m}\ a_{m+1,m} & & 1 & \\ \vdots & & \ddots & \\ a_{m+2,k}\ a_{m+1,k} & & & 1\ \ 0 \\ a_{m+2,1}\ a_{n+1,1} & & & 1 \end{pmatrix} a_{n+1} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.77}$$

Since the system (4.77) is always solvable, α_{n+1} exists for each n, and the Bäcklund transformations are well defined, completing the proof.

For reference, we present the following tables:

	a_{jk}				
j/k	0	1	2	3	4
0	1				
1	$-\lambda$	1			
2	$\frac{3}{2}\lambda^2$	-3λ	1		
3	$-\frac{5}{2}\lambda^3$	$\frac{15}{2}\lambda^2$	-5λ	1	
4	$\frac{35}{8}\lambda^4$	$\frac{-35}{2}\lambda^3$	$\frac{35}{2}\lambda^2$	-7λ	1

	$\alpha_{j,k}$			
j/k	0	1	2	3
1	3λ	1		
2	$\frac{15}{2}\lambda^2$	5λ	1	
3	$\frac{35}{2}\lambda^3$	$\frac{35}{2}\lambda^2$	7λ	1

We next consider the sequence of higher-order Caudrey–Dodd–Gibbon and Kuperschmidt equations. Again, to avoid unnecessary complexity, we consider these equations with a specific scaling. With reference to Sec. 3, we let

$$u \to u/12, \qquad a \to a/3, \tag{4.78}$$

and find the Caudrey–Dodd–Gibbon equation

$$u_t + \frac{\partial}{\partial x}\left(u_{xxxx} + \frac{5}{2}uu_{xx} + \frac{5}{12}u^3\right) = 0 \tag{4.79}$$

and the Kuperschmidt equation

$$a_t + \frac{\partial}{\partial x}\left(a_{xxxx} + 10aa_{xx} + \frac{15}{2}a_x^2 + \frac{20}{3}a^3\right) = 0. \tag{4.80}$$

From Ref. 9 the sequences of conserved covariants (functional gradients of conserved densities) are given by

$$G_{n+2} = J_1(u)\Theta_1(u)G_n, \tag{4.81}$$

$$H_{n+2} = J_2(a)\Theta_2(a)H_n \tag{4.82}$$

for the Caudrey–Dodd–Gibbon and Kuperschmidt equations, respectively, where

$$\Theta_1 = D^3 + 2uD + u_x,$$
$$J_1 = D^3 + \tfrac{1}{2}D^2uD^{-1} + \tfrac{1}{2}D^{-1}uD^2 + \tfrac{1}{8}(u^2D^{-1} + D^{-1}u^2), \tag{4.83}$$

and

$$\Theta_2 = D^3 + 2uD + u_x,$$
$$J_2 = D^3 + 3(uD + Du) + 2(D^2uD^{-1} + D^{-1}uD^2) + 8(u^2D^{-1} + D^{-1}u^2). \tag{4.84}$$

With the normalization that we employ,

$$G_0 = 1, \quad H_0 = 1,$$
$$G_1 = u_{xx} + \tfrac{1}{4}u^2, \quad H_1 = a_{xx} + 4a^2, \tag{4.85}$$

and Eqs. (4.79) and (4.80) are

$$u_t + \Theta_1 G_1(u) = 0,$$
$$a_t + \Theta_2 H_1(a) = 0. \tag{4.86}$$

Furthermore, the respective sequences of higher-order equations are given by

$$u_t + \Theta_1 G_n(u) = 0,$$
$$a_t + \Theta_2 H_n(a) = 0. \tag{4.87}$$

For what follows it is convenient, as was the case for the KdV equations, to "factorize" the recursion operators. That is,

$$\Theta_1 = (D - W)D(D + W), \tag{4.88}$$
$$J_1 = D^{-1}\{(D - W/2)(D + W/2) \times D(D - W/2)(D + W/2)\}D^{-1},$$

and

$$\Theta_2 = (D - V)D(D + V), \tag{4.89}$$
$$J_2 = D^{-1}\{(D - 2V)(D - V) \times D(D + V)(D + 2V)\}D^{-1},$$

where

$$u = W_x - \tfrac{1}{2}W^2,$$
$$a = V_x - \tfrac{1}{2}V^2. \tag{4.90}$$

We now formulate the following.

Theorem 3: The sequences of higher-order Caudrey–Dodd–Gibbon and Kuperschmidt equations

$$u_t + \Theta_1 G_n(u) = 0,$$
$$a_t + \Theta_2 H_n(a) = 0, \tag{4.91}$$

for $n = 1,2,3,\cdots$, have the following Bäcklund transformations:

$$u = 12\frac{\partial^2}{\partial x^2}\ln\varphi + u_2,$$
$$a = \frac{3}{2}\frac{\partial^2}{\partial x^2}\ln\psi + a^2, \tag{4.92}$$

where

$$u_2 = -2\frac{\varphi_{xxx}}{\varphi_x},$$
$$a_2 = -\frac{1}{2}\left(\frac{\psi_{xxx}}{\psi_x} - \frac{3}{4}\frac{\psi_{xx}^2}{\psi_x^2}\right) \tag{4.93}$$

and

$$\frac{\varphi_t}{\varphi_x} + H_n(\{\varphi;x\}) = 0,$$
$$\frac{\psi_t}{\psi_x} + G_n(\{\psi;x\}) = 0. \tag{4.94}$$

Furthermore, Eqs. (4.94) possess the symmetry

$$\psi_x = \varphi_x^{-2}, \tag{4.95}$$

and

$$u_3 = \{\psi;x\},$$
$$a_3 = \{\phi;x\} \tag{4.96}$$

are solutions of Eq. (4.91), respectively.

Proof: (i) The sequences of higher-order modified Caudrey–Dodd–Gibbon and Kuperschmidt equations are given by

$$W_t + M_w G_n(W_x - \tfrac{1}{2}W^2) = 0,$$
$$V_t + M_v H_n(V_x - \tfrac{1}{2}V^2) = 0, \tag{4.97}$$

respectively, where

$$W = \psi_{xx}/\psi_x,$$
$$V = \varphi_{xx}/\varphi_x, \tag{4.98}$$

and

$$M_v = D(D + V). \tag{4.99}$$

Since

$$u_3 = W_x - \tfrac{1}{2}W^2,$$
$$a_3 = V_x - \tfrac{1}{2}V^2, \tag{4.100}$$

the factorizations (4.88) and (4.89) show

$$u_{3t} + \Theta_1 G_n(u_3) = 0,$$
$$a_{3t} + \Theta_2 H_n(a_3) = 0. \tag{4.101}$$

(ii) Now *if* (4.95) is valid, then, as is readily verified,

$$u_2 = \{\psi;x\},$$
$$a_2 = \{\varphi;x\}, \tag{4.102}$$

and by the above (u_2,a_2) solve Eqs. (4.91). Now, the invariance of Eqs. (4.94) under the Moebius group, (4.92) and (4.93) imply

$$u = -2\tilde{\varphi}_{xxx}/\tilde{\varphi}_x,$$
$$a = -\tfrac{1}{2}(\tilde{\psi}_{xxx}/\tilde{\psi}_x - \tfrac{3}{4}\tilde{\psi}_{xx}^2/\tilde{\psi}_x^2), \tag{4.103}$$

where

$$\tilde{\varphi} = 1/\varphi, \quad \tilde{\psi} = 1/\psi \tag{4.104}$$

and $(\tilde{\varphi},\tilde{\psi})$ are solutions of (4.94). By the above, (u,a) are solutions of (4.91), and (4.92) is well defined if (4.95) is verified.

(iii) By (4.98), (4.95) is equivalent to the condition

$$W = -2V, \tag{4.105}$$

or, using (4.97), to

$$2M_v H_n(V_x - \tfrac{1}{2}V^2) + M_{-2v}G_n(-2V_x - 2V^2) = 0. \tag{4.106}$$

We verify (4.106) by induction. Previous calculations demonstrate (4.106) for $n = 1,2$. We assume (4.106) valid for $n = 1,2,\ldots,m$; then, by (4.81) and (4.82),

$$2M_v H_{m+1}(V_x - \tfrac{1}{2}V^2) + M_{-2v}G_{m+1}(-2V_x - 2V^2) = 2M_v J_2(a)\Theta_2(a)H_{m-1}(V_x - \tfrac{1}{2}V^2) + M_{-2v}J_1(\tilde{u})\Theta_1(\tilde{u})G_{m-1}(-2V_x - 2V^2), \tag{4.107}$$

where

$$a = V_x - \tfrac{1}{2}V^2, \quad \tilde{u} = -2V_x - 2V^2.$$

However, (4.88) and (4.89) readily obtain

$$M_v J_1 \Theta_1 = \lambda_v M_v, \quad M_v J_2 \Theta_2 = \Phi_v M_v, \tag{4.108}$$

where

$$\lambda_v = D(D+V)D^{-1}\{(D-V/2)(D+V/2) \times D(D-V/2)(D+V/2)\}D^{-1}(D-V) \tag{4.109}$$

and

$$\Phi_v = D(D+V)D^{-1}\{(D-2V)(D-V) \times D(D+V)(D+2V)\}D^{-1}(D-V). \tag{4.110}$$

The identity [by (4.53)]

$$\lambda_{-2v} = \Phi_v \tag{4.111}$$

and (4.106) for $n = m-1$ imply that (4.107) vanishes, verifying (4.106), (4.105) and completing the proof.

We note that, in another context, the method of factorization of operators has been used to derive Miura transformations and Hamiltonian structures.[10–12]

Remark 3: It is not known whether the sequences of KdV, Caudrey–Dodd–Gibbon, and Kuperschmidt equations exhaust the equations in the class (4.1). Presumably, there may exist a sequence of equations for every index pair, $(m, 1/m)$, $m = -1, -2, -3, \cdots$.

We conclude this section with some remarks concerning the nature of the higher-order poles for the class of equations considered herein. For instance, the sequence of KdV equations, (4.32), can have singularities of the form

$$\varphi = \varphi_0 \epsilon^{-N} + \varphi_1 \epsilon^{-N+1} + \cdots, \tag{4.112}$$

where it is not assumed that (4.112) is Painlevé.

For simplicity we employ the "reduced" expansion[1]

$$\epsilon = x - \psi(t), \quad \varphi_j = \varphi_j(t). \tag{4.113}$$

Now, since (4.32) is invariant under the Moebius group, the transformation

$$\psi = 1/\varphi \tag{4.114}$$

produces a solution which has an expansion

$$\psi = \psi_0 \epsilon^N + \psi_1 \epsilon^{N+1} + \cdots. \tag{4.115}$$

Furthermore, the symmetry

$$\varphi_x = \psi_x^{-1}$$

obtains

$$\varphi_x = \varphi_0 \epsilon^{-N+1} + \cdots, \tag{4.116}$$

$$\varphi = \varphi_0 \epsilon^{-N+2} + \cdots. \tag{4.117}$$

If N is an odd integer, after a finite number of steps, there results

$$\varphi = \varphi_0 \epsilon^{-1} + \cdots. \tag{4.118}$$

However, singularities of the form (4.118) identically possess the Painlevé property. Now $\ln \epsilon$ terms could arise in going from (4.112) to (4.118) [but do not, since (4.118) is Painlevé with the complete set of "arbitrary functions"]. However, no $\ln \epsilon$ terms can occur in going from (4.118) to (4.112). Thus, (4.112), as reconstructed from (4.118), has the Painlevé property (when N is odd). [Note in going from (4.118) to (4.112) Taylor, not Laurent, series are integrated.]

Let us now assume that

$$\varphi \approx \varphi_0 \epsilon^{m+1}. \tag{4.119}$$

Then,

$$\{\varphi; x\} \approx -\tfrac{1}{2}m(m+2)\epsilon^{-2}, \tag{4.120}$$

and using the Lenard formula (4.26) with

$$b^n(\{\varphi; x\}) \approx P^n(m)\epsilon^{-2n}$$

obtains

$$(2n+2)P^{n+1}(m) = 2(2n+1)(\lambda + n(n+2))P^n(m) \tag{4.121}$$

where $\lambda = -\tfrac{1}{2}m(m+2)$.

Thus, each higher-order equation of order $(n+1)$ acquires two new leading orders

$$\lambda = -\tfrac{1}{2}m(m+2) = -n(n+2) \tag{4.122}$$

or

$$m = 2n, -2-2n,$$

where

$$\varphi \approx \varphi_0 \epsilon^{2n+1} \tag{4.123}$$

or

$$\varphi \approx \varphi_0 \epsilon^{-2n-1}.$$

The higher-order KdV equations (in the Schwarzian formulation) can have only odd integral leading orders, and by the previous remarks these have the Painlevé property.

Considerations of a similar nature determine that the higher-order singularities of the Caudrey–Dodd–Gibbon and Kuperschmidt sequences, again, "reduce" to singularities of the (Painlevé) form (4.118). Thus, these equations identically possess the Painlevé property.

5. ITERATIVE CONSTRUCTION OF RATIONAL SOLUTIONS

For the KdV equation

$$u_t + \frac{\partial}{\partial x}\left(\frac{u^2}{2} + u_{xx}\right) = 0, \tag{5.1}$$

the Bäcklund transform

$$u = 12\frac{\partial^2}{\partial x^2}\ln\varphi + u_2 \tag{5.2}$$

implies that

$$\frac{\varphi_t}{\varphi_x} + \{\varphi; x\} = \lambda. \tag{5.3}$$

Equation (5.3) is invariant under:

(i) The Moebius group

$$\varphi = \frac{a\psi + b}{c\psi + d} \tag{5.4}$$

and the transformation

$$\text{(ii)} \quad \varphi_x = \psi_x^{-1}. \tag{5.5}$$

Combining Eq. (5.4), i.e.,

$$\psi = -1/\varphi, \tag{5.6}$$

and Eq. (5.5), there is defined the Bäcklund transformation

$$\varphi_{n+1,x} = \varphi_n^2/\varphi_{n,x}. \tag{5.7}$$

Without loss of generality (modulo a Galilean transformation) we set $\lambda = 0$ in Eq. (5.3). Then setting

$$\varphi_0 = x, \tag{5.8}$$

it is found from Eqs. (5.7) and (5.3) that

$$\varphi_1 = x^3/3 + 4t. \tag{5.9}$$

We normalize (5.9) by setting

$$\varphi_1 = x^3 + 12t. \tag{5.10}$$

From Eq. (5.7) it is found that (after normalization)

$$\varphi_2 = (x^6 + 60tx^3 + ex - 720t^2)/x \tag{5.11}$$

and

$$\varphi_3 = 1/\varphi_1[x^{10} + 180tx^7 + 302\,400t^3x + 7e(x^5 - 60tx^3 - e/3) + f\varphi_1], \tag{5.12}$$

where (e,f) are constants of integration.

Equations (5.8)–(5.12) suggest that

$$\varphi_n = P_n/P_{n-2}, \tag{5.13}$$

where the P_j are polynomials in (x,t). Substitution of (5.13) into (5.7) obtains

$$P_{n-1}P_{n+1,x} - P_{n-1,x}P_{n+1} = P_n^2, \tag{5.14}$$

where

$$\begin{aligned} P_0 &= x, \\ P_1 &= x^3 + 12t, \\ P_2 &= x^6 + 60tx^3 - 720t^2 + ex. \end{aligned} \tag{5.15}$$

The solutions obtained from (5.14) and (5.15) are (essentially) those rational solutions of the KdV equation found by Ablowitz and Segur,[13] using Hirota's method, and are equivalent to rational solutions of Airault, McKean and Moser.[14]

From Eqs. (5.13) and (5.2) we find that

$$\begin{aligned} u &= 12\frac{\partial^2}{\partial x^2}\ln P_n, \\ u_2 &= 12\frac{\partial^2}{\partial x^2}\ln P_{n-2} \end{aligned} \tag{5.16}$$

define rational solution of the KdV equations.

For the Caudrey–Dodd–Gibbon equation (3.1) and the Kuperschmidt equation (3.25), there are defined the following Bäcklund transformations:

$$u = \frac{\partial^2}{\partial x^2}\ln\varphi + u_2 \tag{5.17}$$

and

$$a = \frac{1}{2}\frac{\partial^2}{\partial x^2}\ln\psi + a_2, \tag{5.18}$$

respectively, where

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial^2}{\partial x^2}\{\varphi;x\} + 4\{\varphi;x\}^2 = 0 \tag{5.19}$$

and

$$\frac{\psi_t}{\psi_x} + \frac{\partial^2}{\partial x^2}\{\psi;x\} + \frac{1}{4}\{\psi;x\}^2 = 0. \tag{5.20}$$

Using the transformation

$$\psi_x = \varphi_x^{-2} \tag{5.21}$$

and invariance under the Moebius group, we find the following Bäcklund transformation:

$$\varphi_{n,x} = \psi_n/\psi_{n,x}^{1/2}, \tag{5.22}$$

$$\psi_{n,x} = \varphi_{n-1}^4/\varphi_{n-1,x}^2. \tag{5.23}$$

Letting

$$\varphi_n = P_n/P_{n-1} \tag{5.24}$$

and

$$\psi_n = Q_n/Q_{n-1} \tag{5.25}$$

obtains

$$P_{n-1}P_{n,x} - P_{n-1,x}P_n = Q_n, \tag{5.26}$$

$$Q_{n-1}Q_{n,x} - Q_{n-1,x}Q_n = P_{n-1}^4. \tag{5.27}$$

It is readily found that

$$\begin{aligned} &P_0 = 1, \quad Q_0 = 1, \\ &P_1 = x, \quad Q_1 = 1, \\ &P_2 = x^5 - 720t, \quad Q_2 = x^5 + 180t. \end{aligned} \tag{5.28}$$

are the first terms (after normalization) that satisfy Eqs. (5.26) and (5.27) and define (rational) solutions of Eqs. (5.19) and (5.20).

APPENDIX A: LAX PAIR AND BÄCKLUND TRANSFORMATIONS FOR THE CAUDREY–DODD–GIBBON EQUATION

In Sec. 3 the Caudrey–Dodd–Gibbon equation

$$u_t + \frac{\partial}{\partial x}(u_{xxxx} + 30uu_{xx} + 60u^3) = 0 \tag{A1}$$

was found to have the Bäcklund transformation

$$u = \frac{\partial^2}{\partial x^2}\ln\varphi + u_2, \tag{A2}$$

where u_2 satisfies (A1) and

$$\text{(i)}\quad u_2 = -\frac{1}{6}\frac{\varphi_{xxx}}{\varphi_x}, \tag{A3}$$

$$\text{(ii)}\quad \frac{\varphi_t}{\varphi_x} + \frac{\partial^2}{\partial x^2}\{\varphi;x\} + 4\{\varphi;x\}^2 = 0. \tag{A4}$$

Equations (A3) and (A4) may be rewritten as the following "Lax pair":

$$\varphi_{xxx} + 6u_2\varphi_x = 0, \tag{A5}$$

$$\varphi_t = -18u_{2x}\varphi_{xx} + 6(u_{2xx} - 6u_2^2)\varphi_x. \tag{A6}$$

With the exception that the spectral parameter vanishes, this is the Lax pair found in Ref. 4.

To obtain a Lax pair with the spectral parameter, it is necessary to generalize the procedures introduced in Ref. 2. That is, we define a Bäcklund transformation (A2), where (u,u_2) satisfy (A1). In Sec. 3 the resulting expressions were ordered according to the inverse powers of φ, i.e., (3.6iii, iv, and v). Herein, other than requiring that u_2 satisfy (A1) the various terms are collected into a single equation, obtaining

$$\frac{\partial^2}{\partial x^2}\left(\frac{\varphi_t}{\varphi}\right)+\frac{\partial}{\partial x}\left(\frac{H_5}{\varphi}+\frac{H_4}{\varphi^2}\right)=0, \tag{A7}$$

where

$$H_4=-\varphi_x^2\left\{\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2 + 5(\vartheta_{xx}+\vartheta^2+2\{\varphi;x\}\vartheta)\right\}, \tag{A8}$$

$$H_5=\varphi_x\left\{\frac{\partial^3}{\partial x^3}\{\varphi;x\}+4\frac{\partial}{\partial x}\{\varphi;x\}^2+5\vartheta\frac{\partial}{\partial x}\{\varphi;x\}\right\} + \vartheta_{xx}\left\{\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2+5\vartheta_{xx} + 5\vartheta^2+10\{\varphi;x\}\vartheta\right\}, \tag{A9}$$

$$\vartheta=\{\varphi;x\}+6W, \tag{A10}$$

and

$$W=u_2+\frac{1}{4}\frac{\varphi_{xx}^2}{\varphi_x^2}. \tag{A11}$$

Now, letting

$$\vartheta=6\lambda\varphi/\varphi_x, \tag{A12}$$

it is found from (A10) and (A11) that

$$\varphi_{xxx}+6u_2\varphi_x=6\lambda\varphi. \tag{A13}$$

From (A7)–(A9) and (A12) there results

$$\frac{\partial^2}{\partial x^2}\left\{\frac{\varphi_t}{\varphi}+\frac{\varphi_x}{\varphi}\left(\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2+30\lambda\frac{\varphi\varphi_{xxx}}{\varphi_x^2} - 30\lambda\frac{\varphi\varphi_{xx}^2}{\varphi_x^3}-30\lambda\frac{\varphi_{xx}}{\varphi_x}-180\lambda^2\frac{\varphi^2}{\varphi_x^2}\right)\right\}=0. \tag{A14}$$

Setting the term inside the bracket equal to 0,

$$\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2+30\lambda\frac{\varphi\varphi_{xxx}}{\varphi_x^2} - 30\lambda\frac{\varphi\varphi_{xx}^2}{\varphi_x^3}-30\lambda\frac{\varphi_{xx}}{\varphi_x}-180\lambda^2\frac{\varphi^2}{\varphi_x^2}=0. \tag{A15}$$

Using (A13),

$$\varphi_t=(54\lambda-18u_{2x})\varphi_{xx}+6(u_{2xx}-6u_2^2)\varphi_x+216\lambda u_2\varphi. \tag{A16}$$

Equations (A13) and (A16) constitute the Lax pair for the Caudrey–Dodd–Gibbon equation,[4] where λ is the spectral parameter. We note that Eq. (A15) is not invariant under the Moebius group.

APPENDIX B: SOME SEVENTH-ORDER EQUATIONS

We consider when the equation

$$\frac{\varphi_t}{\varphi_x}+\frac{\partial^4}{\partial x^4}\{\varphi;x\}+\alpha\{\varphi;x\}\frac{\partial^2}{\partial x^2}\{\varphi;x\} +\beta\left(\frac{\partial}{\partial x}\{\varphi;x\}\right)^2+\lambda\{\varphi;x\}^3=0 \tag{B1}$$

has a transformation

$$\varphi_x=\psi_x^m \tag{B2}$$

preserving the form of (B1).

Directly,

$$\{\varphi;x\}=m\frac{\psi_{xxx}}{\psi_x}-\left(\frac{m^2}{2}+m\right)\frac{\psi_{xx}^2}{\psi_x^2} \tag{B3}$$

and

$$\varphi_{xt}=m\psi_x^{m-1}\psi_{xt}. \tag{B4}$$

We note that

$$m\psi_x^{m-1}\psi_{xt}=\frac{\partial}{\partial x}(\psi_x^m F) =\psi_x^m\frac{\partial}{\partial x}F+m\psi_x^{m-1}\psi_{xx}F \tag{B5}$$

or

$$m\psi_{xt}=\psi_x\frac{\partial}{\partial x}F+m\psi_{xx}F. \tag{B6}$$

Therefore, for Eq. (B6) to be of the form (B1)

$$\psi_x\frac{\partial}{\partial x}F+m\psi_{xx}F=\frac{\partial}{\partial x}G, \tag{B7}$$

where G is a functional of ψ_x. Expressions on the lhs of (B7) that are not "gradients" must vanish. In this case, we find:

(i) Term $\psi_{xx}\psi_{xxxx}^2/\psi_x^2$ obtains the condition

$$2m+7+2m(\alpha-\beta)=0. \tag{B8}$$

(ii) Term $\psi_{xx}\psi_{xxx}^3/\psi_x^3$ obtains the condition

$$17m+42+\tfrac{1}{2}\alpha m(9m+28)-6\beta m(m+3)-3\lambda m^2=0. \tag{B9}$$

(iii) Term $\psi_{xxx}^3\psi_{xxx}^2/\psi_x^4$ obtains the condition

$$-39m-84+\alpha m(3m^2-\tfrac{9}{2}m-25) -2\beta m(m^2-5m-16)+3\lambda m^2(m+2)=0. \tag{B10}$$

(iv) Term ψ_{xx}^7/ψ_x^6 obtains the condition

$$60(m+2)-\tfrac{1}{2}\alpha m(13m^2-8m-68)+2\beta m(2m^2-7m-22) +\tfrac{3}{4}\lambda m^2(m^3+m^2-8m-12)=0. \tag{B11}$$

Equation (B8)–(B11) have the following solutions:

(i) $m=-1,\quad \alpha=\beta+\frac{5}{2},\quad 6\lambda=5\beta+\frac{5}{2},$ (B12)

(ii) $m=-2,\quad \alpha=\beta+\frac{3}{4},\quad 6\lambda=\beta+\frac{1}{4},$ (B13)

(iii) $m=-\frac{1}{2},\quad \alpha=12,\quad \beta=6,\quad \lambda=\frac{32}{3},$ (B14)

(iv) $m=-\frac{1}{3},\quad \alpha=26,\quad \beta=\frac{33}{2},\quad \lambda=48,$ (B15)

(v) $m=-\frac{2}{5},\quad \alpha=5,\quad \beta=\frac{3}{4},\quad \lambda=\frac{3}{4}.$ (B16)

Further calculation obtains that Eq. (B6) will be of the form (B1) when

(i) $m=-1,\quad \alpha=5,\quad \beta=\frac{5}{2},\quad \lambda=\frac{5}{2},$

(ii) $m=-2,\quad \alpha=\frac{3}{2},\quad \beta=\frac{3}{4},\quad \lambda=\frac{1}{6},$ (B17)

(iii) $m=-\frac{1}{2},\quad \alpha=12,\quad \beta=6,\quad \lambda=\frac{32}{3}$

The transformations defined by (B15) and (B16) do not preserve the form of Eq. (B1).

[1] J. Weiss, M. Tabor, and G. Carnevale, "The Painlevé property for partial differential equations," J. Math. Phys. **24**, 522 (1983).

[2] M. J. Ablowitz, A. Ramani, and H. Segur, "A connection between nonlinear evolution equations and ordinary differential equations of *P*-type. I," J. Math. Phys. **21**, 715 (1980).

[3]J. Weiss, "The Painlevé property for partial differential equations. II Bäcklund transformation, Lax pairs and the Schwarzian derivative," J. Math. Phys. **24**, 1405 (1983).
[4]P. J. Caudrey, R. K. Dodd, and J. D. Gibbon, "A New Heirarchy of Korteweg–de Vries Equations," Proc. Roy Soc. Lond. A **351**, 407 (1976).
[5]R. K. Dodd and J. D. Gibbon, "The Prolongation Structure of a Higher Order Korteweg–de Vries Equation," Proc. Roy. Soc. Lond. A **358**, 287 (1977).
[6]A. P. Fordy and John Gibbons, "Some Remarkable Nonlinear Transformations," Phys. Lett. A **75**, 325 (1980).
[7]Y. F. Chang, J. M. Greene, M. Tabor, and J. Weiss, "The Analytic Structure of Dynamical Systems and Self-Similar Natural Boundaries," Physica D **8**, 183 (1983).
[8]P. Lax, "Almost Periodic Solutions of the KdV Equation," SIAM Rev. **18**, 351 (1976).
[9]B. Fuchssteiner and W. Oevel, "The bi-Hamiltonian structure of some nonlinear fifth- and seventh-order differential equations and recursion formulas for their symmetries and conserved covariants," J. Math. Phys. **23**, 358 (1982).
[10]A. Fordy and J. Gibbons, "Factorization of operators. I. Miura transformations," J. Math. Phys. **21**, 2508 (1980).
[11]A. Fordy and J. Gibbons, "Factorization of operators. II," J. Math. Phys. **22**, 1170 (1981).
[12]B. Kuperschmidt and G. Wilson, "Modifying Lax Equations and the Second Hamiltonian Structure," Invent. Math. **62**, 403 (1981).
[13]M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM Stud. Appl. Math. (SIAM, Philadelphia, 1981).
[14]H. Airault, H. P. McKean, and J. Moser, "Rational and Elliptic Solutions of the Korteweg-de Vries Equation and a Related Many-Body Problem," Comm. Pure Appl. Math. **30**, 95 (1977).

Expansions over the "squared" solutions and difference evolution equations

V. S. Gerdjikov[a] and M. I. Ivanov[a]
Joint Institute for Nuclear Research, Dubna, USSR

P. P. Kulish
Leningrad Branch of the Steklov Mathematical Institute, Leningrad, USSR

(Received 3 February 1981; accepted for publication 3 December 1982)

The completeness relation for the system of "squared" solutions of the discrete analog of the Zakharov–Shabat problem is derived. It allows one to rederive the known statements concerning the class of difference evolution equations related to this linear problem and to obtain additional results. These include: (i) the expansion of the potential and its variations over the system of "squared" solutions, the expansion coefficients being the scattering data and their variations, respectively; thus the interpretation of the inverse scattering transform (IST) as a generalized Fourier transform becomes obvious; (ii) compact expressions for the trace identities through the operator Λ, for which the "squared" solutions are eigenfunctions; (iii) brief exposition of the spectral theory of the operator Λ; (iv) direct calculation of the action-angle variables based on the symplectic form of the completeness relation; (v) the generating functional of the M operators in the Lax representation; (vi) the quantum version of the IST.

PACS numbers: 02.30.Hq

I. INTRODUCTION

The intensive development of the inverse scattering transform (IST) has led to the discovery of a vast number of completely integrable Hamiltonian systems. For such physically important nonlinear evolution equations (NLEE) as the KdV, nonlinear Schrödinger, sine–Gordon equations, etc., the classes of soliton solutions, the infinite series of conserved quantities, the Bäcklund transformations, the explicit form of the action-angle variables, etc. (see Ref. 1 and the review papers, Refs. 2–4) have been constructed and investigated.

The investigation of the class of NLEE, related to the one-dimensional Zakharov–Shabat system

$$\left[i\sigma_3\frac{d}{dx}+\begin{pmatrix}0 & q(x)\\ r(x) & 0\end{pmatrix}-\lambda\right]\psi(x,\lambda)=0,$$
$$\sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \tag{1.1}$$

has revealed the importance of: (i) the expansions over the "squared" solutions of (1.1)[2,5–8] and (ii) the operator for which the "squared" solutions of (1.1) are eigenfunctions. The spectral theory of the operator Λ [7] enables one to justify the suggested in Ref. 2 interpretation of the IST as a generalized Fourier transform, linearizing the corresponding NLEE. An important property of the operator Λ consists also of the fact that it generates the hierarchy of Hamiltonian structures for the NLEE.[9]

Besides the NLEE there also exist a number of important difference evolution equations (DEE), solvable by the IST.[1,3] An example of such system is the Toda chain.[10]

The main result of the present paper consists in the derivation of the complete integrability, the construction of the hierarchy of symplectic structures and the quantization of the DEE, related to the discrete analog of the Zakharov–Shabat system[11]:

$$\psi(n+1,z)=L(n,z)\psi(n,z),\quad L(n,z)=E(z)+Q(n), \tag{1.2}$$
$$E(z)=\begin{pmatrix}z & 0\\ 0 & z^{-1}\end{pmatrix},\quad Q(n)=\begin{pmatrix}0 & q(n)\\ r(n) & 0\end{pmatrix}.$$

Our construction is based on the completeness relation for the "squared" solutions of the system (1.2).

Ablowitz and Ladik have considered in Ref. 11 the more general at first sight system (we put it in the form, proposed in Ref. 12):

$$u(n+1,\zeta)=\mathscr{L}(n,\zeta)u(n,\zeta),$$
$$\mathscr{L}(n,\zeta)=(\mu_n\nu_n)^{-1/2}\begin{pmatrix}1 & S_n\\ T_n & 1\end{pmatrix}\begin{pmatrix}\zeta & Q_n\\ R_n & \zeta^{-1}\end{pmatrix},$$
$$\mu_n=1-Q_nR_n,\quad \nu_n=1-S_nT_n. \tag{1.3}$$

The class of DEE related to (1.3) includes the discrete analogs of the nonlinear Schrödinger, KdV, sine–Gordon equations, etc. For these DEE the soliton solutions, conservation laws, the Bäcklund transformations, Hamiltonian structure, and the asymptotic of the solutions for $t\to\infty$ are known.[3,11–15]

It comes out that the systems (1.2) and (1.3) are equivalent. (The authors are grateful to I. T. Khabibulin for this remark.) Indeed, it is easy to see that if we relate the potentials and the solutions of these problems by

$$S_n=q(2n+1),\quad Q_n=q(2n),$$
$$T_n=r(2n+1),\quad R_n=r(2n), \tag{1.4}$$
$$u(n,\zeta)|_{\zeta=z^2}=\prod_{k=-\infty}^{2n-1}h(k)E^{-1/2}(z)\psi(n,z)E^{1/2}(z),$$

where $h(k)=1-q(k)r(k)$, we obtain

$$\mathscr{L}(n,\zeta)|_{\zeta=z^2}=[h(2n)h(2n+1)]^{-1/2}E^{-1/2}(z)L(2n+1,z)$$
$$\times L(2n,z)E^{1/2}(z). \tag{1.5}$$

As a result all the objects related to the system (1.2) such as

[a] On leave of absence from the Institute of Nuclear Energy and Nuclear Research, Sofia, Bulgaria.

 0022-2488/84/010025-10$02.50

DEE, conservation laws, Hamiltonian structures, etc. transfer to the corresponding objects of the system (1.3). Therefore, we confine ourselves to the system (1.2).

The present paper is a further development of our preprint.[16] We regret that when writing this preprint we were not aware of Ref. 12. We thank the referee for calling our attention to this paper.

In Sec. II we derive the completeness relation for the "squared" solutions of (1.2). Starting from it, we easily reproduce the statements from Refs. 11–14, and also obtain additional results. These include: (i) the expansion of the potential of (1.2) and its variation over the "squared" solutions, which justify the interpretation of the IST as a Fourier transform (Sec. III); (ii) compact expressions for the trace identities (Sec. III); (iii) brief exposition of the spectral theory of the operator Λ (2.21) (Sec. II); (iv) direct calculation of the action-angle variables based on the symplectic completeness relation[7,8] (Sec. IV); (v) the generating functional of the M operators in the Lax representation (Sec. III). In Sec. V it is shown that the DEE related to (1.2) with the natural reduction $r(n) = \pm q^+(n)$ may be quantized through the quantum IST.[17–19]

II. COMPLETENESS RELATION OF THE "SQUARED" SOLUTIONS

Let us start with some known facts (see Refs. 3 and 11) from the direct and inverse scattering problem for the system (1.2). In order to make the exposition simpler, we consider the case when the potential $w(n) = \binom{q(n)}{r(n)} \in \mathfrak{S}(\mathbf{Z},\mathbb{C}^2)$, the space of complex-valued vector sequences such that

$$\lim_{n\to\infty} n^k w(n) = 0 \quad \text{for all } k = 0,1,2,\cdots . \tag{2.1}$$

This together with the condition

$$0 < \prod_{k=-\infty}^{\infty} |h(k)| < \infty, \quad h(k) = 1 - q(k)r(k) \tag{2.2}$$

ensures the existence and the analyticity properties of the Jost solutions of (1.2), introduced by

$$\lim_{n\to\infty} \psi(n,z)E^{-n}(z) = \mathbf{1}, \quad \lim_{n\to\infty} \phi(n,z)E^{-n}(z) = \mathbf{1},$$
$$\psi(n,z) = \|\psi^-,\psi^+\|, \quad \phi(n,z) = \|\phi^+,\phi^-\|,$$

where $\psi^+, \phi^+, (\psi^-, \phi^-)$ are analytic for $|z| > 1$ ($|z| < 1$). The transition matrix is introduced by

$$\phi(n,z) = \psi(n,z)S(z), \quad S(z) = \begin{pmatrix} a^+, & -b^- \\ b^+, & a^- \end{pmatrix},$$
$$\det S(z) = v = \prod_{k=-\infty}^{\infty} h(k). \tag{2.3}$$

We shall denote by χ^+ (χ^-) the fundamental solutions of (1.2), analytic for $|z| > 1$ ($|z| < 1$):

$$\chi^+(n,z) = \|\phi^+,\psi^+\|, \quad \chi^-(n,z) = \|\psi^-,\phi^-\|,$$
$$\chi^+(n,z) = \psi S^- = \phi S^+, \quad \chi^-(n,z) = \psi T^+ = \phi T^-, \tag{2.4}$$
$$S^+(z) = \begin{pmatrix} 1 & b^-/v \\ 0 & a^+/v \end{pmatrix}, \quad S^-(z) = \begin{pmatrix} a^+ & 0 \\ b^+ & 1 \end{pmatrix},$$
$$T^+(z) = \begin{pmatrix} 1 & -b^- \\ 0 & a^- \end{pmatrix}, \quad T^-(z) = \begin{pmatrix} a^-/v & 0 \\ -b^+/v & 1 \end{pmatrix};$$

obviously $S^-\hat{S}^+ = T^+\hat{T}^- = S(z)$. Here and in what follows by $\hat{X}$ we shall denote the matrix inverse to X, i.e., $\hat{X} \equiv X^{-1}$. The solutions χ^+ and χ^- satisfy the following relations:

$$\chi^+(n,z)E^{-n}(z) = \chi^-(n,z)E^{-n}(z)G(n,z), \quad |z| = 1,$$
$$G(n,z) = E^n(z)\hat{T}(z)S^-(z)E^{-n}(z), \tag{2.5}$$

on the unit circle S^1. If we consider $G(z)$ as a given matrix-valued function of $z \in S^1$, then this relation may be interpreted as a noncanonical Riemann problem.[20]

The continuous spectrum of the problem (1.2) has multiplicity 2 and fills up S^1. The discrete spectrum $\Delta = \Delta^+ \cup \Delta^-$ is located at the zeroes of $a^\pm(z)$,

$$\Delta^\pm \equiv \{z_{j\pm} : a^\pm(z_{j\pm}) = a^\pm(-z_{j\pm}) = 0, \quad |z_{j\pm}| \geqslant 1,\ j = 1,\ldots,N^\pm\}. \tag{2.6}$$

Here for simplicity we asume that $n^+ = n^- = N$. The fact, that $a^\pm(z)(b^\pm(z))$ are even (odd) functions of z follows from

Remark 1: If $\psi(n,z)$ is a solution of (1.2), then $(-1)^n\sigma_3\psi(n,-z)\sigma_3$ will also be a solution of (1.2).

From the analyticity of $\chi^\pm$ it follows that $a^\pm(z)$ will also be analytic functions of z for $|z| \geqslant 1$. One is able to derive the following dispersion relation for them:

$$\ln a^+(z) = \frac{1}{4\pi i}\oint_{S^1} \frac{d\zeta^2}{\zeta^2 - z^2} \ln[1 + \rho^+\rho^-(\zeta)] + \sum_{j=1}^{N} \ln \frac{z^2 - z_{j+}^2}{z^2 - z_{j-}^2}, \quad |z| > 1,$$
$$-\ln a^-(z) = \frac{1}{4\pi i}\oint_{S^1} \frac{d\zeta^2 \cdot z^2}{\zeta^2(\zeta^2 - z^2)} \ln[1 + \rho^+\rho^-(\zeta)] + \sum_{j=1}^{N} \ln \frac{(z^2 - z_{j+}^2)z_{j-}^2}{(z^2 - z_{j-}^2)z_{j+}^2}, \quad |z| < 1, \tag{2.7}$$

where $\rho^\pm(z) = b^\pm(z)/a^\pm(z)$ are the reflection coefficients for the system (1.2).

We shall not discuss the solution of the inverse scattering problem in detail; see Refs. 3, 11, and 20. Note only that the set of independent scattering data $\mathcal{T} = \mathcal{T}^+ \cup \mathcal{T}^-$

$$\mathcal{T}^\pm \equiv \{\rho^\pm(z) = -\rho^\pm(-z), z \in S^1;$$
$$c_j^\pm, z_{j\pm}, |z_{j\pm}| \geqslant 1,\ j = 1,\ldots,N\},$$
$$\rho^\pm = b^\pm/a^\pm(z), \quad c_j^\pm = b_j^\pm/\dot{a}_j^\pm,$$
$$\dot{a}_j^\pm = \left.\frac{da^\pm}{dz}\right|_{z=z_{j\pm}}, \tag{2.8}$$
$$b_j^\pm: \ \phi^\pm(n,z_{j\pm}) = b_j^\pm\psi^\pm(n,z_{j\pm}),$$

and the dispersion relation (2.7) allow one to reconstruct uniquely the functions $a^+(z)$ ($a^-(z)$) for all z, $|z| \geqslant 1$ ($|z| \leqslant 1$), and also $b^\pm(z)$ for $|z| = 1$.

It is instructive to consider the interrelations between the potential $w(n)$ and the set of scattering data $\mathcal{T}$, (2.8), following from the formulas

$$\hat{\chi}^{\pm}(n,z)\sigma_3\chi^{\pm}(n,z)|_{n=-\infty}^{\infty}$$
$$= 2\sum_{n=-\infty}^{\infty}\hat{\chi}^{\pm}(n+1,z)\sigma_3 Q(n)\chi^{\pm}(n,z), \tag{2.9}$$
$$\hat{\chi}(n,z)\delta\chi^{\pm}(n,z)|_{n=-\infty}^{\infty}$$
$$= \sum_{n=-\infty}^{\infty}\hat{\chi}^{\pm}(n+1,z)\delta Q(n)\chi^{\pm}(n,z),$$

which are direct consequences of (1.2). The lhs of (2.9) are expressed easily through the scattering data $\mathcal{T}$, (2.8), and their variations. Inserting the first line of (2.4) into (2.9) for the matrix elements of the rhs of (2.9) one obtains expressions of the type:

$$\sum_{n=-\infty}^{\infty}\tilde{\Phi}(n,z)w(n)h^{-1}(n),$$
$$\sum_{n=-\infty}^{\infty}\tilde{\Phi}^{\pm}(n,z)\sigma_3\delta w(n)h^{-1}(n), \tag{2.10}$$

where

$$\Phi^{\pm}(n,z) = v(n)\phi^{\pm}(n,z)\circ\phi^{\pm}(n+1,z), \quad \tilde{\Phi} = (\Phi_2, -\Phi_1), \tag{2.11}$$
$$\phi(n,z)\circ\psi(m,z) \overset{\text{def}}{=} \begin{pmatrix}\phi_1(n,z)\psi_1(m,z)\\ \phi_2(n,z)\psi_2(m,z)\end{pmatrix}, \quad v(n) = \prod_{k=n}^{\infty}h(k).$$

If we introduce in the space $\mathfrak{S}(\mathbf{Z},\mathbf{C}^2)$ the skew-scalar product, $X,Y\in\mathfrak{S}(\mathbf{Z},\mathbf{C}^2)$:

$$[X,Y] = \sum_{n=-\infty}^{\infty}\tilde{X}(n)Y(n)$$
$$= \sum_{n=-\infty}^{\infty}[X_2(n)Y_1(n) - X_1(n)Y_2(n)], \tag{2.12}$$

then the matrix elements of the rhs of (2.9) can be interpreted as expansion coefficients of $w(n)$ and $\sigma_3\delta w(n)$ over the "squared" solutions $\Phi^{\pm}(n,z)$ of (1.2), i.e., the terms (2.10) will have the form $[\Phi^{\pm}(n),w(n)h^{-1}(n)]$, $[\Phi^{\pm}(n),\sigma_3\delta w(n)h^{-1}(n)]$.

Let us introduce the system $\{\Phi\}$, $\{\Psi\}$ of "squared" solutions of (1.2) by

$$\{\Phi\}\equiv\{\Phi^{\pm}(n,z), z\in S^1; \Phi_j^{\pm}(n), \dot{\Phi}_j^{\pm}(n), j=1,\ldots,N\},$$
$$\{\Psi\}\equiv\{\Psi^{\pm}(n,z), z\in S^1; \Psi_j^{\pm}(n), \dot{\Psi}_j^{\pm}(n), j=1,\ldots,N\}, \tag{2.13}$$
$$\Psi^{\pm}(n,z) = v(n)\psi^{\pm}(n,z)\circ\psi^{\pm}(n+1,z),$$
$$\Psi_j^{\pm}(n) = \Psi^{\pm}(n,z_{j\pm}),$$
$$\dot{\Psi}_j^{\pm}(n) = \lim_{z\to z_{j\pm}}\frac{d}{dz}\Psi^{\pm}(n,z);$$

$\Phi_j^{\pm}(n)$ and $\dot{\Phi}_j^{\pm}(n)$ are obtained analogously from the definition of $\Phi^{\pm}(n,z)$ in (2.11). The completeness of the systems $\{\Psi\}$, $\{\Phi\}$ is proved by introducing the Green function $G = G^{\pm}(n,m,z)$, $|z|\gtrless 1$,

$$G^{\pm}(n,m,z) = \{2/[a^{\pm}(z)]^2\}\{\Psi^{\pm}(n,z)\tilde{\Phi}(m,z)\theta(n-m)$$
$$+\theta(m-n)[2(\phi^{\pm}(n,z)\circ\psi^{\pm}(n+1,z))$$
$$\times(\phi^{\pm}(m+1,z)\circ\psi^{\pm}(m,z))\tilde{}\,v(n)v(m)$$
$$-\Phi^{\pm}(n,z)\tilde{\Psi}^{\pm}(m,z)]\},$$
$$\theta(n-m) = \begin{cases}1, & n>m,\\ \frac{1}{2}, & n=m,\\ 0, & n<m,\end{cases} \tag{2.14}$$

and applying the contour integration method to the integral,

$$\frac{1}{2\pi i}\oint_{\gamma_+}\frac{dz}{z}G^{+}(n,m,z) - \frac{1}{2\pi i}\oint_{\gamma_-}\frac{dz}{z}G^{-}(n,m,z).$$

Here the contours $\gamma_+ = S^1\cup\bar{S}^{\infty}$, $\gamma_- = S^1\cup\bar{S}^{0}$, where S^1 is the positively oriented unit circle and $\bar{S}^{\infty}$ and $\bar{S}^{0}$ the negatively oriented circles with infinitely large and infinitely small radii resp. The result is

$$h(n)\delta(n-m) = \frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}\left[\frac{\Psi^{+}(n,z)\tilde{\Phi}^{+}(m,z)}{[a^{+}(z)]^2} - \frac{\Psi^{-}(n,z)\tilde{\Phi}^{-}(m,z)}{[a^{-}(z)]^2}\right]$$
$$-2\sum_{j=1}^{N}[X_j^{+}(n,m) + X_j^{-}(n,m)], \tag{2.15}$$
$$X_j^{\pm}(n,m) = \frac{1}{z_{j\pm}(a_j^{\pm})^2}[\Psi_j^{\pm}(n)\dot{\tilde{\Phi}}_j^{\pm}(m) + \dot{\Psi}_j^{\pm}(n)\tilde{\Phi}_j^{\pm}(m)]$$
$$-\frac{\dot{a}_j^{\pm} + z_{j\pm}\ddot{a}_j^{\pm}}{z_{j\pm}^2(\dot{a}_j^{\pm})^3}\Psi_j^{\pm}(n)\tilde{\Phi}_j^{\pm}(m).$$

This completeness relation may be rewritten in the so-called symplectic form:

$$h(n)\delta(n-m) = \oint_{S^1}\frac{dz}{z}[P(n,z)\tilde{Q}(m,z) - Q(n,z)\tilde{P}(m,z)]$$
$$+2\sum_{j=1}^{N}[P_j^{+}(n)\tilde{Q}_j^{+}(m) - Q_j^{+}(n)\tilde{P}_j^{+}(m)$$
$$+P_j^{-}(n)\tilde{Q}_j^{-}(m) - Q_j^{-}(n)\tilde{P}_j^{-}(m)], \tag{2.16}$$

where

$$P(n,z) = -(1/2\pi)(\rho^{+}\Psi^{+} + \rho^{-}\Psi^{-})(n,z)$$
$$= -(1/2\pi v)(\sigma^{+}\Phi^{+} + \sigma^{-}\Phi^{-})(n,z),$$
$$Q(n,z) = (iv/b^{+}b^{-})\left(\rho^{+}\Psi^{+} - \frac{\sigma^{+}}{v}\Phi^{+}\right)(n,z)$$
$$= (iv/2b^{+}b^{-})\left(\frac{\sigma^{-}}{v}\Phi^{-} - \rho^{-}\Psi^{-}\right)(n,z), \tag{2.17}$$
$$P_j^{\pm}(n) = \mp(ic_j^{\pm}/z_{j\pm})\Psi_j^{\pm}(n),$$
$$Q_j^{\pm}(n) = \mp\tfrac{1}{2}i[m_j^{\pm}\dot{\Phi}_j^{\pm}(n) - c_j^{\pm}\dot{\Psi}_j^{\pm}(n)],$$
$$\sigma^{\pm}(z) = b^{\mp}(z)/a^{\pm}(z), \quad m_j^{\pm} = (b_j^{\pm}a_j^{\pm})^{-1}.$$

The two systems $\{\Psi\}$ and $\{\Phi\}$ are biorthogonal with respect to the skew-scalar product (2.12). Indeed, using (1.2), one can verify the following biquadratic relations between any two solutions $\phi(n,z)$ and $\psi(n,\zeta)$ of (1.2):

$$[\Phi(n,z),\Psi(n,\zeta)h^{-1}(n)]$$
$$= \frac{\zeta}{z}\cdot\frac{v^2(n)}{\zeta^2 - z^2}$$
$$\times[z\phi_2(n,z)\psi_1(n,\zeta) - \zeta\phi_1(n,z)\psi_2(n,\zeta)]^2|_{n=-\infty}^{\infty}. \tag{2.18}$$

Making use of (2.4) and of the fact that

$$\text{P.v.}\lim_{n\to\infty}\frac{(z/\zeta)^n}{\zeta - z} = \pi\delta(\arg z - \arg\zeta), \quad z,\zeta\in S^1,$$

we obtain

$$[\Phi^{\pm}(n,\zeta),\Psi^{\pm}(n,\zeta)h^{-1}(n)] = \mp 2\pi[a^{\pm}(\zeta)]^2\delta(\arg z - \arg\zeta), \quad z,\zeta\in S^1,$$
$$[\Phi_j^{\pm}(n),\Psi_k^{\pm}(n)h^{-1}(n)] = 0,$$
$$[\Phi_j^{\pm}(n),\dot{\Psi}_k^{\pm}(n)] = -\tfrac{1}{2}(\dot{a}_j^{\pm})^2 z_{j\pm}\delta_{jk}, \tag{2.19}$$
$$[\dot{\Phi}_j^{\pm}(n),\Psi_k^{\pm}(n)h^{-1}(n)] = -\tfrac{1}{2}(\dot{a}_j^{\pm})^2 z_{j\pm}\delta_{jk},$$
$$[\dot{\Phi}_j^{\pm}(n),\dot{\Psi}_k^{\pm}(n)h^{-1}(n)] = -\tfrac{1}{2}(\ddot{a}_j^{\pm} z_{j\pm} + \dot{a}_j^{\pm})\dot{a}_j^{\pm}\delta_{jk}.$$

From (2.18) and (1.29) we also have

$$[Q(n,z),P(n,\zeta)h^{-1}(n)] = -i\delta(\arg z - \arg\zeta), \quad z,\zeta\in S^1,$$
$$[Q_j^{\pm}(n),P_k^{\pm}(n)h^{-1}(n)] = \tfrac{1}{2}\delta_{jk}, \tag{2.20}$$
$$[Q_j^{\pm}(n),P_k^{\mp}(n)h^{-1}(n)] = 0.$$

Relations (2.19) and (2.20) allow one to conclude that the systems $\{\Psi\}$, $\{\Phi\}$, and $\{P,Q\}$ consist of a linearly independent element.

Now it is natural to introduce the operators $\Lambda_{\pm}$, for which the elements of $\{\Psi\}$ and $\{\Phi\}$ are eigenfunctions, i.e.,

$$(\Lambda_+ - z^2)\Psi^{\pm}(n,z) = 0, \quad (\Lambda_- - z^2)\Phi^{\pm}(n,z) = 0, \quad z\in S^1\cup\Delta,$$
$$(\Lambda_+ - z_{j\pm}^2)\dot{\Psi}_j^{\pm}(n) = 2z_{j\pm}\Psi_j^{\pm}(n), \tag{2.21}$$
$$(\Lambda_- - z_{j\pm}^2)\dot{\Phi}_j^{\pm}(n) = 2z_{j\pm}\Phi_j^{\pm}(n).$$

The explicit form of $\Lambda_{\pm}$ has been known.[11,12,14] For us it will be convenient to factorize them in the form

$$\Lambda_{\pm}X(n) = \Lambda_2^{\pm}\Lambda_1^{\pm}X(n), \quad X(n)\in\mathfrak{S}(\mathbb{Z},\mathbb{C}^2), \tag{2.22}$$

where the operators $\Lambda_i^{\pm}$, $i = 1,2$, are defined by

$$\Lambda_1^{+}\Psi^{\pm}(n,z) = z\overline{\Psi}^{\pm}(n,z), \quad \Lambda_1^{-}\Phi^{\pm}(n,z) = z\overline{\Phi}^{\pm}(n,z),$$
$$z = S^1\cup\Delta, \tag{2.23}$$
$$\Lambda_2^{+}\overline{\Psi}^{\pm}(n,z)z\Psi^{\pm}(n,z), \quad \Lambda_2^{-}\overline{\Phi}^{\pm}(n,z) = z\Phi^{\pm}(n,z),$$
$$\overline{\Psi}^{\pm}(n,z) = v(n)\psi^{\pm}(n,z)\circ\psi^{\pm}(n,z),$$
$$\overline{\Phi}^{\pm}(n,z) = v(n)\phi^{\pm}(n,z)\circ\phi^{\pm}(n,z).$$

The explicit form of $\Lambda_i^{\pm}$, $i = 1,2$ and their inverse is given by

$$\Lambda_1^{\pm}X(n) = \begin{pmatrix} X_1(n) \\ X_2(n-1)\end{pmatrix} \pm \begin{pmatrix} q(n) \\ -r(n-1)\end{pmatrix} \times\sum_{n}^{\pm}[r(k)X_1(k) + q(k)X_2(k)]h^{-1}(k),$$

$$\Lambda_2^{\pm}X(n) = h(n)\begin{pmatrix} X_1(n+1) \\ X_2(n)\end{pmatrix} \pm \begin{pmatrix} q(n) \\ -r(n)\end{pmatrix} \times\sum_{n+1}^{\pm}[r(k-1)X_1(k) + q(k)X_2(k)],$$

$$\hat{\Lambda}_1^{\pm}X(n) = h(n)\begin{pmatrix} X_1(n) \\ X_2(n+1)\end{pmatrix} \mp \begin{pmatrix} q(n) \\ -r(n)\end{pmatrix} \times\sum_{n+1}^{\pm}[r(k)X_1(k) + q(k-1)X_2(k)],$$

$$\hat{\Lambda}_2^{\pm}X(n) = \begin{pmatrix} X_1(n-1) \\ X_2(n)\end{pmatrix} \mp \begin{pmatrix} q(n-1) \\ -r(n)\end{pmatrix} \times\sum_{n}^{\pm}[r(k)X_1(k) + q(k)X_2(k)]h^{-1}(k), \tag{2.24}$$

where

$$\sum_n^{+} \equiv \sum_{k=n}^{\infty}, \quad \sum_n^{-} \equiv \sum_{k=-\infty}^{n-1}.$$

The condition (2.1) ensures that $\Lambda_i^{\pm}X\in\mathfrak{S}(\mathbb{Z},\mathbb{C}^2)$ for any $X\in\mathfrak{S}(\mathbb{Z},\mathbb{C}^2)$.

The operators $\Lambda_i^{\pm}$, $i = 1,2$, and $\Lambda_{\pm}$ satisfy conjugationlike relations with respect to the skew-scalar product (2.12):

$$[Y(n),\Lambda_1^{+}X(n)h(n)] = [\Lambda_2^{-}Y(n),X(n)],$$
$$[Y(n),\Lambda_2^{+}X(n)] = [\Lambda_1^{-}h(n)Y(n),X(n)], \tag{2.25}$$
$$[Y(n),\Lambda_{+}h(n)X(n)] = [\Lambda_{-}h(n)Y(n),X(n)].$$

The first two lines of (2.25) follow directly from the explicit form of $\Lambda_i^{\pm}$, (2.24), and from the definition of $[\ ,\ ]$, (2.12); the third line is a consequence of the first two and (2.22).

The spectral theory of the operators $\Lambda_{\pm}$ can be constructed analogously to Refs. 7 and 21. Here we will only show the interrelation between the Green function (2.14) and the operator Λ_+. Applying the contour integration method to the integral

$$\frac{1}{2\pi i}\oint_{\gamma_+}\frac{d\zeta}{\zeta}\frac{\zeta^2+z^2}{\zeta^2-z^2}G^{+}(n,m,\zeta) - \frac{1}{2\pi i}\oint_{\gamma_-}\frac{d\zeta}{\zeta}\frac{\zeta^2+z^2}{\zeta^2-z^2}G^{-}(n,m,\zeta),$$

we obtain the following spectral decomposition for G:

$$G(n,m,z) = \frac{i}{2\pi}\oint_{S^1}\frac{d\zeta}{\zeta}\frac{\zeta^2+z^2}{\zeta^2-z^2}\left\{\frac{\Psi^{+}(n,\zeta)\widetilde{\Phi}^{+}(m,\zeta)}{[a^{+}(\zeta)]^2} - \frac{\Psi^{-}(n,\zeta)\widetilde{\Phi}^{-}(m,\zeta)}{[a^{-}(\zeta)]^2}\right\} - 2\sum_{j=1}^{N}[Y_j^{+}(n,m) + Y_j^{-}(n,m)],$$
$$Y_j^{\pm}(n,m) = \lim_{\zeta\to z_{j\pm}}\frac{d}{d\zeta}\left[\frac{(\zeta - z_{j\pm})^2(\zeta^2+z^2)}{\zeta(\zeta^2-z^2)[a^{\pm}(\zeta)]^2}\Psi^{\pm}(n,\zeta)\widetilde{\Phi}^{\pm}(m,\zeta)\right]. \tag{2.26}$$

From (2.26), (2.21), and (2.15) it follows that

$$(\Lambda_+ + z^2)^{-1}(\Lambda_+ - z^2)G(n,m,z)h^{-1}(m) = \delta(n-m),$$

i.e., $G(n,m,z)$ is the Green function for the operator $(\Lambda_+ + z^2)^{-1}(\Lambda_+ - z^2)$. This result is esentially different from the one related to the Zakharov–Shabat system (1.1); there the continuous analogs of G and Λ_+ are related by $(\Lambda_+ - \lambda)G(x,y,\lambda) = \delta(x-y)$.

III. THE IST AS A FOURIER TRANSFORM

Let us start by deriving the expansions for $w(n)$ and $\sigma_3\delta w(n)$ over the systems of "squared" solutions $\{\Psi\}$ and $\{\Phi\}$. To do this, we multiply the completeness relations (2.15) and (2.16) by $w(m)h^{-1}(m)$ and $\sigma_3\delta w(m)h^{-1}(m)$ from the right and sum over m. Thus the corresponding expansion coefficients have the form (2.10) and through (2.9) are easily expressed in terms of the scattering data $\mathcal{T}$. The result is

$$w(n) = \frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}(\rho^+\Psi^+ + \rho^-\Psi^-)(n,z) - 2\sum_{j=1}^{N}\left[\frac{c_j^+}{z_{j+}}\Psi_j^+(n) - \frac{c_j^-}{z_{j-}}\Psi_j^-(n)\right], \quad (3.1a)$$

$$w(n) = -i\oint_{S^1}\frac{dz}{z}P(n,z) - 2i\sum_{j=1}^{N}[P_j^+(n) + P_j^-(n)] \quad (3.1b)$$

and

$$\sigma_3\delta w(n) = -\frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}(\delta\rho^+\Psi^+ - \delta\rho^-\Psi^-)(n,z) + 2\sum_{j=1}^{N}[Y_j^+(n) + Y_j^-(n)], \quad (3.2a)$$

$$Y_j^\pm(n) = \delta\left(\frac{c_j^\pm}{z_{j\pm}}\right)\Psi_j^\pm(n) + c_j^\pm\delta\ln z_{j\pm}\,\dot\Psi_j^\pm(n),$$

$$\sigma_3\delta w(n) = \oint_{S^1}\frac{dz}{z}[Q(n,z)\,\delta\hat p(z) - P(n,z)\,\delta\hat q(z)] + 2\sum_{j=1}^{N}[Z_j^+(n) + Z_j^-(n)], \quad (3.2b)$$

$$Z_j^\pm(n) = Q_j^\pm(n)\,\delta\hat p_j^\pm - P_j^\pm(n)\,\delta\hat q_j^\pm,$$

where

$$\hat p(z) = -(1/2\pi)\ln[1+\rho^+\rho^-(z)],$$
$$\hat q(z) = -\tfrac{1}{2}i\ln[b^+(z)/b^-(z)], \quad z\in S^1,$$
$$\hat p_j^\pm = \mp i\ln z_{j\pm}, \quad \hat q_j^\pm = \mp i\ln(b_j^\pm/\sqrt{v}), \quad (3.3)$$
$$\delta\hat p(z) = -[P(n,z),\sigma_3\delta w(n)h^{-1}(n)],$$
$$\delta\hat q(z) = -[Q(n,z),\sigma_3\delta w(n)h^{-1}(n)].$$

Now the parallel between the IST and the Fourier transform is obvious: The expansion coefficients in (3.1) and (3.2) are simply the scattering data $\mathcal{T}$, (2.8), and their variations. As a generalization of the usual "discrete exponent" z^n, one should consider $\{\Psi\}$ or $\{P,Q\}$; the role of the shift operator will be played by the operator Λ_+ (2.21).

From (3.1) and (3.2) there follows a more rigorous proof of the theorem, concerning the description of the DEE related to (1.2).

Theorem 1: Let $f(z^2)$ be a meromorphic function with poles lying outside of a certain neighborhood of the spectrum $S^1\cup\Delta$ of (1.2). Then $w(n,t)$ satisfies the DEE

$$\sigma_3\frac{dw}{dt} + f(\Lambda_+)w(n,t) = 0 \quad (3.4)$$

if and only if the scattering data $\mathcal{T}$, (2.8), satisfy the linear equations:

$$\frac{d\rho^\pm}{dt} \mp f(z^2)\rho^\pm(z,t) = 0,$$
$$\frac{dc_j^\pm}{dt} \mp f(z_{j\pm}^2)c_j^\pm(t) = 0, \quad (3.5)$$
$$\frac{dz_{j\pm}}{dt} = 0.$$

Proof: Let us insert the expansion of $w(n)$, (3.1a), and $\sigma_3(dw/dt)$ over the system $\{\Psi\}$ in the lhs of (3.4). The latter is obtained from (3.2a) by considering variations of the form $\sigma_3\delta w(n) = \sigma_3(dw/dt)\delta t + O((\delta t)^2)$, and differs from (3.2a) only in that the coefficients $\delta\rho^\pm,\cdots$ are replaced by $d\rho^\pm/dt,\cdots$; the same is true also for (3.2b). This gives

$$\sigma_3\frac{dw}{dt} + f(\Lambda_+)w(n,t) = \frac{1}{2\pi i}\oint_{S^1}\frac{dz}{z}\{[\rho_t^+ - f(z^2)\rho^+]\Psi^+(n,z) - (\rho_t^- + f(z^2)\rho^-)\Psi^-(n,z)\} + 2\sum_{j=1}^{N}[U_j^+(n) + U_j^-(n)], \quad (3.6)$$

$$U_j^\pm(n) = \left[\frac{d}{dt}\left(\frac{c_j^\pm}{z_{j\pm}}\right) - f(z_{j\pm}^2)\frac{c_j^\pm}{z_{j\pm}}\right]\Psi_j^\pm(n) + \frac{c_j^\pm}{z_{j\pm}}\frac{dz_{j\pm}}{dt}\dot\Psi_j^\pm(n).$$

In obtaining the rhs of (3.6) we have made use of (2.21). It remains to be noted that the lhs of (3.6) vanishes if and only if all the expansion coefficients on the rhs of (3.6) vanish, which readily gives (3.5). This last step follows also from the fact that the systems $\{\Psi\}$ and $\{\Phi\}$ are biorthogonal [see (2.19)].

Analogously, using the symplectic expansion (2.16), we can prove

Theorem 2: $w(n,t)$ satisfies (3.4) if and only if the set $\{\hat p,\hat q\}$ in (3.3) satisfies the linear equations:

$$\frac{d\hat p(z,t)}{dt} = 0, \quad i\frac{d\hat q(z,t)}{dt} = f(z^2),$$
$$\frac{d\hat p_j^\pm(t)}{dt} = 0, \quad i\frac{d\hat q_j^\pm(t)}{dt} = f(z_{j\pm}^2). \quad (3.7)$$

From (3.5) and (3.7) it follows that the DEE (3.4) has an infinite series of conserved quantities $C^{(p)}, p = 0, \pm 1,\cdots$. As a generating functional of $C^{(p)}$ it is natural to consider $\mathcal{A}(z)$:

$$\mathscr{A}(z) = \ln a^+(z), \quad |z| > 1,$$
$$\mathscr{A}(z) = -\ln a^-(z), \quad |z| < 1, \tag{3.8}$$

$C^{(p)}$ being the expansion coefficients of $\mathscr{A}(z)$:

$$\mathscr{A}(z) = \sum_{p=1}^{\infty} C^{(p)} z^{-2p}, \quad |z| \gg 1,$$
$$\mathscr{A}(z) = -\sum_{p=0}^{\infty} C^{(p)} z^{2p}, \quad |z| \ll 1. \tag{3.9}$$

To derive compact expressions for $C^{(p)}$ as functionals of $w(n)$, we start with the relation

$$z\frac{d\mathscr{A}}{dz} = \tfrac{1}{2}\operatorname{tr}\{[z\hat{\chi}^+(n,z)\dot{\chi}^+(n,z) - \dot{n}\sigma_3](\mathbf{1}+\sigma_3)\}|_{n=-\infty}^{\infty}, \quad |z| > 1, \tag{3.10}$$

which follows from (2.4), (3.8), and (1.2). Using (1.2) once more for the rhs of (3.10), we obtain

$$\sum_{n=-\infty}^{\infty} \{\tfrac{1}{2}\operatorname{tr}[\hat{\chi}^+(n,z)\sigma_3\chi^+(n,z)(\mathbf{1}+\sigma_3)] - 1\} = \sum_{n=-\infty}^{\infty} \sum\nolimits_n^+ \operatorname{tr}[\hat{\chi}^+(k+1,z)Q(k)\sigma_3\chi^+(k,z)\sigma_3], \tag{3.11}$$

which can be put into the form

$$z\frac{d\mathscr{A}}{dz} = -\sum_{n=-\infty}^{\infty} \sum\nolimits_n^+ v(k+1)\tilde{w}(k) \times \frac{\psi^+(k+1)\circ\phi^+(k) + \psi^+(k)\circ\phi^+(k+1)}{a^+(z)}. \tag{3.12}$$

Let us now expand $[v(k)/a^+(z)][\psi^+(k+1)\circ\phi^+(k) + \psi^+(k)\circ\phi^+(k+1)]$ over the system $\{\Psi\}$. The corresponding expansion coefficients are expressed through the scattering data $\mathscr{T}$ by using (2.18). Thus we obtain

$$\frac{v(n)}{a^+(z)}[\psi^+(n+1,z)\circ\phi^+(n,z) + \psi^+(n,z)\circ\phi^+(n+1,z)] = \frac{i}{2\pi}\oint_{S^1}\frac{d\zeta}{\zeta}\frac{\zeta^2+z^2}{\zeta^2-z^2}[\rho^+\Psi^+ + \rho^-\Psi^-](n,z) - 2\sum_{j=1}^{N}\left[\frac{c_j^+}{z_{j+}}\cdot\frac{z_{j+}^2+z^2}{z_{j+}^2-z^2}\Psi_j^+(n) - \frac{c_j^-}{z_{j-}}\cdot\frac{z_{j-}^2+z^2}{z_{j-}^2-z^2}\Psi_j^-(n)\right]$$
$$= (\Lambda_+ + z^2)(\Lambda_+ - z^2)^{-1}w(n). \tag{3.13}$$

Inserting the rhs of (3.13) into (3.12), we arrive at

$$z\frac{d\mathscr{A}}{dz} = -\sum_{n=-\infty}^{\infty}\sum\nolimits_n^+\frac{\tilde{w}(k)}{h(k)}(\Lambda_+ + z^2)(\Lambda_+ - z^2)^{-1}w(k), \tag{3.14}$$

which proves to be valid both for $|z| > 1$ and $|z| < 1$ (the considerations for $|z| < 1$ are analogous). Comparing (3.14) and (3.9) for $C^{(p)}$, we obtain

$$C^{(p)} = \frac{1}{p}\sum_{n=-\infty}^{\infty}\sum\nolimits_n^+\frac{\tilde{w}(k)\Lambda_+^p w(k)}{h(k)}, \quad p = \pm 1, \pm 2, \cdots. \tag{3.15}$$

The dispersion relations (2.7) allow one to express $C^{(p)}$ as functionals of the scattering data $\mathscr{T}$:

$$C^{(p)} = \frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}z^{2p}\ln[1+\rho^+\rho^-(z)] - \frac{1}{p}\sum_{j=1}^{N}(z_{j+}^{2p} - z_{j-}^{2p}), \quad p \neq 0, \tag{3.16}$$
$$C^{(0)} = -\ln v = \frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}\ln[1+\rho^+\rho^-(z)] - \sum_{j=1}^{N}\ln\frac{z_{j+}^2}{z_{j-}^2}.$$

The desired trace identities are obtained after equating the rhs of (3.15) and (3.16). Through the same pattern one can derive compact formulas for the variations of $\delta C^{(p)}$.[12] For this it is enough to note that

$$\delta\mathscr{A}(z) = \tfrac{1}{2}tr[\hat{\chi}^+(n,z)\delta\hat{\chi}(n,z)(\mathbf{1}+\sigma_3)]|_{n=-\infty}^{\infty} = \frac{1}{2}\sum_{n=-\infty}^{\infty}[\sigma_3\delta w(n)]\tilde{}\,\frac{v(n+1)}{a^+(z)} \times [\phi^+(n,z)\circ\psi^+(n+1,z) + \phi^+(n+1,z)\circ\psi^+(n,z)], \quad |z| > 1,$$

which with (3.13) directly leads to

$$\delta C^{(p)} = [\sigma_3\delta w(n)h^{-1}(n), \Lambda_+^p w(n)]. \tag{3.17}$$

We end this paragraph by reproducing in compact form the formulas from the traditional approach[2,3] to the DEE (3.4) as a consistency condition,

$$\frac{dL(n,z)}{dt} + L(n,z)M(n,z) - M(n+1,z)L(n,z) = 0, \tag{3.18}$$

of two linear problems: (1.2) and

$$\frac{d\psi(n,z)}{dt} = M(n,z)\psi(n,z). \tag{3.19}$$

Choosing $M(n,z) = \Sigma_k z^k M^{(k)}(n)$ as polynomial of z and z^{-1} and inserting in (3.18), one obtains recurrent relations for the coefficients $M^{(k)}(n)$ [11]; trying to solve them, one, after somewhat tedious calculations, naturally obtains the Λ operator.

Here we shall use another approach, developed for the NLEE by Gel'fand and Dickey[22]; see also Ref. 21. Let us introduce the resolvent of the system (1.2)[23]

$$\mathscr{R}(n,m,z)=\mathscr{R}^{\pm}(n,m,z),\quad |z|\gtrless 1,$$
$$\mathscr{R}^{\pm}(n,m,z)=\chi^{\pm}(n,z)\Theta(n-m)\hat{\chi}^{\pm}(m+1,z), \tag{3.20}$$
$$\Theta^{+}(n-m)=\begin{cases}\mathrm{diag}(-1,0), & m>n,\\ \mathrm{diag}(0,1), & m<n,\end{cases}$$
$$\Theta^{-}(n-m)=\begin{cases}\mathrm{diag}(0,-1), & m>n,\\ \mathrm{diag}(1,0), & m<n,\end{cases}$$

and define its "diagonal" as
$R(n,z)=\mathscr{R}(n,n-1,z)-\frac{1}{2}=-\frac{1}{2}\chi^{\pm}(n,z)\sigma_3\hat{\chi}^{\pm}(n,z)$. It is easy to verify that $R(n,z)$ satisfies the equation

$$L(n,z)R(n,z)-R(n+1,z)L(n,z)=0. \tag{3.21}$$

Since $R^{\pm}(n,z)$ is analytic in z for $|z|\gtrless 1$, one may consider the asymptotic expansions

$$R^{+}(n,z)=-\tfrac{1}{2}\sigma_3+\sum_{p=1}^{\infty}R^{(p)}(n)z^{-p},\quad |z|\gtrgtr 1,$$
$$R^{-}(n,z)=\tfrac{1}{2}\sigma_3+\sum_{p=1}^{\infty}R^{(-p)}(n)z^{p},\quad |z|\lessgtr 1, \tag{3.22}$$

Note that $R^{(p)}(n)$ and $M^{(p)}(n)$, $p\neq 0$, satisfy the same recurrent relations. From the definition of $R(n)$ and (2.23) and (2.24) we have

$$\sigma_3\begin{pmatrix}R_{12}^{\pm}(n)\\ R_{21}^{\pm}(n)\end{pmatrix}=\pm\frac{v(n)\phi^{\pm}(n,z)\circ\psi^{\pm}(n,z)}{a^{\pm}(z)}$$
$$=\tfrac{1}{2}z\hat{\Lambda}_{2}^{+}\frac{v(n)}{a^{\pm}(z)}[\phi^{\pm}(n+1,z)\circ\psi^{\pm}(n,z)$$
$$+\phi^{\pm}(n,z)\circ\psi^{\pm}(n+1,z)+w(n)], \tag{3.23}$$
$$R_{11}^{\pm}(n)=-R_{22}^{\pm}(n)$$
$$=\mp\frac{1}{2}\frac{v(n)}{a^{\pm}(z)}[\phi_1^{\pm}(n,z)\psi_2^{\pm}(n,z)$$
$$+\phi_2^{\pm}(n,z)\psi_1^{\pm}(n,z)].$$

Making use of (3.12) and (3.13), one obtains compact expressions for $R^{(p)}(n,z)$ through the operator Λ_+:

$$R^{(2p)}(n)-\sigma_3\sum_{n}^{+}\frac{\tilde{w}(k)}{h(k)}\Lambda_{+}^{p}w(k),\quad p=\pm 1,\pm 2,\cdots,$$
$$R^{(2p-1)}(n)=\begin{pmatrix}0, & R_{12}^{(2p-1)}(n)\\ R_{21}^{(2p-1)}(n), & 0\end{pmatrix}, \tag{3.24}$$
$$\begin{pmatrix}R_{12}^{(2p-1)}(n)\\ R_{21}^{(2p-1)}(n)\end{pmatrix}=-\sigma_3\hat{\Lambda}_{\epsilon}^{+}\Lambda_{+}^{p}w(n),\quad \epsilon=\begin{cases}2, & p>0,\\ 1, & p<0,\end{cases}$$

The M operators for the DEE (3.4) are simple linear combinations of $R^{(p)}(n)$. We write down the M operator only for the simplest case, when in (3.4) $f(z^2)=\nu z^{2N}$, $N>0$, $\nu=\mathrm{const}$:

$$M^{(N)}(n,z)=\nu\left[\sum_{p=0}^{2N-1}z^{2N-p}R^{(p)}(n)+\tfrac{1}{2}R^{(2N)}(1+\sigma_3)\right]. \tag{3.25}$$

Thus $R(n,z)$ may be considered as a generating functional of the M operators and also of the conserved quantities of the DEE (3.4). The last statement is obtained by comparing (3.24) and (3.11), which gives

$$z\frac{d\mathscr{A}}{dz}=\mp\sum_{n=-\infty}^{\infty}[\mathrm{tr}(R(n,z)\sigma_3)\pm 1],\quad |z|\gtrless 1. \tag{3.26}$$

IV. HIERARCHIES OF HAMILTONIAN STRUCTURES

The proof of the Hamiltonian structure of the DEE (3.4) is now easy. For this one should introduce the following symplectic form[12]:

$$\Omega^{(0)}=2i\sum_{n=-\infty}^{\infty}\frac{\delta q(n)\wedge\delta r(n)}{h(n)}$$
$$\equiv i[\sigma_3\delta w(n),\ \text{exterior product},\ \sigma_3\delta w(n)h^{-1}(n)], \tag{4.1}$$

where $\delta q\wedge\delta r=\delta_1 q\delta_2 r-\delta_2 q\delta_1 r$ is the usual exterior product. In order that the Hamiltonian equations of motion

$$\Omega^{(0)}\left(\sigma_3\frac{dw}{dt},\cdot\right)=\delta H_f(\cdot) \tag{4.2}$$

coincide with (3.4), one should choose H_f in the form

$$H_f=-i\sum_p f_p C^{(p)}$$
$$=-if_0\sum_{n=-\infty}^{\infty}\ln h(n)$$
$$-i\sum_{n=-\infty}^{\infty}\sum_{n}^{+}\tilde{w}(n)h^{-1}(n)F(\Lambda_+)w(n), \tag{4.3}$$

where

$$f(z^2)=\sum_p f_p z^{2p},\quad F(z^2)=\int^{z^2}\frac{ds}{s}[f(s)-f_0]. \tag{4.4}$$

The complete integrability of the DEE (3.4) becomes obvious after recalculating $\Omega^{(0)}$ and H_f in terms of the scattering data variations. Most simply $\Omega^{(0)}$ is calculated by inserting the symplectic expansion (2.16) into (4.1) and using the third line in (3.3). This immediately casts $\Omega^{(0)}$ in canonical form:

$$\Omega^{(0)}=2i\oint_{S^1}\frac{dz}{z}\delta\hat{p}(z)\wedge\delta\hat{q}(z)$$
$$+4i\sum_{j=1}^{N}[\delta\hat{p}_j^{+}\wedge\delta\hat{q}_j^{+}+\delta\hat{p}_j^{-}\wedge\delta\hat{q}_j^{-}], \tag{4.5}$$

which means that $\{\hat{p},\hat{q}\}$ is a set of canonical coordinates and momenta. From (3.16) and (4.3) we see that

$$H_f=-\oint_{S^1}\frac{dz}{z}f(z^2)\hat{p}(z)+i\sum_{j=1}^{N}[F_1(z_{j+}^2)-F_1(z_{j-}^2)],$$
$$F_1(s)=\int^{s}\frac{ds'}{s'}f(s'),\quad z_{j\pm}^2=\exp(\pm 2i\hat{p}_j^{\pm}), \tag{4.6}$$

i.e., H_f depends only on the set of the new momenta $\{\hat{p}\}$. Thus $\{\hat{p},\hat{q}\}$ in (3.3) is the set of the action-angle variables for the DEE (3.4).[12]

The symplectic structure $\Omega^{(0)}$ is not unique. One can introduce a one-parameter family of symplectic forms $\Omega^{(m)}$, generated from $\Omega^{(0)}$ (4.1) by the operator Λ_+:

$$\Omega^{(m)}=i[\sigma_3\delta w(n)h^{-1}(n),\Lambda_{m}^{+}\sigma_3\delta w(n)]. \tag{4.7}$$

The proof that $\Omega^{(m)}$ are symplectic is most easily performed as in Ref. 9 after recalculating $\Omega^{(m)}$ in terms of the scattering data variations, which now gives

$$\Omega^{(m)} = 2i \oint_{S^1} \frac{dz}{z} z^{2m}\, \delta\hat{p}(z) \wedge \delta\hat{q}(z) + 4i \sum_{j=1}^{N} [z_{j+}^{2m}\, \delta\hat{p}_j^{+} \wedge \delta\hat{q}_j^{+} + z_{j-}^{2m}\, \delta\hat{p}_j^{-} \wedge \delta\hat{q}_j^{-}]. \tag{4.8}$$

From (4.8) it is obvious that $\{\Omega^{(2m)}, m = 0, \pm 1, \pm 2, \cdots\}$ is a hierarchy of compatible symplectic forms, which generate a hierarchy of Hamiltonian structures for the DEE (3.4). Indeed, the choice $\Omega = \Omega^{(m)}, H = H_{f^{(m)}}$ in (4.2) with $f^{(m)}(z^2) = z^{2m} f(z^2)$ lead to the same DEE (3.4) as $\Omega = \Omega^{(0)}, H = H_f$.

In complete analogy to Refs. 7 and 8, one can define the Lagrange manifold for the DEE (3.4) by

$$\mathfrak{m}(t) \equiv \{X(n,t) \in \mathfrak{m}: [X(n,t), P(n,t,z)] = 0,\ z \in S^1 \cup \Delta\}.$$

Let us list without proof the main properties of $\mathfrak{m}(t)$:

(i) if $X \in \mathfrak{m}$, then $\Lambda_+ X = \Lambda_- X \in \mathfrak{m}$;

(ii) $\dim \mathfrak{m} = \operatorname{codim} \mathfrak{m}$;

(iii) $\sigma_3 \delta w(n) \in \mathfrak{m}$ if and only if $\delta\hat{p}(z) = 0$ for all $z \in S^1 \cup \Delta$, i.e., the restriction of $\Omega^{(m)}|_{\mathfrak{m}} \equiv 0$ for all $m = 0, \pm 1, \cdots$.

Remark 2: From (2.17), (2.9), and (2.4) one verifies that $w(n,t) \in \mathfrak{m}(t)$. This together with the property (i) of $\mathfrak{m}$ gives $f(\Lambda_+)w = f(\Lambda_-)w$, i.e., the operators Λ_+ and Λ_- generate the same DEE (3.4).

Remark 3: If $w(n,t)$ satisfies any of the DEE (3.4), then $\sigma_3(dw/dt) \in \mathfrak{m}(t)$ for all t.

At the end of this paragraph let us consider two particular examples of soluble DEE. They are related to the system (1.2) with simple reductions of the potential, which naturally requires a recalculation of the action-angle variables.

A. The difference nonlinear Schrödinger equation (DNLS)

$$i\frac{dq(n,t)}{dt} = -[1 - \epsilon q^*(n)q(n)][q(n+1) + q(n-1)] + 2q(n), \quad \epsilon = \pm 1, \tag{4.9}$$

is obtained from (3.4) with $f(z^2) = i(2 - z^2 - z^{-2})$ provided the reduction $r(n) = \epsilon q^*(n)$ holds. This reduction imposes the following restrictions on the scattering data:

$$a^+(z) = a^-(1/z^*), \quad b^+(z) = -\epsilon b^{-*}(1/z^*), \quad z_{j+} = 1/z_{j-}^*, \quad c_j^- = \epsilon (c_j^+)^*/(z_{j+}^*)^2. \tag{4.10}$$

As a Hamiltonian and 2-form, generating (4.9), one can choose

$$\Omega_{\mathrm{DNLS}} = \epsilon \Omega^{(0)}|_{q = \epsilon r^*} = -\frac{\epsilon}{\pi} \int_{-\pi}^{\pi} d\tau\, \delta(\arg b^+(e^{i\tau})) \wedge \delta \times \ln[1 - \epsilon|\rho^+(e^{i\tau})|^2] - 4\epsilon \sum_{j=1}^{N} [\delta\zeta_j \wedge \delta\rho_j + \delta\omega_j \wedge \delta\xi_j], \tag{4.11}$$

where $\ln z_{j+}^2 = \zeta_j + i\omega_j$, $\ln(b_j^+/\sqrt{v}) = \xi_j + i\rho_j$;

$$H_{\mathrm{DNLS}} = \epsilon(2C^{(0)} - C^{(1)} - C^{(-1)})|_{q = \epsilon r^*} = -\sum_{n=-\infty}^{\infty} \{q^*(n)[q(n+1) + q(n-1)] + 2\epsilon \ln[1 - \epsilon q^*(n)q(n)]\} = 2\epsilon\left\{-\frac{1}{\pi}\int_{-\pi}^{\pi} d\tau \sin^2\tau \ln[1 - \epsilon|\rho^+(e^{i\tau})|^2] + 2\sum_{j=1}^{N} [\cos\omega_j \sinh\zeta_j - \xi_j]\right\}. \tag{4.12}$$

The explicit form of the action-angle variables is obvious from (4.11). Note that from (4.10) and (3.16) one obtains $C^{(p)} = C^{(-p)*}$

B. The difference modified KdV equation (DMKdV)

$$i\frac{dq(n,t)}{dt} = -[1 - \epsilon q^2(n)][q(n+1) - q(n-1)], \quad \epsilon = \pm 1, \tag{4.13}$$

is obtained from (3.4) with $f(z^2) = z^{-2} - z^2$ provided that the reduction $q(n) = \epsilon r(n)$ holds, which means that

$$a^+(z) = a^-(1/z), \quad b^+(z) = -\epsilon b^-(1/z), \quad z_{j+} = 1/z_{j-}, \quad c_j^+ = \epsilon c_j^- z_{j+}^2. \tag{4.14}$$

As it has been noted in Ref. 12, $\Omega^{(0)}$ vanishes identically if this reduction is imposed. Therefore, we should use another symplectic structure from the hierarchy, e.g.,

$$\Omega_{\mathrm{MDKdV}} = i\Omega^{(-1)}|_{q = \epsilon r} = -2\epsilon \times \sum_{n=-\infty}^{\infty} \left[2\delta q(n) \wedge \delta q(n+1) + \delta \ln h(n) \wedge \delta \times \left(\sum\nolimits_n^+ q(k)q(k-1)\right)\right] = -\frac{2}{\pi}\int_0^{\pi} d\tau \sin 2\tau\, \delta \times \ln[1 - \epsilon\rho^+(e^{i\tau})\rho^+(e^{-i\tau})] \wedge \delta \times \left[-\frac{i}{2}\ln\frac{b^+(e^{i\tau})}{b^+(e^{-i\tau})}\right] + 4\sum_{j=1}^{N} \delta\cosh(\zeta_j + i\omega_j) \wedge \delta(\xi_j + i\rho_j), \tag{4.15}$$

$\ln z_{j+}^2 = \zeta_j + i\omega_j$, $\ln(b_j^+/\sqrt{v}) = \xi_j + i\rho_j$.

The corresponding Hamiltonian is

$$H_{\mathrm{MDKdV}} = C^{(0)} - C^{(2)} = -\sum_{n=-\infty}^{\infty} \{\ln[1 - \epsilon q^2(n)] + \epsilon q(n)q(n-2)[1 - q^2(n-1)] - \tfrac{1}{2}q^2(n)q^2(n-1)\} = -\frac{2}{\pi}\int_0^{\pi} d\tau \sin^2 2\tau \times \ln[1 - \epsilon\rho^+(e^{i\tau})\rho^+(e^{-i\tau})] - 2\sum_{j=1}^{N} [\zeta_j + i\omega_j - \tfrac{1}{2}\sinh 2(\zeta_j + i\omega_j)].$$

If besides $q(n) = \epsilon r(n)$ we require $q(n) = \epsilon r^*(n)$, then the scattering data $\mathcal{T}$, (2.7), will satisfy both (4.10) and (4.14). In this case the eigenvalues appear either in four tuples $(z_{j+}, z_{j+}^*,$

$-z_{j+}$, $-z_{j+}^*$)or pairwise if among z_{j+} there occur real or pure imaginary numbers. Let us introduce the notations:

$$\begin{aligned} z_{j+}^2 &= e^{\xi_j + i\omega_j}, \quad b_j^+/\sqrt{v} = e^{\xi_j + i\rho_j}, \quad j = 1,...,N_1, \\ z_{\alpha+}^2 &= e^{\epsilon_\alpha}, \quad b_\alpha^+/\sqrt{v} = e^{\gamma_\alpha}, \quad \alpha = 1,...,N_2, \\ z_{\beta+}^2 &= -e^{\eta_\beta}, \quad b_\beta^+/\sqrt{v} = e^{\theta_\beta}, \quad \beta = 1,...,N_3, \\ & 2N_1 + N_2 + N_3 = N. \end{aligned} \tag{4.16}$$

Then the 2-forms $i\Omega^{(m)} = -i\Omega^{(-m)}$ become real and equal to

$$\begin{aligned} i\Omega^{(m)} = \frac{2}{\pi}\int_0^\pi d\tau \sin 2m\tau\, \delta(\ln[1-\epsilon|\rho^+(e^{i\tau})|^2]) \\ \wedge \delta(\arg b^+(e^{i\tau})) \\ - \frac{8}{m}\sum_{j=1}^{N_1}\{\delta[\cos(m\omega_j)\cosh(m\xi_j)]\wedge\delta\beta_j \\ - \delta[\sin(m\omega_j)\sinh(m\xi_j)]\wedge\delta\rho_j\} \\ - \frac{4}{m}\sum_{\alpha=1}^{N_2}\delta\cosh(m\epsilon_\alpha)\wedge\delta\gamma_\alpha \\ - \frac{4(-1)^m}{m}\sum_{\beta=1}^{N_3}\delta\cosh m\eta_\beta\wedge\delta\theta_\beta. \end{aligned} \tag{4.17}$$

From (4.17) with $m = 1$ we easily get the action-angle variables for the MDKdV (4.13) with real-valued $q(n)$. If in (4.13) we change the variables to $u(n) = \operatorname{arctanh} q(n)$ for $\epsilon = 1$, and $u(n) = \operatorname{arctanh} q(n)$ for $\epsilon = -1$, we obtain another interesting DEE:

$$\begin{aligned} \frac{du(n,t)}{dt} &= \tan u(n+1) - \tan u(n-1), \quad \epsilon = 1, \\ \frac{du(n,t)}{dt} &= \tanh u(n+1) - \tanh u(n-1), \quad \epsilon = -1. \end{aligned} \tag{4.18}$$

The equivalence of (4.13) and (4.18) is obvious only for $\epsilon = -1$; for $\epsilon = 1$ the change of the variables $u(n) = \arctan q(n)$ is singular.

There are more examples of interesting DEE which can be obtained from (3.4). Obviously for all of them one can calculate the Hamiltonian structures and the action-angle variables, following the above considerations.

V. QUANTUM DIFFERENCE NONLINEAR EQUATIONS

The nonlinear DEE mentioned above can be solved by a quantum version of IST. Let us consider quantum DNS (4.9) where now the quantities $q(n)$ and $q^+(n)$ are operators with commutation relations ($m,n = 0, \pm 1, \pm 2,..., \pm N$)

$$[q(m),q^+(n)] = \hbar[1-\epsilon q^+(n)q(n)]\delta(n-m). \tag{5.1}$$

Hereafter we shall use the normal ordering with respect to q and q^+. For finite N we can realize these operators in the state space $\mathscr{H}^{(N)}$:

$$\begin{aligned} \mathscr{H}^N &= \bigotimes_{n=-N+1}^{N}\mathscr{H}_n, \\ \mathscr{H}_n &= \mathscr{L}\{|0\rangle_n,|1\rangle_n,|2\rangle_n,\cdots\}, \end{aligned} \tag{5.2}$$

where $\mathscr{L}$ denotes closure of a linear space $\{\cdots\}$ and $|k\rangle_n = (q^+(n))^k|0\rangle_n, q(n)|0\rangle_n = 0$. As a consequence of (5.1) the norm in $\mathscr{H}^{(N)}$ is positive definite provided ${}_n\langle 0|0\rangle_n = 1$:

$$\langle k|l\rangle_n = \delta_{kl}(\hbar)^k\prod_{m=1}^{k}c_m, \quad c_m = \frac{1-e^{2\eta m}}{1-e^{2\eta}}. \tag{5.3}$$

The parameter $\eta = \frac{1}{2}\ln(1-\epsilon\hbar)$ is more appropriate in the following formulas. In order that the Heisenberg equations of motion coincide with (4.9), we must add the quantum corrections to the classical expression of the Hamiltonian (4.12):

$$\begin{aligned} H = -\sum_n\{q^+(n)[q(n+1)-q(n-1)] \\ -[2\hbar/\ln(1-\epsilon\hbar)]\ln[1-\epsilon q^+(n)q(n)]\}. \end{aligned} \tag{5.4}$$

The quantum version of IST (QIST) also uses an auxiliary linear problem. In this case we can take the same L operator (1.2), $r(n) = \epsilon q^+(n)$, with its entries as operators in $\mathscr{H}_n$ (5.2). The main step of QIST is the determination of commutation relations of the quantum scattering data or, to be more precise, the operator-valued entries of the monodromy matrix

$$T_N(z) = L_N(z)L_{N-1}(z)\cdots L_{-N+1} = \begin{pmatrix} A_N(z) & B_N(z) \\ C_N(z) & D_N(z) \end{pmatrix}, \tag{5.5}$$

$$R(\varphi)[T_N(z)\otimes T_N(\zeta)] = [I\otimes T_N(\zeta)][T_N(z)\otimes I]R(\varphi), \tag{5.6}$$

where $R(\varphi)$ is a 4×4 c-number matrix or intertwining operator, I is the identity operator in C^2, $\exp\varphi = z/\zeta$. The R matrix can be calculated from the very same relation (5.6) but with $L_n(z)$, $L_n(\zeta)$ instead of $T_N(z)$, $T_N(\zeta)$:

$$\begin{aligned} R(\varphi) &= \begin{pmatrix} a & 0 & 0 & 0 \\ 0 & b^- & c & 0 \\ 0 & c & b^+ & 0 \\ 0 & 0 & 0 & a \end{pmatrix}, \\ a &= \sinh(\varphi-\eta), \quad b^\pm = e^{\pm\eta}\sinh\varphi, \\ c &= -\sinh\eta. \end{aligned} \tag{5.7}$$

For finite chain with periodic boundary conditions ($2N + k = k$, mod $2N$) the trace of the monodromy matrix $t_N(z) = A_N(z) + D_N(z)$ is the generating functional of the quantum integrals of the motion. In order to define eigenstates and eigenvalues of $t_N(z)$, we shall need the following commutation relations (5.6):

$$[t_N(z),t_N(\zeta)] = 0, \quad [C_N(z),C_N(\zeta)] = 0, \tag{5.8}$$

$$\begin{aligned} A_N(z)C_N(\zeta) &= [1/b^-(\varphi)]C_N(\zeta)A_N(z) \\ &\quad -[c(\varphi)/b^-(\varphi)]C_N(z)A_N(\zeta), \\ D_N(\zeta)C_N(z) &= [1/b^-(\varphi)]C_N(z)D_N(\zeta) \\ &\quad -[c(\varphi)/b^-(\varphi)]C_N(\zeta)D_N(z). \end{aligned} \tag{5.9}$$

Since, when applied to the vacuum $|0\rangle = \Pi_{n=-N+1}^{N}|0\rangle_n$, $L_n(z)$ becomes triangular, one easily finds for the action of $A_N(z)$, $B_N(z)$, $D_N(z)$ on $|0\rangle$

$$\begin{aligned} A_N(z)|0\rangle &= z^{2N}|0\rangle, \quad D_N(z)|0\rangle = z^{-2N}|0\rangle, \\ B_N(z)|0\rangle &= 0. \end{aligned} \tag{5.10}$$

Using (5.8)–(5.10) via the general scheme of the QIST,[17,18] one constructs the eigenstates of $t_N(z)$:

$$|z_1,...,z_n\rangle = \prod_{k=1}^{n}C_N(z_k)|0\rangle \tag{5.11}$$

provided the quasimomenta z_k satisfy the algebraic equations ($\exp\lambda = z$)

$$(z_k)^{4N} = \prod_{l \neq k} \frac{\sinh(\lambda_k - \lambda_l + \eta)}{\sinh(\lambda_k - \lambda_l - \eta)}, \quad k = 1,2,\ldots,n. \quad (5.12)$$

The corresponding eigenvalue is given by

$$V(z,\{z_k\}_1^n) = z^{2N} \prod_{k=1}^{n} \frac{e^{\eta} \sinh(\lambda - \lambda_k - \eta)}{\sinh(\lambda - \lambda_k)} + z^{-2N} \prod_{k=1}^{n} \frac{e^{\eta} \sinh(\lambda - \lambda_k + \eta)}{\sinh(\lambda - \lambda_k)}. \quad (5.13)$$

For the energy of the state (5.11) we have

$$E(\{z_k\}_1^n) = \sum_{k=1}^{n} \epsilon(z_k), \quad \epsilon(z) = 2\hbar - z^2 - z^{-2}. \quad (5.14)$$

There exist different phases in the limit $N \to \infty$. The phase with finite number of particles is the simplest one. The state space has the Fock type structure with vacuum $|0\rangle$ and creation operators $q^+(n)$, $n = 0, \pm 1, \pm 2, \cdots$, or

$$R^+(z) = \lim_{N\to\infty} C_N(z)/z^{2N} A_N(z), \quad |z| = 1. \quad (5.15)$$

The additional factor z^{2N} is a consequence of the transition matrix definition

$$T(z) = \lim_{N\to\infty} E^{-N}(z) T_N(z) E^{-N}(z) = \lim_{N\to\infty} \begin{vmatrix} z^{-2N} A_N(z) & B_N(z) \\ C_N(z) & z^{2N} D_N(z) \end{vmatrix} = \begin{vmatrix} A(z) & B(z) \\ C(z) & D(z) \end{vmatrix}, \quad (5.16)$$

where $E(z) = \langle 0|L_n(z)|0\rangle = \mathrm{diag}(z,z^{-1})$. It is possible to define operator-valued Jost solutions (in the weak sense) and their analytic properties and relations to the transition matrix $T(z)$. The inverse to $L_n(z)$ is $[\rho_n = 1 - \epsilon q^+(n) q(n)]$

$$L_n^{-1}(z) = \frac{e^{-\eta}}{\rho_n} V L_n(e^{-\eta}/z) V^{-1}, \quad V = \mathrm{diag}(e^{-\eta/2}, -e^{\eta/2}). \quad (5.17)$$

Using $L_n^t(z) = \sigma_1 L_n(1/z)\sigma_1$, we get

$$T_N^{-1}(z) = Q_N^{-1} W T_N^t(e^{\eta} z) W^{-1}, \quad W = V\sigma_1, \quad Q_N = \prod_{n=-N+1}^{N} e^{\eta} \rho_n. \quad (5.18)$$

The operator $R^+(z)$, $R(z) = \epsilon D^{-1}(z) B(z)$ is called quantum scattering data. They are generators of the Zamolodchikov–Faddeev algebra. By means of the formulas (5.17) and (5.18) one can obtain a quantum analog of (2.5), i.e., the quantum Riemann problem. The reconstruction of the local quantum operators $q(n)$ and $q^+(n)$ from the quantum scattering data would enable one to calculate the Green's functions of this model.

ACKNOWLEDGMENTS

The authors are deeply grateful to E. Kh. Khristov and A. G. Reyman for helpful discussions.

[1] V. E. Zakharov, S. V. Manakov, S. P. Novikov, and L. P. Pitaevskii, *Soliton Theory: The Inverse Problem Method* (in Russian) (Nauka, Moscow, 1980).
[2] M. Ablowitz, D. Kaup. A. Newell, and H. Seeger, Stud. Appl. Math. **53**, 249 (1974).
[3] M. Ablowitz, Stud. Appl. Math. **58**, 17 (1978).
[4] L. D. Faddeev. in *Solitons*, edited by R. K. Bullough and P. Candrey, Topics in Current Physics (Springer-Verlag, Berlin, 1980), Vol. 17, p. 339.
[5] D. J. Kaup, Math. Anal. Appl. **54**, 789 (1976).
[6] D. J. Kaup and A. C. Newell, Adv. Math. **31**, 67 (1979).
[7] V. S. Gerdjikov and E. Kh. Khristov, Mat. Zametki **28**, 501 (1980) (in Russian); Bulg. J. Phys. **7**, 28 (1980) (in Russian).
[8] V. S. Gerdjikov and E. Kh. Khristov, Bulg. J. Phys. **7**, 119 (1980) (in Russian).
[9] P. P. Kulish and A. G. Reiman, Zap. Nauchnich Semin. LOMI **77**, 134 (1978) (Leningrad, USSR, in Russian).
[10] S. V. Manakov, Zh. Eksp. Teor. Fiz. **67**, 543 (1974) [Sov. Phys. JETP **40**, 269 (1975)].
[11] M. Ablowitz and J. F. Ladik, J. Math. Phys. **16**, 598 (1978); **17**, 1011 (1976).
[12] F. Kako and N. Mugibayashi, Prog. Theor. Phys. **61**, 776 (1979).
[13] S. C. Chiu and J. F. Ladik, J. Math. Phys. **18**, 690 (1977).
[14] D. Levi and O. Ragnisco, Lett. Nuovo Cimento **22**, 691 (1978); M. Bruschi, D. Levi, and O. Ragnisco, J. Phys. A: Math. Gen. **13**, 2531 (1980).
[15] V. Ju. Novokshenov and I. T. Khabibulin. Dokl. Akad. Nauk SSSR **257**, 543 (1981).
[16] V. S. Gerdjikov, M. I. Ivanov, and P. P. Kulish,, JINR Preprint E2-80-882, Dubna, USSR, 1981.
[17] L. D. Faddeev, Sov. Sci. Rev. Math. Phys. C **1**, 107 (1980).
[18] P. P. Kulish and E. K. Sklyanin, in *Integrable Quantum Field Theories*, edited by J. Hietarinta and C. Montonen, Lecture Notes in Physics (Springer-Verlag, Berlin, 1982), Vol. 151, p. 61.
[19] P. P. Kulish, Lett. Math. Phys.. **5**, 191 (1981).
[20] I. T. Khabibulin, Dokl. Akad. Nauk SSSR **249**, 67 (1979).
[21] V. S. Gerdjikov, JINR Preprint E2-81-652, Dubna, USSR, 1981.
[22] I. M. Gel'fand and L. A. Dickey, Funct. Anal. Appl. **11** (2), 11 (1977) (in Russian).
[23] V. S. Gerdjikov and M. I. Ivanov, JINR Preprint, P5-82-412, Dubna, USSR 1982.

A new integral equation for summing Feynman graph series (general scalar Lagrangian case)

C. Gilain and D. Lévy
Service de Physique Théorique, Division de la Physique, Centre d'Etudes Nucleaires de Saclay, B.P. N° 2, 91190 Gif-Sur-Yvette, France

(Received 30 June 1981; accepted for publication 14 January 1983)

The Schwinger parameter formalism is used to derive a new integral equation verified by the "open" four-point amplitude built from any scalar Lagrangian. This integral equation is a generalization of the one already obtained and studied by the authors in the φ^3 ladder graph case. One of the main results obtained here is a new representation of the Feynman amplitudes: the so-called β-representation, which expresses the Bethe–Salpeter structure of a graph in the Schwinger parameter space. The integrand of the β-representation satisfies a recurrence relation which is used to sum the perturbation series, and which leads to an integral equation for its sum. The expression of this integral equation is also given in some particular cases (particular values of the invariants, particular classes of graphs, etc.). The Mellin transform of the open amplitude satisfies a similar integral equation which may be used to describe the Regge behavior.

PACS numbers: 02.30.Rz, 11.10.Mn, 11.10.Ef

INTRODUCTION

This work takes place in a set of studies whose aim is to obtain, in the framework of Lagrangian field theory, results on the infinite sum of the perturbation series, whatever the value of the coupling constant is. The common feature of this set of studies is that they are performed in the framework of the Schwinger parametrization of Feynman integrals.

Some years ago, powerful results were obtained on the complete asymptotic behavior of each term of the perturbation series (mainly in the Regge limit) for scalar Lagrangians, and on their sum.[1]

Another way, more recently explored, provided results on the four-point amplitude which are not restricted to asymptotic values of the invariants. It relies on the existence of a new integral equation (IE) that does not apply to the amplitude itself, as it is the case for the Bethe–Salpeter (BS) integral equation, but rather to a new quantity: the "open amplitude." The first step has consisted of deriving this new IE in the restricted case of φ^3 ladder subseries.[2] The present work is the generalization of this first step to the complete perturbation series built from any scalar Lagrangian.

The advantages of working with IE are well known: Under conditions of sufficient regularity of the inhomogeneous term and of the kernel, the solution of an IE can be computed. For example, when an IE satisfies the conditions of the Fredholm theorems, its solution is the ratio of two holomorphic functions, and its singularities are poles, given by the zeros of the Fredholm denominator, which depends only on the kernel.

As for the Bethe–Salpeter IE in momentum space, our IE makes use of the Bethe–Salpeter structure of the amplitude, that is to say, its decomposition into generalized ladders whose rungs are t-channel two-particle irreducible subgraphs (t-2PI subgraphs) [see Fig. 1(b)]. The Bethe–Salpeter IE reflects directly the factorization of the integrand when the Feynman amplitude is expressed as an integral over internal 4-momenta.

The Schwinger parametrization of the same amplitude destroys this factorization. For example, the quadratic form $D_G(\alpha)$ which appears in the integrand is a complicated function of all the Schwinger parameters of the graph G. However, the ladder structure of the graph was still reflected, in the φ^3 ladder case, by the open amplitude built in Ref. 2: Inside the set of all integration variables of the Schwinger parametrization [Eq. (1)], we have distinguished there a subset $\alpha_c = \{\alpha_{i_1}, \alpha_{i_2}, \cdots\}$, called the closing variables. The open amplitude $O_G(\alpha_c)$ is then defined by the same integration as the Feynman amplitude I_G itself, except that the closing variable integration is not performed. Of course, the Feynman amplitude of the graph G is the integral of $O_G(\alpha_c)$:

$$I_G = \int d\alpha_c \; O_G(\alpha_c).$$

We have shown that the open amplitude obeys a recurrence law on the number of rungs of the ladder. This recurrence law is the key result from which the existence and the properties of the IE verified by the infinite sum of the open amplitudes is deduced.

We show in the present paper that an analogous work can be done independently of the ladder restriction and for any scalar Lagrangian φ^n.

Although the Bethe–Salpeter IE and our IE, both, reflect the Bethe–Salpeter structure of the amplitude, they are qualitatively different: It is not possible to transform one of them into the other. They concern different amplitudes and different variables:

(i) Our IE is not satisfied by the amplitude, but by the open amplitude.

(ii) The integration variables in the Bethe–Salpeter IE are the external momenta, whereas our IE involves as integration variables the closing variables, i.e., a given subset of the Schwinger parameters.

A consequence of the qualitative difference between the two IE appears in the actual computation: In the φ^3 ladder case, our IE turns to be very appropriate; indeed it provides

 0022-2488/84/010035-13$02.50

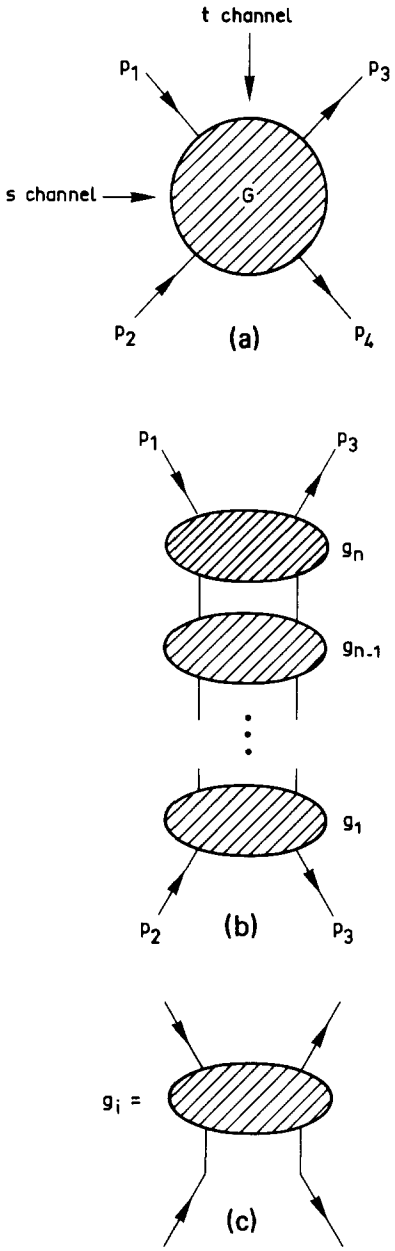

FIG. 1. The kinematics of the four point graph and its Bethe–Salpeter structure: $G = \cup_i g_i$. In the first part of this work, the subgraphs g_i have no special property of reducibility: There is, in general more than one such decomposition of the graph G. In the second part, g_i are restricted to be two particule irreducible in the t channel: There is an only Bethe–Salpeter decomposition of G.

not only the Regge singularities but gives directly the amplitude itself, whereas the Bethe–Salpeter IE has to be studied by two different methods to obtain the same results.[3] From the method initiated by Lee and Sawyer, indeed, the Regge singularity analysis is obtained from an analytic continuation of the partial waves, the problem of the summation of the partial wave expansion, which gives the amplitude, being left over. If one is interested in the amplitude, other methods must be used (such as the perturbation-theoretical integral representation,[3] for instance), and so the complete study of the properties of the amplitude through the Bethe–Salpeter equation is difficult and lengthy.

Though our integral equation is singular, we prove the existence and uniticity of its solution and make explicit its singularity structure. Our fundamental result is that, for each given value of the coupling constant λ, the solution can be written as a finite sum of solutions of Fredholm equations plus a function which is the sum of a convergent series in λ. Moreover, our IE allows simple approximate quantitative computation: For example, the trace approximation gives good results for the dominant trajectory.[4]

To achieve the generalization of our IE, we split the problem into two steps: First we neglect the UV divergences and focus our attention on the algebraic structure. This step is completely done for all scalar Lagrangians (Secs. I–IV). Then we have to face the ultraviolet divergence problem, namely, in our approach the compatibility of the Bergère–Zuber[5] renormalization procedure and of the structure of the recurrence relation [see Sec. I A, Eqs. (32)]. This is done here only for the φ^3 interaction.

The price to pay for the generality of our result is, of course, the formal character of the equation obtained. The kernel, which governs the properties of the solution, is given in terms of an infinite series. The logical following step of our program is the link between the properties of this series and those of the four-point amplitude.

We conclude this introduction with a more precise presentation of the content of this paper. We obtain two new results: the first one, presented in Sec. I, is a scalar integral representation for the Feynman amplitudes, which is an alternative to the Schwinger one. The Schwinger α-parametrization gives in fact the amplitude associated with a given graph as a multiple scalar integral involving as many scalar variables as internal lines in the graph. There appears in the integrand no factorization according to the "rungs" of the generalized ladder [see Fig. 1(b)]. Our aim is to make explicit on the Schwinger integrand the BS structure of a graph. This requires, as presented in subsection I A, a change in the choice of the invariants and consequently of the topological functions which are their coefficients in the quadratic from $D_G(\alpha)$. In subsection I B, important properties of quasifactorization and of recurrence of these topological functions of the graphs are given. In subsection I C, the structure of the quadratic form $D_G(\alpha)$ is made precise. In subsection I D, we are then led to establish our alternative parametrization for the Feynman amplitude: the β-parametrization. Let us consider the graph G of Fig. 1(b), which is a generalized ladder with n rungs g_i. The β-parametrization is an integral over $6 \times n$ scalar variables (the β variables) and its integrand is a product of two factors:

(1) The first one is completely factorized, and is a product of n functions of six variables, each one being attached to one rung g_i.

(2) The other one is a global factor, which depends only on n and is independent of the structure of each rung: It is the skeleton of BS structure of the graph.

The β variables represent appropriate combinations of the topological polynomials associated with each g_i. Their variation domains are always explicitly indicated by means of θ step functions.

We have then the adequate tools for proving the existence of the integral equation, which is the second new result

of this work. It is the aim of Sec. II. From the recurrence relation obeyed by the open amplitudes (defined in II A) we deduce the integral equation satisfied by their sum (II B and II C).

Then we discuss in this framework the Regge limit, that is to say, the structure of the IE in the Mellin space (Sec. III).

Some physically interesting particular situations are grouped in the fourth section: forward scattering, bound states equation,... . In this section one can find also the simplified expression of the IE for a special class of kernels (the ladder with generalized rungs; see Fig. 4), or for particular values of the variables β.

Finally, the renormalization problem is achieved for the φ^3 interaction Lagrangian in Sec. V. Some technical points are grouped in the Appendix.

I. BETHE–SALPETER STRUCTURE AND TOPOLOGICAL POLYNOMIALS

We consider here the scalar Lagrangian field theories. With any graph G is associated its Feynman amplitude, whose Schwinger integral representation is

$$I_G^{\epsilon}(P) = \lambda^{N(G)}(ie^{-i\epsilon})^{-\omega(G)/2}\int_0^{\infty}\prod_{a=1}^{l(G)} d\alpha_a \times\exp\left(-ie^{-i\epsilon}\sum_{a=1}^{l(G)}\alpha_a m^2\right) \times R\left(\frac{e^{ie^{-i\epsilon}D_G(\alpha)}}{P_G^2(\alpha)}\right). \tag{1}$$

In (1), $\omega(G)$ is the superficial degree of divergence of the graph G:

$$\omega(G) = 4L(G) - 2l(G),$$

where $L(G)$, $l(G)$, and $N(G)$ are, respectively, the number of independent loops, of internal lines, and of vertices of G. P is the set of external 4-momenta, and λ is the coupling constant of the theory. There is a scalar variable α_a attached to each internal line of the graph. The set $\{\alpha_1,\alpha_2,...,\alpha_{l(G)}\}$ will be noted α or α_G every time an ambiguity is possible. The operator R is the Bergère–Zuber[5] subtraction operator which ensures the ultraviolet (UV) convergence of the Feynman amplitude. In this work we will pay no attention to the UV convergence problems, but for the case of the interaction φ^3 which we treat exhaustively (see Sec. V).

In Minkowsky space the amplitude is the limit $\epsilon\to 0_+$ of I_G^{ϵ}. As we are mainly interested with the algebraic structure of the integrand, and not with the convergence conditions of the integral, we place our problem in Euclidean space, in which the amplitude is given from (1) with $\epsilon = \pi/2$:

$$I_G(P) = \int_0^{\infty} d\mu_G(\alpha_G) e^{D_G(\alpha)}, \tag{1'}$$

where

$$d\mu_G(\alpha_G) = \lambda^{N(G)}\prod_{a=1}^{l(G)} d\alpha_a \frac{\exp(-\Sigma_{a=1}^{l(G)}\alpha_a m^2)}{P_G^2(\alpha)}. \tag{2}$$

The function $D_G(\alpha)$ is a quadratic form built from the external 4-momenta. In 2 particles $\to$2 particles case which we are studying, it is equal to

$$D_G(\alpha) = s\frac{A_G^s(\alpha)}{P_G(\alpha)} + t\frac{A_G^t(\alpha)}{P_G(\alpha)} + u\frac{A_G^u(\alpha)}{P_G(\alpha)} + \sum_{i=1}^{4} p_i^2\frac{A_G^i(\alpha)}{P_G(\alpha)}. \tag{3}$$

s,t,u are the Mandelstam invariants built from the external momenta p_i [see Fig. 1(a)], and $P_G(\alpha)$, $A_G^s(\alpha)$, $A_G^t(\alpha)$, $A_G^u(\alpha)$, $A_G^i(\alpha)$ $(i = 1,...,4)$ are the topological polynomials, characteristic of the graph G. Their definition can be found in the Appendix A of Ref. 2. Let us only say that they are polynomials, homogeneous in the set α, and of degree $L(G)$ for P_G, $(L(G)+1)$ for the other ones.

The problem we solve here is the adaptation of this formalism in order to make use of the Bethe–Salpeter structure of the four-point amplitude: Any graph composed of at least n two-particle irreducible subgraphs in the t channel may be drawn as the generalized ladder of Fig. 1(b). As we consider the two vertical lines attached under each bubble as internal lines of the corresponding subgraph, the graph G is exactly the union of each subgraph g_i:

$$G = \{g_1,...,g_n\}.$$

In this first section, except for the existence of the two additional vertical lines, the graphs g_i can have absolutely any structure: They can be reducible or irreducible.

The problem stands of course in the fact that the integrand $e^{D_G(\alpha)}/P_G^2(\alpha)$ in (1') is not factorized in functions, each attached to each subgraph g_i. As we want to build G as a ladder of rungs g_i, we are faced with the necessity of performing loop integrals to link two subgraphs: In the following paragraph a change of external momenta is performed in order to make easier this integration.

A. Alternative expression for the quadratic form $D_G(\alpha)$

The first step consists in modifying the usual form of $D_G(\alpha)$. We choose as external momenta the three combinations,

$$\begin{aligned} q_1 &= \tfrac{1}{2}(p_1+p_3),\\ q_2 &= \tfrac{1}{2}(p_2+p_4),\\ q &= (p_1-p_3) = (p_4-p_2), \end{aligned} \tag{4}$$

and build their associated invariants,

$$\begin{aligned} s_{11} &= q_1^2 = \tfrac{1}{2}(p_1^2+p_3^2) - \tfrac{1}{4}t,\\ s_{12} &= 2q_1q_2 = \tfrac{1}{2}(s-u) = s + \frac{1}{2}\left(t - \sum_{i=1}^{4} p_i^2\right),\\ s_{22} &= q_2^2 = \tfrac{1}{2}(p_2^2+p_4^2) - \tfrac{1}{4}t,\\ s_1 &= 2qq_1 = p_1^2 - p_3^2,\\ s_2 &= 2qq_2 = p_4^2 - p_2^2,\\ s_t &= q^2 = t. \end{aligned} \tag{5}$$

The seven invariants s,t,u,p_i^2, $i = 1,2,3,4$, are not independent $(s+t+u = \Sigma p_i^2)$, so it is enough to define the six independent invariants s_j, $j\in K$, where K is the set of indices:

$$K = \{11,12,22,1,2,t\}.$$

Putting in (3) the inverse relations of (5), which gives the Mandelstam invariants in terms of the s_j variables, we find

$$D_G(\alpha) = \sum_{j\in K} s_j \beta^j_G(\alpha). \tag{6}$$

The $\beta^j_G(\alpha)$ are the combinations of topological polynomials associated with s_j:

$$\begin{aligned}
\beta^{11}_G(\alpha) &= [1/P(\alpha)][A^s_G(\alpha) + A^u_G(\alpha) + A^1_G(\alpha) + A^3_G(\alpha)],\\
\beta^{12}_G(\alpha) &= [1/P_G(\alpha)][A^s_G(\alpha) - A^u_G(\alpha)],\\
\beta^{22}_G(\alpha) &= [1/P_G(\alpha)][A^s_G(\alpha) + A^u_G(\alpha) + A^2_G(\alpha) + A^4_G(\alpha)],\\
\beta^1_G(\alpha) &= \tfrac{1}{2}[1/P_G(\alpha)][A^1_G(\alpha) - A^3_G(\alpha)],\\
\beta^2_G(\alpha) &= \tfrac{1}{2}[1/P_G(\alpha)][A^4_G(\alpha) - A^2_G(\alpha)],\\
\beta^t_G(\alpha) &= [1/P_G(\alpha)]\{A^t_G(\alpha)\\
&\quad + \tfrac{1}{4}[A^1_G(\alpha) + A^2_G(\alpha) + A^3_G(\alpha) + A^4_G(\alpha)]\}.
\end{aligned} \tag{7}$$

The set of the six functions β^j_G, $j\in K$, will be noted β_G.

Let us now give the variation domain of β_G, when the α parameters vary from zero to infinity. For the most general graph, the topological functions $A^j_G(\alpha)/P_G(\alpha)$, $j = s,t,u,1,2,3,4$ are independent and vary from zero to plus infinity. Then, using (7), one obtains the bounded domain:

$$|\beta^{12}_G| + 2|\beta^1_G| \leqslant \beta^{11}_G, \tag{8a}$$

$$|\beta^{12}_G| + 2|\beta^2_G| \leqslant \beta^{22}_G, \tag{8b}$$

$$|\beta^1_G| + |\beta^2_G| \leqslant 2\beta^t_G. \tag{8c}$$

In opposition with $A^j_G(\alpha)/P_G(\alpha)$, some of the β^j_G may become negative.

In fact, we will see in the following that the six β^j, $j\in K$, do not play an equivalent role: We have to group them into two sets:

$$\gamma = \{\beta^{12}, \beta^2, \beta^{22}\} \tag{9a}$$

and

$$\bar\gamma = \{\beta^{11}, \beta^1, \beta^t\}. \tag{9b}$$

Thus, the variation domain (8) may be built in two steps: the variation domain of $\bar\gamma$, γ being kept fixed and the domain for γ, whatever $\bar\gamma$ is. These variation domains play an important role in the following. To each of them are attached, respectively, the function Θ_1, Θ_2, Θ_3 with

$$\Theta_1(\beta) = \Theta_2\cdot\Theta_3, \tag{10a}$$

where

$$\Theta_2(\gamma,\bar\gamma) = \theta(\beta^{11} - |\beta^{12}| - 2|\beta^1|)\cdot\theta(2\beta^t - |\beta^1| - |\beta^2|), \tag{10b}$$

$$\Theta_3(\gamma) = \theta(\beta^{22} - |\beta^{12}| - 2|\beta^2|), \tag{10c}$$

with θ the usual step function.

B. Bethe–Salpeter structure of the β functions

The theorem we establish now concerns the Bethe–Salpeter structure of the topological polynomials. It is the result upon which the whole work relies.

Theorem 1: Let us consider a graph G which can be written as a generalized ladder with n rungs [Fig. 1(b)]:

$$G = \{g_1, g_2, \ldots, g_n\}.$$

Then there exists seven functions of $6\times n$ variables, S^j_n, $j\in K$, and S^0_n, verifying the three following properties:

—They are independent of the graph g_i, depending only on the number n of such subgraphs,

$$-\beta^j_G(\alpha_G) = S^j_n(\beta_{g_1}(\alpha_{g_1}), \beta_{g_2}(\alpha_{g_2}), \ldots, \beta_{g_n}(\alpha_{g_n})), \tag{11}$$

and

$$P_G(\alpha_G) = \left(\prod_{i=1}^n P_{g_i}(\alpha_{g_i})\right) S^0_n(\beta_{g_1}(\alpha_{g_1}), \ldots, \beta_{g_n}(\alpha_{g_n})). \tag{12}$$

—The functions S^j_n, $j\in K$, and S^0_n verify the following recurrence relations:

$$S^j_n(\beta_1,\ldots,\beta_n) = S^j_{n-1}(\beta_1,\ldots,\beta_{n-2}, S_2(\beta_{n-1},\beta_n)), \tag{13}$$

$$S^0_n(\beta_1,\ldots,\beta_n) = S^0_{n-1}(\beta_1,\ldots,\beta_{n-2}, S_2(\beta_{n-1},\beta_n)) S^0_2(\beta_{n-1},\beta_n). \tag{14}$$

The meaning of this theorem is the following: The β functions associated with the graph are themselves functions of the β functions associated with each subgraph g_i in a way which is independent of the graph G except for the number of subgraphs g_i.

It is this property which replaces the factorization property of the integrand in the momentum space.

Proof: The proof proceeds through two stages: first we show it directly for the case $n = 2$. Then the proof works by recurrence.

$n = 2$ case: Let us consider a graph G which is two-particule reducible in the t channel (see Fig. 2): $G = \{g_1, g_2\}$.

We write the amplitude I_G in terms of the convolution of the two amplitudes I_{g_1} and I_{g_2}:

$$I_G(q_1,q_2,q) = cst\int d^4q'\, I_{g_1}(-q',q_2,q)\cdot I_{g_2}(q_1,q',q), \tag{15}$$

where

$$q' = \tfrac{1}{2}(p'_4 + p'_2).$$

In the two members of Eq. (15), we use for I the expression (1′), where $D(\alpha)$ is given by (6). After having done the integration over q', we can identify on the two sides the denominators and the coefficients of the invariants. We remark that β_G depends on α_{g_1} and α_{g_2} only through β_{g_1} and β_{g_2}. We thus obtain explicitly the functions S_2:

$$\begin{aligned}
S^{11}_2(\beta_1,\beta_2) &= \beta^{11}_2 - (\beta^{12}_2)^2/(\beta^{11}_1 + \beta^{22}_2),\\
S^{12}_2(\beta_1,\beta_2) &= \beta^{12}_1\beta^{12}_2/(\beta^{11}_1 + \beta^{12}_2),\\
S^{22}_2(\beta_1,\beta_2) &= \beta^{22}_1 - (\beta^{12}_1)^2/(\beta^{11}_1 + \beta^{22}_2),\\
S^1_2(\beta_1,\beta_2) &= \beta^1_2 - \beta^{12}_2(\beta^2_2 - \beta^1_1)/(\beta^{11}_1 + \beta^{22}_2),\\
S^2_2(\beta_1,\beta_2) &= \beta^2_1 + \beta^{12}_1(\beta^2_2 - \beta^1_1)/(\beta^{11}_1 + \beta^{22}_2),\\
S^t_2(\beta_1,\beta_2) &= \beta^t_1 + \beta^t_2 - (\beta^2_2 - \beta^1_1)^2/(\beta^{11}_1 + \beta^{22}_2),
\end{aligned} \tag{16}$$

and finally

$$S^0_2(\beta_1,\beta_2) = \beta^{11}_1 + \beta^{22}_2. \tag{17}$$

n subgraph case: Let us turn now to the graph of Fig. 1(b). We build by recurrence the set of functions S^j_n, $j\in K$ [see (13)]. Inside the graph G we can group together the two last subgraphs g_{n-1} and g_n:

$$G = \{g_1, g_2, \ldots, g_{n-2}, g'_{n-1}\},$$

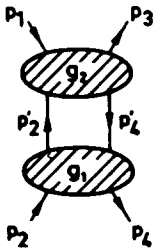

FIG. 2. The graph $G = \{ g_1, g_2 \}$.

with

$$g'_{n-1} = \{ g_{n-1}, g_n \}.$$

If we assume that the S^j_{n-1} functions are known, we have

$$\beta^j_G(\alpha_G) = S^j_{n-1}(\beta_{g_1}(\alpha_{g_1}),..., \beta_{g_{n-2}}(\alpha_{g_{n-2}}), \beta_{g'_{n-1}}(\alpha_{g'_{n-1}})).$$

then using Eq. (16) to compute $\beta_{g'_{n-1}}$, we obtain

$$\beta^j_G(\alpha_G) = S^j_{n-1}(\beta_{g_1}(\alpha_{g_1}),..., \beta_{g_{n-2}}(\alpha_{g_{n-2}}), S_2(\beta_{g_{n-1}}(\alpha_{g_{n-1}}), \beta_{g_n}(\alpha_{g_n}))). \tag{18}$$

The comparison of (18) with (11) proves the existence of S^j_n ($j \in K$) and gives us their recurrence law. With the same procedure, we deduce the recurrence law (14).

This achieves the proof of Theorem 1.

C. Bethe–Salpeter structure of the quadratic form

In this subsection, the dependence of the quadratic form in function of the variables $\bar{\gamma}_n$ is studied. In the recurrence relations (13) and (14), the six variables β_n do not play an equivalent role. The dependence in function of three of them ($\bar{\gamma}_n$) is linear and does not depend on all the $6 \times n - 3$ other variables. It will be seen further that this property allows to obtain an IE with only three integration variables and not six. To lighten the notations, we write $\beta_{(n)} = \{ \beta_1,..., \beta_n \}$. The functions $S^j_n, j \in K$, are homogeneous functions of degree one in the set of the $6 \times n$ variables $\beta_{(n)}$, and S^0_n is homogeneous of degree $(n-1)$ in the same set. We recall that we have defined the two subsets [see (9a) and (9b)]: $\gamma = \{ \beta^{12}, \beta^{22}, \beta^2 \}$ and $\bar{\gamma} = \{ \beta^{11}, \beta^1, \beta^t \}$; we define also the two subsets of indices:

$$K' = \{12,22,2\} \quad \text{and} \quad \bar{K}' = \{11,1,t\}.$$

From (16) and (13), one can show by recurrence that it is possible to define a set of functions $\hat{S}^j_n$ such that

$$\begin{aligned} S^j_n(\beta_{(n)}) &= \hat{S}^j_n(\beta_{(n-1)}, \gamma_n), \quad \text{for } j \in K' \text{ and } j = 0, \\ S^j_n(\beta_{(n)}) &= \hat{S}^j_n(\beta_{(n-1)}, \gamma_n) + \beta^j_n, \quad \text{for } j \in \bar{K}'. \end{aligned} \tag{19}$$

As a direct consequence of Theorem 1, we are led to define a function D_n:

$$D_n(\beta_1,..., \beta_n) \equiv \sum_{j \in K} s_j S^j_n(\beta_1,..., \beta_n). \tag{20}$$

The quadratic form $D_G(\alpha_G)$ has a simple expression in function of D_n:

$$D_G(\alpha_G) = D_n(\beta_{g_1}(\alpha_{g_1}), \beta_{g_2}(\alpha_{g_2}),..., \beta_{g_n}(\alpha_{g_n})). \tag{21}$$

The useful properties of D_n are given in the following theorem.

Theorem 2: The dependence of the D_n function upon the three variables $\bar{\gamma}_n$ of the last graph g_n is explicit and linear:

$$D_n(\beta_{(n)}) = \hat{D}_n(\beta_{(n-1)}, \gamma_n) + \sum_{j \in \bar{K}'} s_j \beta^j_n, \tag{22}$$

where the $\hat{D}_n$ function depends, as far as the last subgraph is concerned, only on the set γ_n. The $\hat{D}_n$ function verifies the recurrence law:

$$\begin{aligned} &\hat{D}_n(\beta_{(n-1)}, \gamma_n) \\ &\quad = \hat{D}_{n-1}(\beta_{(n-2)}, \hat{S}_2(\beta_{n-1}, \gamma_n)) + d(\bar{\gamma}_{n-1}, \gamma_n), \end{aligned} \tag{23a}$$

where

$$\begin{aligned} d(\bar{\gamma}_1, \bar{\gamma}_2) = &- s_{11} \frac{(\beta^{12}_2)^2}{\beta^{11}_1 + \beta^{22}_2} - s_1 \frac{\beta^{12}_2 (\beta^2_2 - \beta^1_1)}{\beta^{11}_1 + \beta^{22}_2} \\ &+ s_t \left(\beta^t_1 - \frac{(\beta^2_2 - \beta^1_1)^2}{\beta^{11}_1 + \beta^{22}_2} \right) \end{aligned} \tag{23b}$$

and where $\hat{S}_2$ represents the set of functions $\{ \hat{S}^j_2, j \in K' \}$.

Proof: The relation (22) follows immediately from Eqs. (19) and (20). The function D_n and the term $\Sigma_{j \in K'} s_j \beta^j_n$, follow the same recurrence law (13) as S^j_n. Thus, using (22), one can obtain the recurrence law (23) for $\hat{D}_n$.

Let us remark that the function d, and the term $(\Sigma_{j \in \bar{K}'} s_j \beta^j_n)$ in (22) correspond exactly to the violation of the * law in the framework of our work on the φ^3 ladder.[2]

D. β-parametrization of the Feynman amplitudes

We are now able to proceed any further and to propose an alternative form for the Schwinger parametrization, form which reflects the Bethe Salpeter structure of the amplitude:

Theorem 3: The amplitude I_G attached to the graph of Fig. 1(b) may be written as

$$I_G = \int \prod_{i=1}^{n} (d\beta_i j_{g_i}(\beta_i)) \frac{e^{D_n(\beta_1,..., \beta_n)}}{[S^0_n(\beta_1,..., \beta_n)]^2}, \tag{24}$$

where

$$j_g(\beta) = \Theta_1(\beta) \int_0^\infty d\mu_g(\alpha_g) \prod_{j \in \bar{K}} \delta(\beta^j - \beta^j_g(\alpha_g)). \tag{25}$$

Proof: Theorem 3 is easily proved if, inside expression (1') where $D_G(\alpha)$ is given by Eq. (21), we insert

$$1 = \int \prod_{i=1}^{n} \prod_{j \in \bar{K}} \delta(\beta^j_i - \beta^j_{g_i}(\alpha_{g_i})) \, d\beta^j_i. \tag{26}$$

Let us make three remarks:

(1) We purposely make explicit the integration region for the β via the factor Θ_1 [see Eq. (10)].

(2) For a given graph G, the decomposition $G = \{ g_1,..., g_n \}$ is not unique, as far as the irreducibility of the subgraphs g_i is not required. In particular, to any graph is associated its β-parametrization with $n = 1$:

$$I_G = \int d\beta j_G(\beta) e^{D_1(\beta)}.$$

(3) The strength of the expression (24) is that the integrand appears as the product of two qualitatively different factors:

(a) $e^{D_n(\beta_{(n)})}/[S_n^0(\beta_{(n)})]^2$ is independent of the characteristics of the graph G but the number n of subgraphs g_i.

(b) The n functions $j_{g_i}(\beta_i)$ depend on the subgraphs g_i.

This factorized structure is the main property which is used to build the integral equation derived in the next section.

II. INTEGRAL EQUATION

The Bethe–Salpeter integral equation is written for the amplitude. It is not the case here. Our work relies upon the properties of the partially integrated integrand. The first subsection is devoted to define this "open amplitude." Then a first form of the integral is given. The third subsection gives the final form of this equation.

From now on and up to the end of the work we consider for each graph its unique decomposition in the generalized ladder [see Fig. 1(b)] of t-2PI subgraphs: Here the notation g_i will always refer to such a two-particle irreducible subgraph.

A. The recurrence relation obeyed by the open amplitude

The open amplitude $O_{G_{n-1}}(\gamma_n)$ is defined by the relation

$$O_{G_{n-1}}(\gamma_n)=\int\prod_{i=1}^{n-1}[d\beta_i\, j_{g_i}(\beta_i)]\,\frac{e^{\hat{D}_n(\gamma_{(n)})}}{[S_n^0(\gamma_{(n)})]^2}, \tag{27}$$

where $\gamma_{(n)}$ is a condensed notation:

$$\gamma_{(n)}=\{\beta_{(n-1)},\gamma_n\}. \tag{28}$$

Equation (27) is nothing but Eq. (24) where we let remain the last six integrations $d\beta_n$:

$$I_{G_n}=\int d\beta_n\, j_{g_n}(\beta_n)\exp\Big(\sum_{j\in\bar{K}'}s_j\,\beta_n^j\Big)O_{G_{n-1}}(\gamma_n). \tag{29}$$

The open amplitude is only dependent on the $(n-1)$ subgraphs $G_{n-1}=\{g_1,\ldots,g_{n-1}\}$, and not on the last subgraph g_n. From a given open amplitude $O_{G_{n-1}}$, it is possible, using (29), to reconstruct all the amplitudes of the family of graphs G_n which have the sames $(n-1)$ first subgraphs and a different nth subgraph g_n. Such graphs, which are generalized ladder with n rungs, but with only the $(n-1)$ first subgraphs G_{n-1} specified, will be called n-open graphs (see Fig. 3).

The integration (29) can be simplified: Inside the set of the six variables β_n, the three integrations $d\bar{\gamma}_n$ can be performed:

$$I_{G_n}=\int d\gamma_n\,\bar{j}_{g_n}(\gamma_n)O_{G_{n-1}}(\gamma_n) \tag{30}$$

with

$$\bar{j}_g(\gamma)=\int d\bar{\gamma}\, j_g(\beta)\exp\Big(\sum_{j\in\bar{K}'}s_j\,\beta^j\Big).$$

Inserting definition (25) for $j_g(\beta)$, we find

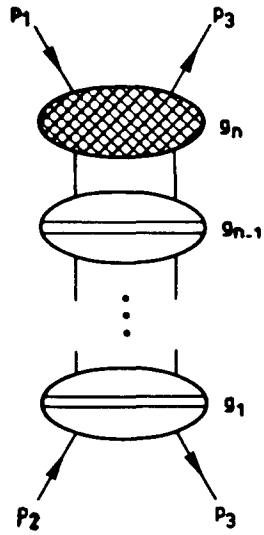

FIG. 3. The n-open graph $G_{n-1}=\{g_1,\ldots,g_{n-1}\}$. An n-open graph, and the open amplitude which is associated with it, depend on the $(n-1)$ subgraphs g_i, $i=1,\ldots,n-1$, but not on the nth subgraph. The nth bubble is a skeleton which may be dressed by any graph g_n.

$$\bar{j}_g(\gamma)=\Theta_3(\gamma)\int d\mu_g(\alpha_g)\exp\Big(\sum_{j\in\bar{K}'}s_j\,\beta_g^j(\alpha_g)\Big)$$
$$\times\prod_{j\in\bar{K}'}\delta(\beta^j-\beta_g^j(\alpha_g)). \tag{31}$$

The way to build the recurrence relation on the number of subgraphs of the open amplitude is straightforward: In the expression $O_{G_n}(\gamma_{n+1})$, we make use of the recurrence relations (14) and (23).

We recognize the open amplitude $O_{G_{n-1}}(\hat{S}_2(\beta_n,\gamma_{n+1}))$ in the integrand and so

$$O_{G_n}(\gamma_{n+1})=\int d\beta_n\, j_{g_n}(\beta_n)\,\frac{e^{d(\bar{\gamma}_n,\gamma_{n+1})}}{[\hat{S}_2^0(\beta_n,\gamma_{n+1})]^2}$$
$$\times O_{G_{n-1}}(\hat{S}_2(\beta_n,\gamma_{n+1})). \tag{32}$$

In the relation (32), only the variables with an index equal to n or $n+1$ appear. Thus the notations can be simplified: instead of $\beta_n=\{\gamma_n,\bar{\gamma}_n\}$ and $\beta_{n+1}=\{\gamma_{n+1},\bar{\gamma}_{n+1}\}$, we will use in the remainder $\beta'=\{\gamma',\bar{\gamma}'\}$ and $\beta=\{\gamma,\bar{\gamma}\}$.

Let us remark that the open amplitude has been defined in perfect analogy with the φ^3 ladder case.[2] We recall that, in this latter work, the closing variables (see the Introduction) were the three Schwinger parameters attached to the last rung and the last vertical lines of the ladder, whose correspondent here is exactly the last subgraph g_n [see Figs. 1(b) and (c)]. One can get convinced that the elements of γ_n concern the same topological polynomials of g_n that the closing variables of the ladder.

B. Summing the series over all graphs

As we already said, we may draw all graphs generated by any scalar Lagrangian as a generalized ladder (see Fig. 1(b)], where each subgraph g_i is t-2PI. Now, the crucial point when one wants to face the whole perturbation is to organize the infinite sum. The results already obtained [the factorization of Eq. (24) on one hand, the definition of the open amplitude on the other] lead us to the following four steps:

(i) For any n-open graph, we define its open amplitude $O_{G_{n-1}}$.

(ii) We group together all the n-open graphs and define the quantity

$$O_n(\gamma) = \sum_{G_{n-1}} O_{G_{n-1}}(\gamma). \tag{33}$$

From Eqs. (32) and (33), we see that $O_n(\gamma)$ verifies the recurrence relation

$$O_n(\gamma) = \int d\beta'\, k(\gamma,\beta')O_{n-1}(\hat{S}_2(\beta',\gamma)), \tag{34}$$

with

$$k(\gamma,\beta') = \sum_g k_g(\gamma,\beta'), \tag{35a}$$

where Σ_g is the sum over all the t-2PI graphs and where

$$k_g(\gamma,\beta') = j_g(\beta')\frac{e^{d(\bar{\gamma}',\gamma)}}{[\hat{S}_2^0(\beta',\gamma)]^2}. \tag{35b}$$

(iii) The following step consists in summing over each such set of graphs; we define

$$O(\gamma) = \sum_{n=1}^{\infty} O_n(\gamma). \tag{36}$$

then $O(\gamma)$ verifies the integral equation

$$O(\gamma) = O_1(\gamma) + \int d\beta'\, k(\gamma,\beta')O(\hat{S}_2(\beta',\gamma)). \tag{37}$$

with

$$O_1(\gamma) = e^{\hat{D}_1(\gamma)} = e^{s_{12}\beta^{12} + s_{22}\beta^{22} + s_2\beta^2}. \tag{38}$$

Equation (37) is essentially the integral equation we are looking for.

(iv) The last step consists of performing the integration on the variables γ in order to get the four-point amplitude I:

$$I = \int d\gamma\, \bar{j}(\gamma)O(\gamma), \tag{39}$$

where

$$\bar{j}(\gamma) = \sum_g \bar{j}_g(\gamma) \tag{40}$$

with $\bar{j}_g$ given by (31).

C. Final form for the integral equation

We will now proceed a little further in order to get the integral equation verified by $O(\gamma)$ in a more classical form, and see whether it falls under the scope of classical theorems.

We define the change of variables

$$\gamma' \to \gamma^* \tag{41}$$

such that

$$\begin{aligned}
&\beta^{12'} \to \beta^{12*} = \hat{S}_2^{12}(\beta',\gamma) = \beta^{12'}\beta^{12}/(\beta^{11'} + \beta^{22}),\\
&\beta^{22'} \to \beta^{22*} = \hat{S}_2^{22}(\beta',\gamma) = \beta^{22'} - (\beta^{22'})^2/(\beta^{11'} + \beta^{22}),\\
&\beta^{2'} \to \beta^{2*} = \hat{S}_2^{2}(\beta',\gamma)\\
&\qquad = \beta^{2'} + \beta^{12'}(\beta^2 - \beta^{1'})/(\beta^{11'} + \beta^{22}).
\end{aligned} \tag{42}$$

This change of variables does not concern the variables $\bar{\gamma}'$.

We define

$$u = \beta^{12*}/\beta^{12} = \beta^{12'}/(\beta^{11'} + \beta^{22}). \tag{43}$$

One can immediately see that the variation domain of γ^* is at most as large as the domain defined by $\Theta_3(\gamma^*)$, as γ^* is nothing but the γ variables of the graph of Fig. 2 with $\{g_1, g_2\} = \{g', g\}$.

The computation of the actual variation domain of γ^* is given in Appendix A. It is given by the following function:

$$\Theta_4(\gamma,\gamma^*) = \theta(1 - |u|)\theta(\beta^{22*} - |u|\beta^{22} - 2|\beta^{2*} - u\beta^2|). \tag{44}$$

The Jacobian of the transformation (41) is

$$J(\gamma' \to \gamma^*) = (\beta^{11'} + \beta^{22})/\beta^{12}. \tag{45}$$

Among the six integrations of the integral (37), three can be done explicitly and a new kernel K can be defined by

$$K(\gamma,\gamma^*) = \sum_g K_g(\gamma,\gamma^*) \tag{46}$$

with

$$K_g(\gamma,\gamma^*) = \int d\bar{\gamma}'\, k_g(\gamma,\beta')J(\gamma' \to \gamma^*).$$

Using (35b) and the expression (25) of j_g, one obtains

$$\begin{aligned}
K_g(\gamma,\gamma^*) = \Theta_4(\gamma,\gamma^*) &\int d\nu_g(\alpha_g)\exp[d(\delta\bar{\gamma}(\alpha_g),\gamma)]\\
&\times \prod_{j\in K'} \delta(\beta^{j*} - \hat{S}_2^j(\beta_g(\alpha_g),\gamma))
\end{aligned} \tag{47}$$

with

$$d\nu_g(\alpha_g) = \frac{d\mu_g(\alpha_g)}{[\hat{S}_2^0(\beta(\alpha_g),\gamma)]^2} = \frac{d\mu_g(\alpha_g)}{[\delta^{11}(\alpha_g) + \delta^{22}]^2}.$$

We finally have

$$O(\gamma) = O_1(\gamma) + \int d\gamma^*\, K(\gamma,\gamma^*)O(\gamma^*), \tag{48}$$

with $O_1(\gamma)$ given by (38) and $K(\gamma,\gamma^*)$ by Eqs. (46) and (47). Of course, we obtain the amplitude 1 from Eq. (39).

Let us make three last remarks about the IE:

—The dependence of $O(\gamma)$ as function of $s_j, j\in K$, has two sources: $O_1(\gamma)$ depends upon s_{12}, s_{22}, and s_2, and the kernel K depends on the three other invariants s_{11}, s_1, and s_t.

—Whereas the number of integration variables was six in the IE (37), it is only three in (48). This difference reflects exactly the difference between the recurrence relation verified by S_n and D_n and which concerns six variables [see (13)], and the one verified by $\hat{D}_n$, where only the three variables γ_n are concerned.

—The inhomogeneous term $O_1(\gamma)$ is a simple explicit function [Eq. (38)] which is independent of the Lagrangian.

III. MELLIN TRANSFORM AND REGGE POLES

A. The integral equation verified by the open amplitude of the Mellin transform

The reasons for working with the Mellin transform of the amplitudes are of two different types:

—First, there are technical reasons which are linked to the Wick rotation problem and to the Landau singularities. These points have been discussed in Ref. 2 for the φ^3 ladder case, and we shall not come back to it in the present paper.

—On the other hand, it is well known that the Mellin

transform is very well adapted for the study of the amplitude at high energy, where the Regge model is relevant. The singularities of the Mellin transform are linked to the Regge singularities in the angular momentum space.

In term of the invariants $s_j, j\in K$, the Regge limit is defined by

$$s_{12}\to\infty, \quad s_j = cst \quad \text{for } j\neq 12;$$

so we are going to perform the Mellin transform of the amplitude for the variable s_{12}. The Mellin transform $\bar f(x)$ of a function $f(s)$, which is integrable and regular when s goes to zero, is defined by

$$e^{-i\pi x}\Gamma(-x)\bar f(x) = \int_0^\infty ds\, s^{-x-1} f(s)$$

$$\text{for } -1 < \mathrm{Re}(x) < 0. \tag{49}$$

For the values of x where the integral (49) does not exist, $\bar f(x)$ can be defined by analytic continuation. If one uses the β-representation (24) of I_G, it is possible to perform the integration (49) over the variable s_{12} explicitly, and one obtains

$$\bar I_G(x) = \int \prod_{i=1}^{n} [d\beta_i\, j_{g_i}(\beta_i)] \times \frac{[S_n^{12}(\beta_1,\ldots,\beta_n)]^x}{[S_n^0(\beta_1,\ldots,\beta_n)]^2} \times \exp\left(\sum_{\substack{j\in K\\ j\neq 12}} s_j S_n^j(\beta_1,\ldots,\beta_n)\right).$$

Using the relations (13) and (16), it can be shown that

$$S_n^{12}(\beta_1,\ldots,\beta_n) = \frac{\Pi_{i=1}^n \beta_i^{12}}{S_n^0(\beta_1,\ldots,\beta_n)};$$

the Mellin transform $\bar I_G(x)$ becomes

$$\bar I_G(x) = \int \prod_{i=1}^{n} [d\beta_i\, j_{g_i}(\beta_i)(\beta_i^{12})^x] \times \frac{1}{[S_n^0(\beta_1,\ldots,\beta_n)]^{x+2}} \exp\left(\sum_{\substack{j\in K\\ j\neq 12}} s_j S_n^j(\beta_1,\ldots,\beta_n)\right). \tag{50}$$

The factors $(\beta_i^{12})^x$ and $[S_n^0(\beta_1,\ldots,\beta_n)]^{-x}$ introduce no new singularity in the integral when $x > -1$, and $\bar I_G(x)$ is defined when I_G is defined.

The open amplitude of the Mellin transform is a function of three variables $\gamma = (\gamma^{12},\gamma^{22},\gamma^2)$ defined by suppressing in (50) the last integration $d\beta_n$ and the factor $j_{g_n}(\beta_n)(\beta_n^{12})^x \exp(\Sigma_{j\in\bar K} s_j \beta_n^j)$, which depends only on the variables β_n:

$$\bar O_{G_{n-1}}(\gamma_n) = \int \prod_{i=1}^{n} [d\beta_i\, j_{g_i}(\beta_i)(\beta_i^{12})^x] \times \frac{1}{[S_n^0(\beta_1,\ldots,\beta_n)]^{x+2}} \times \exp\left(\sum_{\substack{i\in K\\ i\neq 12}} s_i \bar S_n^i(\beta_1,\ldots,\beta_n)\right).$$

The important property of the function $\bar O_{G_n}(\gamma)$ is the fact that it follows nearly the same recurrence relation than $O_{G_n}(\gamma)$. The only difference comes from a factor

$$[\beta_n^{12}/S_2^0(\beta_n,\gamma)]^x = (\beta^{12*}/\beta^{12})^x$$

which appears in the kernel

$$\bar O_{G_n}(\gamma) = \int d\beta_n\, j_{g_n}(\beta_n)\left(\frac{\beta^{12*}}{\beta^{12}}\right)^x \times \frac{\exp[d(\bar\gamma_n,\gamma)]}{[S_n^0(\beta_1,\ldots,\beta_n)]^{x+2}} \bar O_{G_{n-1}}(\hat S_2(\beta_n,\gamma)).$$

Then, in the same manner as for the function $O(\gamma)$ [see Eq. (48)], one can show that the sum $\bar O$ of all the open amplitudes of the Mellin transforms verifies an IE:

$$\bar O(\gamma) = \bar O_1(\gamma) + \int d\gamma^*\, \bar K(\gamma,\gamma^*) O(\gamma^*), \tag{51}$$

with

$$\bar K(\gamma,\gamma^*) = (\beta^{12*}/\beta^{12})^x K(\gamma,\gamma^*). \tag{52}$$

Due to the factor $(S_n^{12})^x$, or $(\beta_i^{12})^x$, all the expressions derived here would be well defined only if S_n^{12} or β_i^{12} would never become negative. It is not true in general [see Eqs. (7) and (11)].

So it is necessary to replace $(S_n^{12})^x$ by

$$(S_n^{12})^x \to \theta(S_n^{12})(S_n^{12})^x + \theta(-S_n^{12}) e^{i\pi x}(-S_n^{12})^x,$$

and similarly for $(\beta_i^{12})^x$. It is known that the step functions θ are the origin of the Mandelstam cut.

B. Particular value of γ: $\beta^{12} = 0$

Usually, when an IE is written for a particular values of a variable, the number of integration variables does not vary. Here, if we put $\beta^{12} = 0$, then

$$S_2^{12}(\beta',\beta) = 0,$$

the interval of integration for the variable β^{12*} disappears and the IE becomes a two-variable IE. Actually it is not possible to put $\beta^{12} = 0$ directly in the IE of Eq. (51) because the Jacobian $J(\gamma'\to\gamma^*)$ becomes infinite and one must come back to Eq. (37). The change of variables $\gamma'\to\gamma^*$, being not allowed when $\beta^{12} = 0$, we replace it by the change $\gamma'\to(\beta^{22*},\beta^{2*},u)$ with

$$u = \hat S_2^2(\beta',\gamma)/\beta^{12} = \beta^{12\prime}/(\beta^{11\prime} + \beta^{22}). \tag{53}$$

For any value of β^{12}, the IE can be written as

$$\bar O(\gamma) = \bar O_1(\gamma) + \int d\beta^{22*}\, d\beta^{2*}\, du \times \bar L(\gamma,\beta^{22*},\beta^{2*},u)\bar O(\beta^{22*},\beta^{2*},u\beta^{12}) \tag{54}$$

with

$$\bar L(\gamma,\beta^{22*},\beta^{2*},u) = \Theta_4(\beta^{22},\beta^2,\beta^{22*},\beta^{2*},u)u^x \int dv_g \exp(d) \times \prod_{j=22,2} \delta(\beta^{j*} - \hat S_2^j)\delta\left(u - \frac{\beta^{12}(\alpha)}{\hat S_2^0}\right), \tag{55}$$

where d is defined by (23b).

It is now possible to put $\beta^{12} = 0$ in the previous equation, and we find

$$\bar{O}^0(\beta^{22},\beta^2)=\bar{O}_1^0(\beta^{22},\beta^2)+\int d\beta^{22*}\,d\beta^{2*}\times\bar{K}(\beta^{22},\beta^2;\beta^{22*},\beta^{2*})\bar{O}^0(\beta^{22*},\beta^{2*}) \tag{56}$$

with $\bar{K}^0=\Sigma\bar{K}_g^0$ and

$$\bar{K}_g^0(\beta^{22},\beta^2,\beta^{22*},\beta^{2*})=\int d\nu_g(\alpha_g)\left(\frac{\beta_g^{12}(\alpha_g)}{\hat{S}_2^0(\beta(\alpha),\gamma)}\right)^x\times\exp\left[s_t\left(\beta^t(\alpha)-\frac{(\beta^2-\beta^1(\alpha))^2}{\beta^{11}(\alpha)+\beta^{22}}\right)\right]\times\prod_{j=22,2}\delta(\beta^{j*}-\hat{S}_2^j(\beta(\alpha),\beta^{22},\beta^2)) \tag{57}$$

and $\bar{O}^0(\beta^{22},\beta^2)=\bar{O}(\beta^{12}=0,\beta^{22},\beta^2)$.

When $\beta^{12}=0$, two invariants s_{11} and s_1 disappear in the expression of the IE. The solution $\bar{O}^0$ depends on the three remaining invariants s_{22}, s_2, and s_t and on the Mellin variable x. The kernel itself depends only on $s_t=t$ and x.

C. Expansion of the IE. Leading Regge poles

In Ref. 2, this reduction of the number of integration variables, when $\beta^{12}=0$, was the basis of a method of computing the amplitudes and its singularities by means of an expansion, the first term of which is precisely the function $\bar{O}_0(\beta^{22},\beta^2)$. Each term of this expansion was the solution of a Fredholm type IE, and so its singularities were given by the annulation of the determinant of the kernel. This expansion classifies the singularities, which give the Regge singularities of the amplitude, in a simple manner: Only the first term of the expansion contributes to the leading Regge pole, only the two first terms contribute to the subleading poles, and so on... . In the general case we study here, it is again possible to perform such an expansion which has the same formal structure. Of course, the nature of the kernel depends on the Lagrangian, and on the particular graphs one actually keeps in the kernel. For the complete perturbation it will be difficult to verify if we are or not in the Fredholm case.

Let us expand the function $\bar{O}(\beta^{12},\beta^{22},\beta^2)$ as a series of β^{12}:

$$\bar{O}(\beta^{12},\beta^{22},\beta^2)=\sum_n\bar{O}^n(\beta^{22},\beta^2)(\beta^{12})^n/n! \tag{58}$$

Using Eq. (51), it can be easily verify that each function $\bar{O}^n$ is the solution of an IE with a kernel $\bar{K}^n=\Sigma_g\,\bar{K}_g^n$, where $\bar{K}_g^n$ is equal to $\bar{K}_g^0$, with x replaced by $x+n$.

If we note explicitly the dependence of these kernels, they verify

$$\bar{K}^n(x)=\bar{K}^0(x+n). \tag{59}$$

If the kernel $\bar{K}_n^x$ are of the Fredholm type (bounded, squared-integrable, kernel of a compact operator,...), the relation (59) shows that all the Regge poles are given by the first kernel $\bar{K}^0(x)$. The poles coming from the other kernels $\bar{K}^n(x)$ are obtained by a simple translation: $x\to x-n$. In the φ^3 ladder case, the kernel $\bar{K}^0(x)$ are not of the Fredholm type. The expansion (58) must be slightly modified in order to obtain Fredholm type kernels, and the degenerescence of the daughter spectrum is lost (the exact degenerescence is true only in the limit $\lambda\to 0$).

As the kernels $\bar{K}^n(x)$ depend only on $s_t=t$ and x and, of course, on the coupling constant λ, the Regge poles, when they exist, depend only on t and λ. We recover here the well-known property that the Regge poles are independent of external squared four momenta p_i^2.

IV. PARTICULAR CASES

A. Particular values of the invariants

When some of the invariants are equal to zero, the structure of the integral equations changes: The number of integration variables is reduced from three to two, and even only one in one case.

1. Forward elastic scattering

The elastic scattering in the forward direction is defined by

$$p_1^2=p_3^2,\quad p_2^2=p_4^2,\quad t=0,$$

in term of the Mandelstam invariants or by

$$s_1=s_2=s_t=0$$

in term of the s_j variables [Eq. (5)].

The kernel K of Eq. (48) depends on β_2 and β_2^* only through the combination $\beta^{2*}-u\beta^2$. In particular, one of the three δ functions contains this combination. Thus, if one integrates the kernel with a function $f(\beta^{12},\beta^{22})$ which is independent of β^2, the result is also independent of β^2:

$$\int d\gamma^*\,K(\gamma,\gamma^*)f(\beta^{12*},\beta^{22*})=\int d\beta^{12*}d\beta^{22*}\,k(\beta^{12},\beta^{22};\beta^{12*},\beta^{22*})f(\beta^{12*},\beta^{22*})$$

with

$$k=\sum_g k_g$$

and

$$k_g(\beta^{12},\beta^{22};\beta^{12*},\beta^{22*})=\theta(1-|u|)\theta(\beta^{22*}-|u|\beta^{22}-2|\beta^2(\alpha_g)-u\beta^1(\alpha_g)|)\times\int d\nu_g\exp\left(-s_{11}\frac{(\beta^{12})^2}{\beta^{12}(\alpha_g)+\beta^{22}}\right)\times\prod_{j=12,22}\delta(\beta^{j*}-S_2^j(\beta(\alpha_g);\beta^{12},\beta^{22})). \tag{60}$$

So, since the first term O_1 does not depend on β^2 when s_2 is equal to zero, $O_2,O_3,\ldots,O_n$, and thus their sum O does not depend on β^2. This last function is a function of only two variables, $O=O(\beta^{12},\beta^{22})$ and it verifies an IE, the kernel of which is $k(\beta^{12},\beta^{22};\beta^{12*},\beta^{22*})$.

The reduction from three to two of the number of integration variables is a consequence of the well-known result[6] that in the equal mass case and at $t=0$ the BS IE have a supplementary symmetry.

2. Threshold in the t channel

The annulation of the three invariants s_{11}, s_{12}, and s_1 corresponds to the threshold in the t channel:

$$p_1 = -p_3 \quad \text{or} \quad p_1^2 = p_3^2 = \tfrac{1}{4}t \quad \text{and} \quad s = u.$$

When $s_{11} = s_{12} = s_1 = 0$, it can be shown, in the same manner as in the previous subsection, that the amplitude verifies an IE of only two variables β^{22} and β^2. The first term is

$$O_1 = O_1(\beta^{22}, \beta^2) = \exp(s_{22}\beta^{22} + s_2\beta^2),$$

and the kernel corresponding to the graph g becomes

$$k_g(\beta^{22}, \beta^2; \beta^{22*}, \beta^{2*}) = \int d\nu_g(\alpha_g) \exp\left[s_t\left(\beta^t(\alpha_g) - \frac{\beta^2 - \beta^1(\alpha_g)}{\beta^{11}(\alpha_g) + \beta^{22}}\right)\right] \times \prod_{j=22,2} \delta(\beta^{j*} - \hat{S}_2^j(\beta(\alpha_g); \beta^{22}, \beta^2)) \times \theta(\beta^{22*} - |u(\alpha)|\beta^{22} - 2|\beta^{2*} - u(\alpha)\beta^2|) \tag{61}$$

with

$$u(\alpha) = \beta^{12}(\alpha)/[\beta^{11}(\alpha) + \beta^{22}].$$

3. Elastic scattering with some external momentum equal to zero

Here we consider the case where

$$p_1 = p_3 = 0$$

or

$$p_1^2 = p_3^2 = t = 0$$

and

$$s = u = p_2^2 = p_4^2.$$

This case contains, as an even more particular case, the scattering when all the momentum and all the invariants are equal to zero:

$$p_i = 0, \quad i = 1,2,3,4.$$

The simplifications of the two previous paragraphs can be done together, and one obtains an IE of only one variable β^{22}, with a kernel which is given by $k = \Sigma_g k_g$ and

$$k_g(\beta^{22}, \beta^{2*}) = \int d\nu_g \, \delta(\beta^{22*} - \hat{S}_2^{22}(\beta(\alpha_g); \beta^{22})) \times \theta(\beta^{22*} - |u(\alpha)|\beta^{22} - 2|\beta^2(\alpha) - u(\alpha)\beta^1(\alpha)|). \tag{62}$$

4. Bound states

It is possible to give another interesting interpretation of the cases studied in subsections 2 and 3. Using, for example, the Schwinger representation of the graphs, one can see that the vertex function of a bound state (the "relativistic wavefunction"), of squared mass t and which contains two particles of momentum p_2 and p_4, has exactly the same structure in the invariant $s_t = t$, $s_2 = \frac{1}{2}(p_4^2 - p_2^2)$, and $s_{22} = \frac{1}{2}(p_2^2 + p_4^2) - \frac{1}{4}t$ as the $2 \to 2$ amplitudes in the t channel when $p_1 + p_3 = 0$. Thus one can define on "open" relativistic wavefunction $\varphi(\beta^{22}, \beta^2)$ which verifies a homogeneous IE:

$$\varphi(\beta^{22}, \beta^2) = \int d\beta^{22*} d\beta^{2*} \, k(\beta^{22}, \beta^2; \beta^{22*}, \beta^{2*}) \varphi(\beta^{22*}, \beta^{2*}). \tag{63}$$

The wavefunction itself can be computed by integrating the open wavefunction φ over β^{22} and β^2.

Similarly, if now one considers a bound state of mass equal to zero ($t = 0$), its "open" relativistic wavefunction $\varphi(\beta)$ verifies a homogeneous IE of one variable:

$$\varphi(\beta^{22}) = \int d\beta^{22*} \, k(\beta^{22}; \beta^{22*}) \varphi(\beta^{22*}). \tag{64}$$

B. Particular value of γ: $\beta^{12} = 0$

It has been seen in the previous section that the IE verified by the Mellin transform became simpler when β^{12} was equal to zero. It is also the case for the IE (48). The amplitudes $O^0(\beta^{22}, \beta^2) = O(\beta^{12} = 0, \beta^{22}, \beta^2)$ verifies an IE with a kernel $K^0 = \Sigma_g K_g^0$ defined by

$$K_g^0(\beta^{22}, \beta^2; \beta^{22*}, \beta^{2*}) = \int d\nu_g(\alpha) \exp\left[s_t\left(\beta^t(\alpha) - \frac{(\beta^2 - \beta^1(\alpha))^2}{\beta^{11}(\alpha) + \beta^{22}}\right)\right] \times \prod_{j=22,2} \delta(\beta^{2*} - \hat{S}_2^j(\gamma(\alpha), \beta^{22}, \beta^2)). \tag{65}$$

When $\beta^{12} = 0$, three invariants, s_{12}, s_{11}, and s_1, disappear in the expression of the IE. The solution depends only on the three remaining invariants s_{22}, s_2, and s_t (the kernel depends on s_t and the first term on s_{22} and s_2).

C. Particular class of graphs

In this section, we consider a particular class of graphs: the generalized rung ladder graph (GRLG), where the rungs are made with subgraphs which are linked to each upright by only one vertex (see Fig. 4, where different examples of such generalized rungs are given). To this class belongs the ladder graph of φ^3, which has been already widely studied in our previous paper.[2] When one considers the GRLG, four among the seven topological polynomials [Eq. (3)] of the rungs are equal to zero,

$$A^t = A^u = A^1 = A^3 = 0, \tag{66}$$

and the formalism which has been worked up in the first sections become simpler: All the kernels [Eqs. (25), (31), (35), and (46)] depend on the graph by the same function $\tilde{j}_g$ of only one variable, and, except for this function, they are explicit functions. In the φ^3 ladder graph case, the function $\tilde{j}_g$ is a constant, and, of course, we find the IE of our previous paper again.

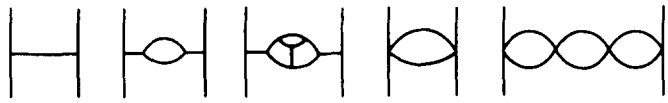

FIG. 4. Some example of generalized rungs. The three first graphs come from a φ^3 Lagrangian; the two last ones, from a φ^4 Lagrangian.

In this subsection we note g the generalized rung itself and $\bar{g}$ the same graph with two vertical lines added (see Fig. 5). If α and α' are the Schwinger parameters attached to these two lines, the measure $d\mu_{\bar{g}}$ becomes

$$d\mu_{\bar{g}}(\alpha_{\bar{g}}) = d\alpha\, d\alpha' \exp[-(\alpha+\alpha')m^2]\, d\mu_g(\alpha_g).$$

In addition to the relations (66), the particular structure of the graphs lead to simple expressions for the other topological polynomials of $\bar{g}$:

$$P_{\bar{g}} = P_g, \quad A^{\,2}_{\bar{g}} = \alpha A^{\,2}_g,$$
$$A^{\,4}_{\bar{g}} = \alpha' A^{\,4}_g, \quad A^{\,s}_{\bar{g}} = A^{\,s}_g.$$

If one computes the $\beta^{\,j}_{\bar{g}}$ functions, one obtains

$$\begin{aligned}
&\beta^{\,1}_{\bar{g}}(\alpha_{\bar{g}}) = 0,\\
&\beta^{\,11}_{\bar{g}}(\alpha_{\bar{g}}) = \beta^{\,12}_{\bar{g}}(\alpha_{\bar{g}}) = \beta^{\,12}_{g}(\alpha_g),\\
&\beta^{\,22}_{\bar{g}}(\alpha_{\bar{g}}) = \alpha+\alpha'+\beta^{\,12}_{g}(\alpha_g),\\
&\beta^{\,2}_{\bar{g}}(\alpha_{\bar{g}}) = \tfrac{1}{2}(\alpha'-\alpha), \quad \beta^{\,t}_{\bar{g}}(\alpha_{\bar{g}}) = \tfrac{1}{4}(\alpha+\alpha').
\end{aligned} \tag{67}$$

We are not going to transform all the results of the previous sections, but only the main ones.

Taking into account the relations (67), we can give the new expression of the kernel $j_{\bar{g}}(\beta)$ [see Eq. (25)]

$$j_{\bar{g}}(\beta) = \delta(\beta^{\,1})\delta(\beta^{\,11}-\beta^{\,12})\delta(\beta^{\,t}-\tfrac{1}{4}(\beta^{\,22}-\beta^{\,12})) \times \exp(-\beta^{\,22}m^2)\tilde{j}_g(\beta^{\,12}),$$

where the new function $\tilde{j}_g$ is defined by

$$\tilde{j}_g(\beta^{\,12}) = \theta(\beta^{\,12})\exp(\beta^{\,12}m^2)\int d\mu_g\,\delta(\beta^{\,22}-\beta^{\,12}(\alpha_g)). \tag{68}$$

In the particular case of the φ^3 ladder graphs, the function $\tilde{j}_g$ is a constant:

$$\tilde{j}_g(\beta^{\,12}) = \lambda^{\,2}.$$

The β-representation [Eq. (24)] can be written

$$I_G = \int \prod_{i=1}^{n} [d\beta^{\,12}_i \tilde{j}_{g_i}(\beta^{\,12}_i)]\, Q_n(\beta^{\,12}_{(n)}), \tag{69}$$

where Q_n is an explicit function of n variables, $\beta^{\,12}_{(n)} = (\beta^{\,12}_1, \beta^{\,12}_2, \ldots, \beta^{\,12}_n)$, independent of the graph and equal to

$$Q_n(\beta^{\,12}_{(n)}) = \int \prod_{i=1}^{n} [d\beta^{\,22}_i\, d\beta^{\,2}_i \exp(-\beta^{\,22}_i m^2)] \times \frac{\exp[D_n(\beta_1,\ldots,\beta_n)]}{[S^0_n(\beta_1,\ldots,\beta_n)]^2}\Bigg|_{\substack{\beta^{\,1}_i=0;\ \beta^{\,11}_i=\beta^{\,12}_i;\\ \beta^{\,t}_i=(\beta^{\,22}_i-\beta^{\,12}_i)/4;\\ i=1,\ldots,n.}}$$

For example, the β-representation of dimension 1 is

$$I_G = \int d\beta^{\,12}\, \tilde{j}_G(\beta^{\,12}) Q_1(\beta^{\,12}),$$

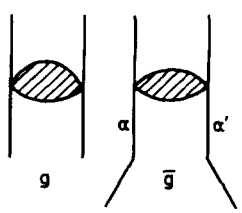

FIG. 5. Definition of g and $\bar{g}$. In subsection IV C, g is the rung itself and $\bar{g}$ is the rung to which two vertical lines have been added. α and α' are the Schwinger parameters attached to these two additional lines.

with

$$Q_1(\beta^{\,12}) = \frac{\exp[-\beta^{\,12}(m^2-s_{11}-s_{12}-s_{22})]}{(m^2-s_{22}-\frac{1}{4}s_t)^2-(\frac{1}{2}s_2)^2} = \frac{\exp[-(m^2-s)\beta^{\,12}]}{(m^2-p_2^2)(m^2-p_4^2)}.$$

The denominator of Q_1 represents the propagators of the two extra lines which have been added to the graph.

Let us now come to the integral equation itself [Eq. (48)]. The product of the three δ functions in Eq. (47) can be written as

$$[\beta^{\,22}/\beta^{\,12}(1-u)^2]\delta(\beta^{\,12}_g(\alpha_g) - [u/(1-u)]\beta^{\,22}) \times \delta(\alpha+\alpha'-\beta^{\,22*}+u\beta^{\,22})\delta(\tfrac{1}{2}(\alpha'-\alpha)-\beta^{\,2*}+u\beta^{\,2}). \tag{70}$$

In the definition of K_g [Eq. (47)], the integrations over the two variables α and α' can be done, using the two last δ functions. The remaining integrations of the first δ function give the $\tilde{j}_g$ function with an argument equal to $[u/(1-u)]\beta^{\,22}$. As $\beta^{\,12}$ and thus u are always positive variables, the Θ_4 function becomes simpler:

$$\Theta_4(\gamma,\gamma^*) = \theta(U-u)$$

with

$$U = \inf\left(1, \frac{\beta^{\,22*}-\beta^{\,2*}}{\beta^{\,22}-\beta^{\,2}}, \frac{\beta^{\,22*}+\beta^{\,2*}}{\beta^{\,22}+\beta^{\,2}}\right).$$

Finally we obtain

$$K_g(\gamma,\gamma^*) = \frac{1}{\beta^{\,12}\beta^{\,22}}\exp\left(-\beta^{\,22*}m^2 - \frac{u^2}{1-u}\beta^{\,22}m^2\right) \times \exp(d)\,\tilde{j}_g\left(\frac{u}{1-u}\beta^{\,22}\right) \times \theta(U(\beta^{\,22},\beta^{\,2},\beta^{\,22*},\beta^{\,2*})-u)$$

with

$$d = -[s_{11}(\beta^{\,12})^2 + s_1\beta^{\,12}\beta^{\,2} + s_t(\beta^{\,2})^2](1-u)/\beta^{\,22} + s_t(\beta^{\,22*}-u\beta^{\,22})/4.$$

In order to verify that, in the φ^3 ladder case ($\tilde{j}_g = \lambda^{\,2}$), this kernel is actually identical to the one of Ref. 2, two changes must be done. First we must perform the change of variables $\beta^{\,12}, \beta^{\,2}, \beta^{\,22} \to \alpha, \alpha', \beta$, defined by the relations

$$\begin{aligned}
&\alpha = \tfrac{1}{2}(\beta^{\,22}-\beta^{\,12}) - \beta^{\,2},\\
&\alpha' = \tfrac{1}{2}(\beta^{\,22}-\beta^{\,12}) + \beta^{\,2},\\
&\beta = \beta^{\,12}.
\end{aligned}$$

The other change comes from the different normalization of the amplitudes. Here the pole term is $O_1(\gamma)$ [see Eq. (38)] when in Ref. 2 it would be defined by

$$F_1(\alpha,\alpha',\beta) = \exp(s\beta + p_2^2\alpha + p_4^2\alpha').$$

V. RENORMALIZATION: φ^3 INTERACTION LAGRANGIAN CASE

As soon as some graphs of the theory are divergent, we have to take into account the renormalization operator R. We do this here only for the most simple case, namely the φ^3 Lagrangian case.

The general definition of the renormalization operator can be found in Refs. 1(d) or 5.

When we restrict ourself to φ^3 interaction, there is only one connected divergent subgraph in the theory (see Fig. 6).[1(d)] Such subgraphs are logarithmically divergent, and they are always disjoint. The renormalization operator for a given graph G reduces to

$$R^G = \prod_{g^d} (1 - \tau_{g^d}^{-4}) \tag{71}$$

where g^d are all the connected divergent subgraphs described by Fig. 6 of G and $\tau_{g^d}^{-4}$ is the generalized subtraction Taylor operator. If the graph G is partitioned into a set of subgraphs $g_i : G = (g_1, g_2, \ldots, g_n)$, for example, if one considers the Bethe–Salpeter structure of G [see Fig. 1(b)], then R^G appears as a product of renormalization operators, each acting on a given subgraph g_i:

$$R^G = \prod_{i=1}^{n} R^{g_i}. \tag{72}$$

It is this property which makes easy the demonstration of the compatibility of the renormalization and of the β-representation of the Feynman integral. Before going on, we give the expression of R^{g^d} for a simple divergent graph. The operator R^{g^d} is an operator which acts on a function $f(\alpha,\alpha')$ which depends on the two Schwinger parameters of the graph g^d (see Fig. 6). Putting $\lambda = 4$ in Eqs. (I.9) and (I.10) of Ref. 1(d) and using the integral representation of the Taylor remainder [see, for example, Eq. (III.15) of Ref. 2], R^{g^d} can be written as

$$\begin{aligned} R^{g^d} f(\alpha,\alpha') &= \int_0^1 \frac{dg^*(u)}{du}\, du \\ &= f(\alpha,\alpha') - \lim_{u\to 0} [u^4 f(\alpha u^2, \alpha' u^2)], \end{aligned} \tag{73}$$

where $g(u)$ is defined by

$$g(u) = u^4 f(\alpha u^2, \alpha' u^2).$$

The generalization to the case of several divergent subgraphs is straightforward, but we are not going to write it because the only property we need is actually the factorization property of Eq. (72).

In the Euclidian space, the amplitude I_G of a graph G is [see Eq. (1)]

$$I_G = \int d\lambda_G(\alpha_G)\, R^G \left(\frac{e^{D_G(\alpha_G)}}{[P_G(\alpha_G)]^2} \right),$$

with

$$d\lambda_G(\alpha_G) = P_G^2(\alpha_G)\, d\mu(\alpha_G).$$

The functions D_G and P_G verify the structure property of Eqs. (12) and (21). Then, using (26), one obtains

$$\frac{e^{D_G(\alpha_G)}}{[P_G(\alpha_G)]^2} = \frac{1}{[\Pi_{i=1}^n P_{g_i}(\alpha_{g_i})]^2} \int \prod_{i=1}^{n} \prod_{j\in K} d\beta_i^j \times \delta(\beta_i^j - \beta_{g_i}^j(\alpha_{g_i})) \frac{e^{D_n(\beta_{(n)})}}{[P_n(\beta_{(n)})]^2}.$$

If one commutes the integration on the variable β_i^j and the renormalization operator R^G, then one finds that the β-representation (24) remains unchanged, except that $j_g(\beta)$ needs to be renormalized and becomes

$$j_g^R(\beta) = \int d\lambda_g(\alpha_g)\, R^g \left(\frac{\Pi_{j\in K}\, \delta(\beta^j - \beta_g^j(\alpha_g))}{[P_g(\alpha_g)]^2} \right). \tag{74}$$

One sees now a supplementary advantage of the β-representation: the different singularities of the Schwinger representation are disconnected:

—The UV divergences appear only in the α_g integration which are contained in the expression of j_g.

—The Landau singularities can come only from the β integration because only the function $D_n(\beta_{(n)})$ depends on the Lorentz invariants s_i.

VI. CONCLUSION

In the present paper, it has been shown that the Schwinger parameter formalism, could be modified in such a way that the Bethe-Salpeter structure of the amplitude becomes explicit. This is done through the introduction of a new scalar representation of the Feynman amplitudes, the β-representation [Eq. (24)]. The fundamental feature of this β-representation is the quasifactorization property of Theorem 3. Reflecting the generalized ladder structure of the graphs, the β-representation naturally exhibits a recurrence law in the number of "rungs" [Eq. (32)]. We are then able to build the infinite sum of the "open amplitudes" as the solution of a three-variable integral equation [Eq. (48)]. The last step to obtain the four-point amplitude is to perform the closing integration [Eq. (39)].

We conclude and indicate the next steps that this program should follow. The treatment of the renormalization is, of course, one of them. It has been shown in the framework of asymptotic behavior studies[1(e)] that in the case of a strictly renormalizable theory (such as φ^3 in dimension six or φ^4 in dimension four) the renormalization procedure can be split into two steps: On the one hand, the divergent subgraphs occuring inside the t-2PI subgraphs have to be subtracted: a behavior predicted by the renormalization group is thus generated for the infinite sum of graphs building each "rung" of the generalized ladder; on the other hand, UV divergences arising from the ladder structure itself have to be treated. Obviously we have to look for such a two-step treatment within our framework. Already, for the φ^3 Lagrangian, we have shown (Sec. V) that the R operator respects the factorization property of Theorem 3 (see Sec. I D).

The next point of our program after renormalization has to do with the fact that the actual properties of the solution of our integral equation, of course, depends on the analytic structure of the kernel [the inhomogeneous term is explicit; see Eq. (38)]. This structure is not known in general for the complete perturbative expansion of the kernel.

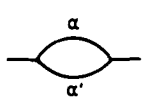

FIG. 6. The only connect divergent subgraph of φ^3.

However, our approach allows to reach many exact results even in cases where infinite subseries of the perturbation series are kept: The structure of the kernel is actually entirely explicit whenever it is restricted to a finite sum. It is then possible to classify the cases where global theorems (such as Fredholm theorems) may be used: quantitative work, such as in the φ^3 ladder case,[2] can be done.

In this paper we have paid attention essentially to the four-point amplitude. As outlined in Sec. IV, it is possible to exhibit an analogous integral equation for the three-point amplitude (vertex). This can also be obtained for the propagator.

Let us end this conclusion by a remark concerning the contested interest of the study of the φ^3 ladder subseries presented in Ref. 2. The results we have obtained in the present paper, taking into account the whole perturbation series, indeed show that essential properties of the perturbation series are already present in the ladder. For example, as in the ladder case, we find a three-variable integral equation and this equation happens to be simpler under the same circumstances (reduction of the number of variables in various particular cases). Also, the β^{12} expansion, the analog of the γ expansion for the ladder case, allows us to classify the singularities in the Mellin space. We even obtain a complete analogy between the ladder and the "generalized rung ladder" (see Sec. IV C and Fig. 4).

As a last statement, we want to stress the importance of the kind of factorization property of a Feynman amplitude into a "skeleton," which exhibits its BS structure and contains its external momentum dependence, and a "dressing," which carries the whole information concerning the dynamics attached to the interaction Lagrangian.

APPENDIX: VARIATION DOMAINS

In the integral (37), the integration domain of the variable β' is determined by the factor $\Theta_1(\beta')$, which is present in the kernel (see Eqs. (35) and (25)]. If one performs the change of variables $\gamma' \to \gamma^*$, the new integration domains of the integration variables $(\gamma^*, \bar{\gamma}')$ must be determined. As the change of variables depends on γ [see Eq. (42)], the new domain also depends on γ. We are going to describe this domain in two steps: the variation domain of $\bar{\gamma}'$ when γ and γ^* are fixed; the variation domain of γ^* and γ is fixed.

A. Variation domain of $\bar{\gamma}'$ with γ and γ^* fixed

Using Eq. (42), one calculates γ' as a function of γ, γ^* and $\bar{\gamma}'$:

$$\begin{aligned} \beta^{12'} &= u(\beta^{11'} + \beta^{22}), \\ \beta^{22'} &= \beta^{22*} + u^2(\beta^{11'} + \beta^{22}), \\ \beta^{2'} &= \beta^{2*} - u(\beta^2 - \beta^{1'}) \end{aligned} \tag{A1}$$

with

$$u = \beta^{12*}/\beta^{12}.$$

Then one writes that $\beta' = (\gamma', \bar{\gamma}')$ verifies the three conditions (8):

$$(8a) \Rightarrow -2|\beta^{1'}| + (1 - |u|)\beta^{11'} - |u|\beta^{22} \leqslant 0, \tag{A2a}$$

$$\begin{aligned} (8b) \Rightarrow & 2|\beta^{1'} - \beta^2 + \beta^{2*}/u| + (1 - |u|)\beta^{11'} \\ & - \beta^{22*}/|u| + \beta^{22}(1 - |u|) \leqslant 0 \end{aligned} \tag{A2b}$$

$$(8c) \Rightarrow |\beta^{1'}| + |\beta^{2*} - u(\beta^2 - \beta^{1'})| - 2\beta'' \leqslant 0. \tag{A2c}$$

These three inequalities define the variation domain of $\bar{\gamma}' = (\beta^{11'}, \beta^{2'}, \beta'')$.

B. Variation domain of γ^* with γ fixed

This domain is defined by the condition that the previous domain for $\bar{\gamma}'$ is not empty. A necessary condition for the inequality (A2a) to be verified is

$$|u| < 1. \tag{A3}$$

It can be easily shown that the compatibility of the relations (A2a) and (A2b) needs the fact that γ^* verifies the inequality

$$2|\beta^{2*} - u\beta^2| \leqslant \beta^{22*} - |u|\beta^{22}. \tag{A4}$$

The two relations (A3) and (A4) determine the variation domain of γ^*:

$$\Theta_4(\gamma, \gamma^*) = \theta(1 - |u|)\theta(\beta^{22*} - |u|\beta^{22} - 2|\beta^{2*} - u\beta^2|). \tag{A5}$$

[1](a) O. I. Zav'Yalov, Zh. Eksp. Teor. Fiz. **47**, 1099 (1964) [Sov. Phys. JETP **20**, 736 (1965)]; (b) O. I. Zav'Yalov and B. M. Stepanov, Yad. Fiz. **1**, 922 (1965) [Sov. J. Nucl. Phys. **1**, 658 (1965)]; (c) M. C. Bergère and Y. M. P. Lam, Comm. Math. Phys. **39**, 1 (1974), and Freie Universitat Berlin Preprint HEP May 74/9, 1974, unpublished; (d) M. C. Bergère and C. Gilain, J. Math. Phys. **19**, 1495 (1978); (e) M. C. Bergère and C. de Calan, Phys. Rev. D **20**, 2047 (1979).
[2]C. Gilain and D. Levy, J. Math. Phys. **22**, 1787 (1981).
[3]N. Nakanishi, Suppl. Prog. Theor. Phys. **43**, 1 (1969).
[4]C. Gilain, thesis, Université de Paris-Sud-Centre d'Orsay, 1981.
[5]M. C. Bergère and J. B. Zuber, Comm. Math. Phys. **35**, 113 (1974).
[6]G. Domokos and P. Suranyi, Nucl. Phys. **54**, 529 (1964).

Hamiltonian operators with maximal eigenvalues

Evans M. Harrell II[a)]
School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332

(Received 2 August 1983; accepted for publication 16 September 1983)

The potentials $V(x)$ with a given L^p norm that maximize the lowest eigenvalue of $-\Delta + V$ are characterized.

PACS numbers: 02.30.Tb, 03.65. – w

I. INTRODUCTION

How large can a given eigenvalue of a differential operator be? This question has implications for many topics in mathematical physics, especially when the operator has the form

$$H = -\Delta + V(x),$$

where V is a real-valued multiplication operator. Self-adjoint realizations of H are the fundamental mathematical objects of quantum mechanics. The eigenvalues are the energy levels of quantum-mechanical particles, and V is the potential energy. Here the variable will range over a finite open domain D in $\mathbb{R}^m$ with a smooth boundary (an assumption much stronger than necessary), and V will be a nonnegative function in $L^1(D)$. Nonnegativity of V is assumed only to avoid confronting questions of self-adjointness. Except for that it would follow automatically that a potential maximizing an eigenvalue would be nonnegative. H can be defined as a self-adjoint operator on L^2 by either of the following methods.

(a) Let $-\Delta$ be the usual self-adjoint Laplacian with Dirichlet boundary conditions on the boundary of D, and define $-\Delta + V(x)$ via the sum of the associated quadratic forms.[1] Alternatively, equip $-\Delta$ with Neumann or mixed boundary conditions.

(b) Extend H to the infinite domain $\mathbb{R}^m$ by forcing the potential outside D to equal an appropriate fixed function. With the assumptions to be imposed on V, it suffices to have the exterior potential be bounded below, locally integrable, and greater than a positive constant outside some compact set (the constant need only be large enough to ensure the existence of an eigenvalue).

Let $Q(p, c)$ denote the set of potentials V defined on D such that $\|V\|_p \leqslant c$. Let $E(V)$ denote the ground-state (lowest) eigenvalue of H. The question asked above can now be made specific: What is the supremum of $E(V)$ over the set $Q(p, c)$ and for what V is it attained, if any? The answer turns out to be that there is a maximizing potential, and that it is of a very special form, ordinarily the maximal eigenvalue times a characteristic function,

$$V_*(x) = E_{\max}\, \chi_S(x).$$

Indeed, the techniques of this paper also allow one to characterize the function $V(x)$ that maximizes the bottom of the spectrum of a rather general semibounded operator of the form $T + V$, where T represents a closed, semibounded operator on $L^2(D)$ with a few simple properties. Specifically, the domain of self-adjointness of $T + V$ should be the same for all V in $Q(p, c)$ and T should be local in the sense that if f is constant (a.e.) on an open subset U of D, then $Tf = 0$ a.e. on U. For example, T could be a positive higher-order differential operator with no zeroth order term. The maximizing potential function V is still ordinarily of the form $E_{\max}\, \chi_S(x)$, subject to qualifications analogous to the ones spelled out below for the case $T = -\Delta$.

This problem was raised most recently in a list of open problems in mathematical physics at a meeting of the American Mathematical Society.[2] Prominent among the reasons for interest in it are its implications for inverse spectral theory, where for practical as well as theoretical reasons it is important to know what properties of a potential are determined by incomplete spectral information. The result mentioned above would be read by an inverse-spectral theorist the other way around, as stating that if the lowest eigenvalue is larger than a certain amount, then the L^p norms of V are larger than something, and that if a potential has L^p norm equal to c and maximizes the eigenvalue, then it has a particular very simple form. From the latter point of view the statement is reminiscent of Levitan and Gasymov's striking version of Ambarzumian's theorem, viz., for $V \in L^1[0,1]$ and Neumann boundary conditions imposed at 0 and 1, if $E_0 = 0$,

$$E_n - n^2 \to 0,$$

where E_n is the nth eigenvalue, then necessarily $V(x) = 0$ a.e.[3]

II. MAXIMIZING POTENTIALS

Let H be as above, and suppose that V belongs to $Q(p, c)$ for some fixed p, c, and D. In the case $p = \infty$ it is obvious that the lowest or any other eigenvalue is maximized by $V = c$, so $p = \infty$ will not be considered further. It will first be established that there exists a V in $Q(p, c)$ that maximizes the lowest eigenvalue, at least for certain p.

Proposition 1: There is a bound on the lowest eigenvalue depending only on p, c, and D. Consequently there exists a maximizing sequence $V_n \in Q(p, c)$ such that

$$\lim_{n \to \infty} E(V_n) = E_{\max} \equiv \sup_{Q(p,c)} E(V).$$

Proof: The normalized ground-state eigenfunction f_0 of $-\Delta$ is bounded and hence in the quadratic-form domain of H. Therefore an upper bound for $E(V)$ is given by the Rayleigh–Ritz inequality as

a) Partially supported by NSF grant MCS 7926408.

 0022-2488/84/010048-04$02.50

$$E(V) \leqslant E(0) + (f_0, Vf_0)$$
$$\leqslant E(0) + \|f_0\|_\infty^2 \|V\|_1$$
$$\leqslant E(0) + \|f_0\|_\infty^2 \|V\|_p [Vol(D)]^{1/q}, \quad 1/p + 1/q = 1,$$

which depends only on c, p, and D. [$E(0)$ is just the lowest eigenvalue of $-\Delta$.] ∎

Remark: Any sufficiently smooth function f_0 in the quadratic-form domain of H will furnish an upper bound. The normalized ground-state eigenfunction gives a good estimate to compare with the exact answer for simple special cases.

Proposition 2: For all $N > 0$ there exists a $V \in Q(p, c) \cap Q(\infty, N)$ that maximizes $E(V)$ within that class. If $p > \max(2, m/2)$, then there exists a maximizing potential V within $Q(p, c)$.

Proof: By interpolation $Q(p, c) \cap Q(\infty, N)$ lies within $Q(r, c')$ for all $r \geqslant p$ and some c' depending on r. Choose $r > 2$ and $> m/2$; this ensures that the eigenvalue depends continuously on V in the $\| \|_r$ norm.[1,4] The maximizing sequence V_k within $Q(p, c) \cap Q(\infty, N)$ has a subsequence that converges weakly in L^r to some limit V_*. By a theorem of Mazur[5] there is a sequence of convex combinations of V_k that converges strongly to V_*. Since $Q(p, c) \cap Q(\infty, N)$ is convex, the new sequence remains within that class. By the Rayleigh–Ritz inequality, the replacement of V_k by convex combinations can only increase $E(V)$, i.e., if

$$\sum_i a_i = 1, \quad a_i \geqslant 0,$$

and f now denotes the normalized ground-state eigenfunction of

$$-\Delta + \sum_i a_i V_i,$$

then

$$E\left(\sum_i a_i V_i\right) = \left(f, \left(-\Delta + \sum_i a_i V_i\right) f\right)$$
$$= \sum_i a_i (f, (-\Delta + V_i) f)$$
$$\geqslant \sum_i a_i E(V_i).$$

It follows that $E(V_*) = E_{\max}$. Observe that the relevance of Mazur's theorem is more convex combination than the nature of the convergence. The latter takes place in a somewhat arbitrary L^r. Of course, if $p > \max(2, m/2)$, then the truncation to $Q(\infty, N)$ in this proof is unnecessary. ∎

Definition: The potential function V is a local eigenvalue extremizer for the set $Q(p, c)$ iff

(a) $\|V\|_p = c$;

(b) H (or its restriction to a given connected subset of D) has a nondegenerate eigenvalue Λ;

(c) for every bounded multiplicative function $W(x)$ such that

$$\left.\frac{d\|V_t\|_p}{dt}\right|_{t=0} = 0,$$

where $V_t = V + tW$, the eigenvalue $\Lambda(V_t)$ such that $\Lambda(V_0) = \Lambda$ satisfies

$$\left.\frac{d\Lambda(V_t)}{dt}\right|_{t=0} = 0.$$

Remarks: (a) Perturbation theory guarantees the existence and differentiability of $\Lambda(V_t)$ for sufficiently small real values of t.[4]

(b) This is a necessary condition for V to maximize the lowest eigenvalue, which is known to be nondegenerate (after restriction to a connected component of D, if necessary); if it were false, then W could be given some higher-order dependence on t so that $V + tW \in Q(p, c)$, but dE/dt would still differ from 0.

Proposition 3: Any local eigenvalue maximizer in $Q(p, c)$ is equivalent almost everywhere to a function satisfying the nonlinear partial differential equation

$$\Delta V^{(p-1)/2} = (V - \Lambda) V^{(p-1)/2} \tag{1}$$

on the interior of its support.

Remark: This curious equation has the obvious solution $V = \Lambda$ on $S = \text{int supp}(V)$, which is the only solution when $p = 1$. It would be surprising if other conceivable solutions were relevant, but they might arise if either the shape of D or the boundary conditions were peculiar enough. While (1) is trivially satisfied away from S, it is *not* satisfied on the boundary of S, and so does not hold throughout D in the usual distributional sense.

Proof: Let y and z be points in S at the centers of small balls of radius d, denoted Y and Z. Let

$$W(x) = \chi_Y(x) - k\chi_Z(x),$$

where k is chosen to satisfy the condition in (c) of the definition. Since for almost every y and z the averages of V^p over Y and Z approach $V^p(y)$ and $V^p(z)$ as $d \to 0$,[6] from the definition of the L^p norm, k can be taken arbitrarily close to the value

$$(V(y)/V(z))^{p-1}$$

for almost every y and z (write the integrand for $\|V_t\|_p^p$ to first order in t). Let $\psi(x)$ be the normalized eigenfunction for $\Lambda(V)$. By the Feynman–Hellmann theorem,[4]

$$\left.\frac{d\Lambda(V_t)}{dt}\right|_{t=0} = \int \chi_X \psi^2(x)dx - \int \chi_Y k\psi^2(x)dx.$$

For V to be a local eigenvalue extremizer it is necessary for the derivative to be 0 regardless of y, z, and d. By letting d tend to 0, it follows that for almost every y and z in S,

$$\psi^2(y) = (V(y)/V(z))^{p-1}\psi^2(z),$$

or, in other words, that

$$\psi(x) = CV^{(p-1)/2}(x) \text{ almost everywhere on } S \tag{2}$$

for some constant C. Since $\Delta\psi = (V - \Lambda)\psi$ (sense of distributions), Eq. (2) implies Eq. (1). ∎

Actually, Eq. (2) holds almost everywhere on supp(V) (the distinction is the possible existence of nowhere dense sets of positive measure), since the balls can be replaced with appropriate sets that "shrink nicely."[6]

Proposition 4: Let $V \geqslant 0$, $V \in L^p(D)$, $p \geqslant 1$, D as above and moreover assumed connected. Define $V_T = \min(V, T)$.

Let $E_0(H)$ and $\phi(H)$ denote the ground-state eigenvalue and eigenfunction of an operator H. Then $E_0(-\Delta + V_T)$ tends monotonically to $E_0(-\Delta + V)$ and $\phi(-\Delta + V_T)$ tends to $\phi(-\Delta + V)$ in L^2.

Remark: Connectedness just ensures nondegeneracy of the ground state.

Proof: For simplicity of notation, let $f = \phi(-\Delta + V)$ and $E = E_0(-\Delta + V)$. Monotonicity of the eigenvalue is an immediate and well-known consequence of the min-max principle, or the Rayleigh–Ritz inequality. From straightforward corollaries of the spectral theorem it suffices to show that

$$\|(-\Delta + V_T - E)f\|_2 \to 0.$$

Actually, this just ensures that some point of the spectrum of $-\Delta + V_T$ tends to E and the associated eigenfunction converges. But since the ground-state eigenfunctions are characterized by positivity, that point has to be the ground state. Also, set $p = 1$, which includes all the other cases.

Since $V_T(x)f(x)$ increases monotonically to $V(x)f(x)$, the distribution $(-\Delta + V_T)f$, which is only in L^1 *a priori*,[1] increases to $(-\Delta + V)f = Ef \in L^2$. Therefore

$$\|(-\Delta + V_T - E)f\|_2^2 = \int_D ((-\Delta + V_T - E)f)^2 d^m x$$

is finite, and hence tends to zero by the monotone convergence theorem. ∎

Theorem 1: For $p = 1$ or $p > 2$, $m/2$, there is a potential in $Q(p,c)$ that maximizes the lowest eigenvalue, and it satisfies (1) with $\Lambda = E_{\max}$ on S. In particular, when $p = 1$, $V_* = E_{\max}$ and ψ equals its maximum almost everywhere on S.

Proof: The foregoing propositions cover all p other than $p = 1$. If $p = 1$, then consider the set $Q(1, c) \cap Q(\infty, N)$ in place of $Q(1, c)$, where N is larger than the upper bound on $E(V)$ from Proposition 1. The proof of Proposition 3 goes through unchanged, so that on $\text{supp}(V)$, $\psi(x) = C$ (a fixed constant) and $V = E(V)$ almost everywhere, independently of N as $N \to \infty$. But truncation of V at high values affects the ground-state eigenvalue continuously by Proposition 4. Hence there cannot be an unbounded $V \in Q(1, c)$ with a higher eigenvalue than the maximum on $Q(1, p) \cap Q(\infty, N)$. ∎

Theorem 2: If $p = 1$, or if $p \neq 1$, but it is known that V_* exists and is constant on its support, then V_* is unique a.e.

Proof: Suppose that there were two distinct sets S. Then, as in the proof of Proposition 2, the eigenvalue corresponding to the average of the two maximizing potentials would be no less than $E_{\max}$, since the average is a convex combination. This is a contradiction, since the averaged potential would equal $E_{\max}/2$ on a set of positive measure. ∎

What makes the proof of Theorem 1 work is that all the maximizing potentials within $Q(1, c) \cap Q(\infty, N)$ satisfy a pointwise bound independent of N. If the same were known for all p, then the restriction to values for which V is relatively bounded could be dispensed with. It would suffice, for example, to know that the only solution of (1) of interest is the obvious one. In principle, these arguments leave open the possibility that different solutions are relevant for different N, and do not have a uniform bound.

III. EXAMPLES

The one-dimensional case of an interval is rather easy to analyze in detail, since there are no geometrical complications and since all eigenvalues are automatically nondegenerate. By a change of variable it suffices to consider only the interval $[0,1]$. The case of a sphere is similar.

Scholium: Let H be the one-dimensional operator $-d^2/dx^2 + V(x)$ on $L^2[0,1]$, with Dirichlet boundary conditions, and denote the nth eigenvalue E_n, $n = 0,1,2,\ldots$. Let V range over $Q(1, c)$. The eigenvalue E_n is maximized by potentials of the form

$$V_n(x) = E_{n,\max}\, \chi_{S_n}(x),$$

uniquely determined only for $n = 0$. If $n > 0$, then there are uncountably many distinct choices of S_n, which can consist of any number of subintervals from 1 to $n + 1$. The subintervals are constrained only by their total length and the distances between them and from them to the endpoints 0 and 1.

The somewhat informal proof will be given by constructing the possible potentials. In one dimension there is no possibility of $\bar{S}$ differing from $\text{supp}(V_*)$, since $\text{supp}(V_*)$ is the set on which the corresponding eigenfunction ψ has its maximum or minimum value, and on the complement ψ is a simple exponential function. Since ψ is not maximized at 0, V must equal 0 on some interval beginning at 0. Since $\psi \in C^1$, its first chance to attain its maximum occurs when $\sin(\sqrt{E_n}\, x) = 1$, i.e., at

$$x = \pi\sqrt{E_n}/2.$$

At that point the eigenfunction may either be constant for a while or continue oscillating until some later maximum or minimum. It is a matter of utter indifference how long the eigenfunction remains constant after reaching a sinusoidal maximum or minimum, so long as the total length of constancy has the correct value. By the Sturmian theorem, the nth eigenfunction must make $(n + 1)/2$ complete sinusoidal oscillations punctuated by intervals on which it is constant. The total length of the oscillations is $(n + 1)\pi/\sqrt{E_n}$, while from the condition that $V_n \in Q(1, c)$ the total length of the intervals of constancy of ψ is c/E_n (see Fig. 1). Therefore

$$(n + 1)\pi/\sqrt{E_n} + c/E_n = 1.$$

The solution of this is

$$E_n = ((n + 1)\pi + ((n + 1)^2\pi^2 + 4c)^{1/2})^2/4.$$

For instance, the first several maximum eigenvalues are

n	$E_{n,\max}$ $c = 1$	$c = 10$
0	11.784 7490	26.027 5168
1	41.454 2947	57.746 7175
2	90.815 4283	107.899 653
3	159.907 417	177.349 813
4	248.736 090	266.364 685
5	357.302 960	375.039 120

The asymptotic form is $E_{n,\max} \sim ((n + 1)\pi)^2$. For compari-

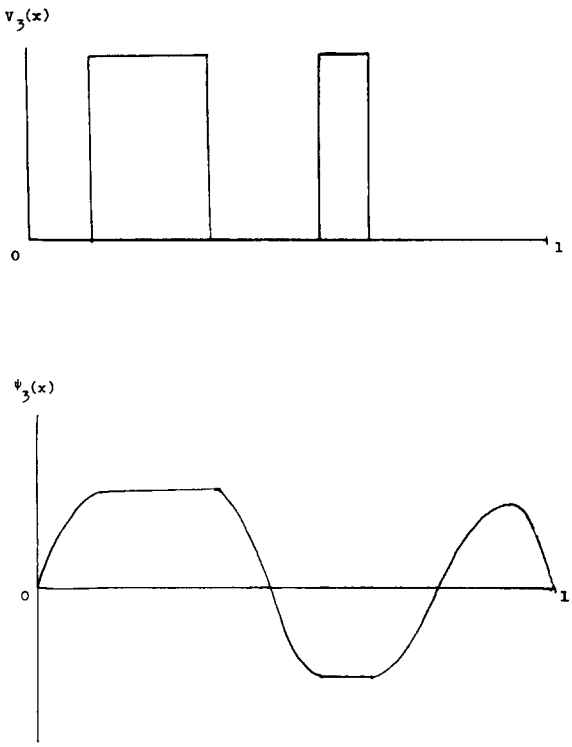

FIG. 1. Typical maximizing potential and eigenfunction for a higher eigenvalue.

son, the bounds on $E_{0,\max}$ from Proposition 1 are, respectively, 11.869 6044 and 29.869 6044 (i.e., $\pi^2 + 2c$), and the lowest eigenvalue with $V = 0$ is $\pi^2 = 9.869\ 6044$. $E_{0,\max}$ has been found by an independent method by Farris.[7]

The maximizing potential for the lowest eigenvalue with Neumann boundary conditions is $V(x) = c$, and the maximizers of the higher Neumann eigenvalues are obtained by an argument analogous to the above.

Similarly, if $n > 1$ and D is a regular figure, such as a cube, sphere, ellipsoid, etc., it is highly probable that the maximal lowest eigenvalue is attained when S is a smaller concentric figure of similar shape, and the maximum eigenvalues can be obtained explicitly in terms of the special functions associated with the separated Laplacian.

This is certainly true of the sphere. Let $p = 1$ and let D be the unit sphere in $\mathbb{R}^m$. The maximizing potential for the lowest eigenvalue is of the form

$$V_*(x) = E_{\max}\, \chi_S(x).$$

The set $\bar{S}$ in this case is again equal to $\mathrm{supp}(V_*)$ and must be a concentric sphere. This is because a spherical average of all rotations of any putative V_* would lead to at least as high an E_*, as seen above. Yet $\mathrm{supp}(V_*)$ cannot be hollow without violating the minimum principle for the superharmonic ground-state eigenfunction on $\mathrm{supp}(V_*)^c$.

It follows that the eigenvalue equation is separable and reduces to the one-dimensional equation

$$-R''(r) - (m-1)R'(r)/r + (V_*(R) - E_{\max})R(r) = 0,$$

which is just a form of Bessel's equation, with solutions

$$R(r) = r^{1-m/2}\mathscr{C}_{m/2-1}(\sqrt{E_{\max} - V_*}\, r)$$

on the interval $[0, r_0]$ on which V_* is constant, where $\mathscr{C}$ is any of the usual Bessel functions of index $m/2 - 1$. Consequently, $E_{\max}$ is the unique solution of the following triple of equations in three unknowns, $E_{\max}$, r_0, and a:

$$E_{\max}\, \omega_m\, r_0^m = c,$$

$$J_{m/2-1}(\sqrt{E_{\max}}) + aY_{m/2-1}(\sqrt{E_{\max}}) = 0 \quad \text{(first zero)},$$

$$\frac{d}{dr}\left(r^{1-m/2}(J_{m/2-1}(\sqrt{E_{\max}}\, r) + aY_{m/2-1}(\sqrt{E_{\max}}\, r))\right)\Big|_{r=r_0} = 0,$$

where ω_m is the volume of the m-sphere. In dimension $m = 3$, the Bessel functions reduce to circular functions, and the equations may be written

$$4\pi r_0^3\, E_{\max}/3 = c,$$

$$\sqrt{E_{\max}} + \phi = \pi,$$

$$\tan(\sqrt{E_{\max}}\, r_0 + \phi) = \sqrt{E_{\max}}\, r_0.$$

These are easy to solve numerically. For example, with $c = 1$,

$$E_{\max} \doteq 11.024\ 7609.$$

(The lowest eigenvalue with $V = 0$ is $\pi^2 \doteq 9.869\ 6044$, and the upper bound from Proposition 1 is $\pi^2 + \pi/2 \doteq 11.440\ 4007$.)

[1]M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, in four volumes (Springer, New York, 1972–1979).
[2]Problem list of A. G. Ramm in H. Samelson, "Queries," Notices Am. Math. Soc. **29**, 326–329 (1982).
[3]B. M. Levitan and M. G. Gasymov, "Determination of a differential equation by two of its spectra," Uspehi Mat. Nauk **19**(2), 3–63 (1964) [Russ. Math. Surveys **19**(2), 1–63 (1964)]. Curiously, Levitan and Gasymov state the theorem in Ambarzumian's original form, assuming that all $E_n = n^2$, although they prove this much more powerful version.
[4]T. Kato, *Perturbation Theory for Linear Operators, Die Grundlehren der mathematischen Wissenschaften*, Vol. 132 (Springer, New York, 1966). The Feynman–Hellmann theorem is not identified as such, but is equation (3.18) on p. 391.
[5]K. Yôsida, *Functional Analysis, Die Grundlehren der mathematischen Wissenschaften*, Vol. 123 (Springer, Heidelberg, 1965).
[6]W. Rudin, *Real and Complex Analysis*, 2nd ed. (McGraw–Hill, New York, 1974).
[7]M. Farris, "A Sturm–Liouville problem with maximal first eigenvalue," preprint, 1982.

Stochastic path-ordered exponentials

Jürgen Potthoff
Fakultät für Physik, Universität Bielefeld, D-4800 Bielefeld 1, Federal Republic of Germany

(Received 20 July 1982; accepted for publication 10 December 1982)

We prove convergence of an approximation of the stochastic product integral for conditional Wiener paths to the solution of a certain stochastic integral equation. This is used to establish the Wiener–Itô representation for the kernel of the semigroup $\exp t\Delta_A$, where $\Delta_A = \Sigma_\mu (\partial_\mu \mathbb{1} + A_\mu)^2$ for functions A_μ with values in the space of anti-Hermitian matrices.

PACS numbers: 02.50.Ey, 02.30. – f

I. INTRODUCTION

The aim of this paper is to construct a (symmetrized) stochastic product integral w.r.t. the D-dimensional conditional Wiener path $\mathbf{Z}$ starting at $\mathbf{x}$ at time zero, ending at $\mathbf{y}$ at time t. The product integral is defined as the limit of a polygonal approximation and we show convergence of this approximation to the solution of a certain stochastic integral equation w.r.t. $\mathbf{Z}$.

The existence of this product integral, which we suggestively denote by $\overleftarrow{\prod}_{s<t} \exp \mathbf{A}(Z_s)\cdot d\mathbf{Z}_s, \overleftarrow{\prod}$ denoting a product whose factors are ordered with increasing time to the left, allows writing the Wiener–Itô representation of the kernel of the semigroup $\exp t\Delta_{\mathbf{A}}$, $t \geqslant 0$, with $\Delta_{\mathbf{A}} = \Sigma_{\mu=1}^{D} (\partial/\partial x_\mu + A_\mu)^2$, on $L^2(\mathbb{R}^D, \mathbb{C}^m)$:

$$(\exp t\Delta_{\mathbf{A}})(x,y) = \int dP_{\mathbf{xy}}^t \overleftarrow{\prod_{s<r}} \exp(\mathbf{A}(\mathbf{Z}_s)\cdot d\mathbf{Z}_s), \tag{1.1}$$

where A is a D-tuple of continuous functions such that div $\mathbf{A}$ is continuous, with values in the space of anti-Hermitian $m \times m$ matrices and $dP_{\mathbf{xy}}^t$ is the conditional Wiener measure.

This formula, which turned out to be very useful in Euclidean quantum field theory and whose proof for the case $m = 1$ can be found in Ref. 1, Chap. V, appears already in several papers.[2–6] A discussion of the proof for $m \geqslant 1$ is found in Refs. 2 and 3; however, there both authors construct the product integral for the Brownian path without fixed endpoint and restrict the integration over these paths [cf. (1.1)] to those with fixed endpoint. Unfortunately the product integral for Brownian paths is defined only up to sets of measure of zero, so that the validity of their discussions is not clear, since the conditioned paths $\mathbf{Z}$ from a set of measure zero.

Stochastic product integrals for Brownian motion have been studied by several authors (see Refs. 7–9 and literature quoted there).

A basic tool of these works is to use the independence of the increments of the Wiener process of the past, i.e., its martingale property, which does not hold for the $\mathbf{Z}$-process. Although Simon[1] has shown how one can overcome this difficulty for defining stochastic integrals w.r.t. $\mathbf{Z}$ by an appropriate decomposition of the increments, this is not sufficient to generalize the proofs presented in Refs. 7 and 9.

Actually, in this paper we have to make use of the ideas of the Strasbourg school[10–13]—in particular Emery has already developed a theory of stochastic product integrals w.r.t. semimartingales and their related integral equations[10] in a very general framework.

On the other hand the $\mathbf{Z}$-process is simple enough (e.g., it is almost surely continuous) to allow for a detailed treatment without going through all the complications provided by the general situation. In this sense part of the present paper can be understood as an illustration (with some modifications) of the ideas found in Ref. 10 and in the beautiful book of Métivier and Pellaumail.[12]

Instead of working directly with the $\mathbf{Z}$-process we prefer to work with the D-dimensional Brownian bridge $\mathbf{W}$, which is related to $\mathbf{Z}$ via

$$\mathbf{Z}_s \doteq (1 - s/t)\mathbf{x} + s/t\,\mathbf{y} + \sqrt{t}\,\mathbf{W}_{s/t}, \quad 0 \leqslant s \leqslant t,$$
$$\int dP_{\mathbf{xy}}^t \doteq (2\pi t)^{-D/2} \exp\left(-\frac{|\mathbf{x}-\mathbf{y}|^2}{2t}\right) E(\cdot), \tag{1.2}$$

where $\doteq$ means equality in sense of probability distributions and $E(\cdot)$ denotes expectations; i.e., $\mathbf{W}_s$, $0 \leqslant s \leqslant 1$, is the Gaussian process [over a probability space $(\Omega, \mathcal{F}, P)$] of mean zero and covariance matrix $E(\mathbf{W}_s \mathbf{W}_t) = \mathbb{1}_D s(1-t)$ for $0 \leqslant s \leqslant t \leqslant 1$, $\mathbb{1}_D$ denoting the D-dimensional unit matrix.[14]

The paper is organized as follows. In Sec. II we discuss some preliminary material; in Sec. III we study integral equations w.r.t. $\mathbf{W}$ and show convergence of the product integral. Finally in Sec. IV we prove the Wiener–Itô representation for the kernel of $\exp t\Delta_{\mathbf{A}}$ as given in (1.1).

II. PRELIMINARY RESULTS[15]

As mentioned in the Introduction the problem in dealing with the Brownian bridge W comes from the dependence of its increments of the past. Simon[1] has shown how to bypass this difficulty using the decomposition

$$W_{t+\Delta t} - W_t = \left(W_{t+\Delta t} - \frac{1-(t+\Delta t)}{1-t} W_t\right) - \Delta t \frac{1}{1-t} W_t, \tag{2.1}$$

so that the increment in ( ) on the rhs is past independent. However, in this paper we need some more detailed information about the dW-integral than is available in Ref. 1, such as continuity of $\int_0^t \cdot dW_s$ in t.

 0022-2488/84/010052-05$02.50

We note that an "integrated version" of (2.1) reads

$$W_t = B_t - \int_0^t ds(1-s)^{-1}W_s, \quad 0\leqslant t\leqslant 1, \tag{2.2}$$

where B_t has the same probability distribution as the standard Brownian motion b_t ($B_t \doteq b_t$), as one easily checks.

Let the underlying probability space of the theory be denoted by $(\Omega,\mathcal{F},P)$ and let the filtration of σ-subalgebras generated by b_t be $(\mathcal{F}_t')_{t>0}$ (i.e., b is an $(\mathcal{F}_t')$-martingale). Then Jeulin[11] shows that B_t is measurable with respect to the enlarged filtration $(\mathcal{F}_t)_{t>0}$, where $\mathcal{F}_t = \mathcal{F}_t' \vee \sigma(b_1)$, $\sigma(b_1)$ denoting the subalgebra generated by b_1. Thus the filtration $(\mathcal{F}_t)_{t>0}$ is the "natural" one in this framework and in fact B is an $(\mathcal{F}_t)$-martingale, so that by (2.2) W_t is an $(\mathcal{F}_t)$-semimartingale.[11,16]

Henceforth measurability is understood w.r.t. $\mathcal{F}$ or $(\mathcal{F}_t)$ depending on the context.

The representation (2.2) allows now for an easy adaption of the construction of stochastic integrals $\int_0^t X_s\,dW_s$ as, e.g., in McKean's book[9] for nonanticipating functionals X of W (i.e., X_s is $\mathcal{F}_s$-measurable for $0\leqslant s\leqslant 1$) satisfying some suitable boundedness condition (see below).

Obviously we have the bound

$$E\left(\left(\int_0^t X_s\,dW_s\right)^2\right)\leqslant 2\left[E\left(\int_0^t X_s^2\,ds\right) + E\left(\left(\int_0^t X_s W_s(1-s)^{-1}ds\right)^2\right)\right], \tag{2.3}$$

and using Hölder's inequality it can be shown that for X such that $E(\int_0^1 |X_s|^{2+\epsilon}ds) < \infty$, for any $\epsilon > 0$, one can define $\int_0^t X_s dW_s$, $0\leqslant t\leqslant 1$ as an integral continuous in t.
All this generalizes now naturally to the case of D-dimensional Brownian bridge $\mathbf{W}$ (i.e., D independent copies of W) and X taking values in some Banach space $\mathcal{H}$ with norm $\|\cdot\|$.

We shall have to use the following

Definition 2.1: A stopping time u is a map $u:\Omega\to[0,1]$ so that $\{\omega;u(\omega)\leqslant t\}\in\mathcal{F}_t$ for every $t\in[0,1]$.

A stochastic interval $[u,v)$, for two stopping times u,v is the set $\{(\omega,t);u(\omega)\leqslant t<v(\omega)\}\subset\Omega\times[0,1]$. $[u,v]$, (u,v), $(u,v]$ are defined similarly.

If X is a process with values in $\mathcal{H}$ and if u is a stopping time, denote $X_u^* = \sup_{0<t<u}\|X_t\|$.

For a D-tuple of processes $\mathbf{X}$, whose components X_μ take values in $\mathcal{H}$, we let $\|X_t\|^2 \equiv \Sigma_\mu \|X_{\mu t}\|^2$ and define X_u^* similarly.

Using the fact that B is a continuous $(\mathcal{F}_t)$-martingale the results of Sec. 6.9 of Ref. 12 imply for $Z_t := \int_0^t \mathbf{X}_s\cdot d\,\mathbf{B}_s$[17] the bound

$$E\left(\sup_{t<u}\|Z_t\|^2\right)\leqslant 8E\left(\int_0^u \|X_s\|^2 ds\right). \tag{2.4}$$

The following theorem is a generalization of the preceding consideration.

Theorem 2.2: Let $\mathbf{X}$ be a D-tuple of $\mathcal{H}$-valued processes so that $E(\int_0^1 \|\mathbf{X}_s\|^{2+\epsilon}ds)$ is finite; then one can define the stochastic integral $\int_0^t \mathbf{X}_s\cdot d\,\mathbf{W}_s$ as a continuous function of t. Moreover, one has the estimate

$$E\left(\sup_{t<u}\left|\left|\int_0^t \mathbf{X}_s\cdot d\,\mathbf{W}_s\right|\right|^2\right) \leqslant E\left(Q_u\int_0^u \|\mathbf{X}_s\|^2(1-s)^{-1/2}ds\right) \tag{2.5}$$

for any stopping time u, where Q denotes the continuous, increasing process

$$Q_t = 16\left(1+\int_0^t |\mathbf{W}_s|^2(1-s)^{-3/2}ds\right), \quad 0\leqslant t\leqslant 1.$$

Remark: Continuity of Q is due to the fact that the integral exists for all $t\in[0,1]$ as a consequence of Hölder continuity of the Brownian bridge W.

(2.5) is similar to what is called "π^*-property" in Ref. 12.

Let us conclude this section by the observation that if $\mathbf{X}$ has the form $\mathbf{X}_s = \mathbf{X}(\mathbf{W}_s)$ then the condition $E(\int_0\|\mathbf{X}_s\|^{2+\epsilon}ds) < \infty$ (in order to define $\int_0^t \mathbf{X}_s\cdot d\,\mathbf{W}_s$) can be replaced by $\mathbf{X}\in L^p_{\text{loc}}$, $p>2$ if $D=1$, and $p>D$ if $D\geqslant 2$, as Hölder's inequality and the use of continuity of $\mathbf{W}$ show.

III. CONVERGENCE OF THE PRODUCT INTEGRAL

For the rest of the paper we let $\mathcal{H}$ be the Banach space of complex $m\times m$ matrices, $\mathbb{1}$ representing the unit matrix equipped with the operator norm $\|\cdot\|$ on $\mathbb{C}^m$.

The central result of this section is to define the product integral by an approximation which is shown to converge to the (unique) solution of a certain stochastic integral equation. Let us begin with a short study of a class of integral equations, which is an adaptation of the very general theory in Ref. 12 to our simple situation.

Consider the equation (let $D=1$ for notational convenience)

$$X_t = \mathbb{1} + \int_0^t dW_s A_s X_s, \quad t\in[0,1] \tag{3.1}$$

for (nonanticipating) A with values in $\mathcal{H}$.

We can state the following

Lemma 3.1: Let A be such that $E(\int_0^1 \|A_s\|^2(1-s)^{-1/2}ds) < \infty$.

Then the integral equation (3.1) admits a unique solution.

The proof of this lemma has two steps. First one shows that (3.1) has a unique solution on a sufficiently small stochastic interval $[0,u]$ (using the Banach fixed point theorem). Then one extends the solution globally by $[0,1]$.

Define a stopping time u by[18]

$$u = \inf\left\{t;t\geqslant 0, Q_t\int_0^t \|A_s\|^2(1-s)^{-1/2}ds > \tfrac{1}{2}\right\}\wedge 1.$$

Let $\mathcal{X}$ be the complete metric space of continuous H-valued processes defined on $[0,u]$, with $X_0 = \mathbb{1}$ for $X\in\mathcal{X}$ and $|||X|||^2 := E(\sup_{t<u}\|X_t\|^2)$ finite. We define a mapping $U:\mathcal{X}\to\mathcal{X}$ by

$$(UX)_t = \mathbb{1} + \int_0^t dW_s A_s X_s. \tag{3.2}$$

By Theorem 2.2, one easily verifies that $\mathcal{D}(U) = \mathcal{X}$:

$$|||UX|||^2 \leqslant 2\left(1 + E\left(\sup_{t<u} Q_t \int_0^t \|A_s\|^2 \|X_s\|^2 (1-s)^{-1/2} ds\right)\right)$$
$$\leqslant 2 + |||X|||^2 < \infty.$$

To prove that U is a contraction let $X, X' \in \mathscr{X}$. Then

$$|||U(X-X')|||^2$$
$$\leqslant E\left(\sup_{t<u} Q_t \int_0^t \|A_s\|^2 \|X_s - X'_s\|^2 (1-s)^{-1/2} ds\right)$$
$$\leqslant \tfrac{1}{2}|||X-X'|||^2,$$

again by Theorem 2.2.

Finally we note that $u > 0$. The condition $E(\int_0^t \|A_s\|^2 (1-s)^{-1/2} ds) < \infty$ implies that $P(\int_0^t \|A_s\|^2 (1-s)^{-1/2} ds > 2^k) \leqslant 2^{-k} \times$ const for every t and k, so that the Borel–Cantelli lemma implies that $\|A_s\|^2 (1-s)^{-1/2}$ is integrable on [0,1] and hence by continuity of Q_t and $\int_0^t \|A_s\|^2 (1-s)^{-1/2} ds$ in t $u > 0$. This concludes the first step.

Note that $|||X|||^2 < \infty$ clearly implies that $\|X_t\| < \infty$ for $t \leqslant u$. Hence, choosing some large $B > 0$, one can extend the solution by the same method as before for all those $\omega \in \Omega$, so that $X^*_u \leqslant B$ and for a new stopping time $u' > u$, so that $Q_t \int_u^t \|A_s\|^2 (1-s)^{-1/2} ds \leqslant \frac{1}{2}$ for $t \leqslant u'$.

This is systematized in the following construction. Define recursively a sequence of stopping times $\{u_k\}_{k>0}$ as follows: $u_0 = 0$; given u_k choose B_k large enough such that $P(X^*_{u_k} > B_k) \leqslant 2^{-k}$. Then if $X^*_{u_k} > B_k$ put $u_{k+1} = u_k$; if $X^*_{u_k} \leqslant B_k$ let

$$u_{k+1} = \inf\left\{t; t \geqslant u_k, Q_t \int_{u_k}^t \|A_s\|^2 (1-s)^{-1/2} ds > \tfrac{1}{2}\right\} \wedge 1.$$

On each stochastic interval one can now apply the contraction mapping principle as above. But as $k \to \infty$ $u_k \to 1$, which proves the lemma.

The lemma is easily generalized to

Theorem 3.2: Consider the D-dimensional Brownian bridge $\mathbf{W}$. Let a D-tuple of nonanticipating functionals $\mathbf{A}$ and a nonanticipating B, $\mathbf{A}$ and B taking values in $\mathscr{H}$, be such that $E(\int_0^1 \|A_s\|^2 (1-s)^{-1/2} ds)$ and $E(\int_0^1 \|B_s\|^2 ds)$ are finite. Then the integral equation

$$X_t = \mathbb{1} + \int_0^t d\mathbf{W}_s \cdot \mathbf{A}_s X_s + \int_0^t ds B_s X_s \tag{3.3}$$

has a unique continuous solution on [0,1].

Remark: As before for $\mathbf{A}_s = \mathbf{A}(\mathbf{W}_s)$, $B_s = B(\mathbf{W}_s)$ the preceding conditions can be replaced by $\mathbf{A} \in L^p_{\text{loc}}$, $B \in L^q_{\text{loc}}$, p as in the remark after Theorem 2.1, $q = 2$ if $D = 1$, $q > D$ if $D \geqslant 2$.

In the following we assume $\mathbf{A}_s = \mathbf{A}(\mathbf{W}_s, s)$ and that $\mathbf{A}$ is C^2 on $\mathbb{R}^D \times [0,1]$, bounded with bounded first and second derivatives.

Define a family of processes on [0,1], indexed by $n \in \mathbb{N}$, as follows:

$$X^n_t = \exp\left[\tfrac{1}{2}(\mathbf{A}_t + \mathbf{A}_{(m-1)/2^n}) \cdot (\mathbf{W}_t - \mathbf{W}_{(m-1)/2^n})\right] \times \prod_{k=1}^{m-1} \exp\left[\tfrac{1}{2}\mathbf{A}_{k/2^n} + \mathbf{A}_{(k-1)/2^n}) \cdot \Delta_k \mathbf{W}\right] \tag{3.4}$$

for

$$t \in \Delta_m := \left[\frac{m-1}{2^n}, \frac{m}{2^n}\right] \quad \text{and}$$
$$\Delta_k \mathbf{W} := \mathbf{W}_{k/2^n} - \mathbf{W}_{(k-1)/2^n}.$$

For later convenience we introduce the following notations: $\Delta_k \mathbf{A} := \frac{1}{2}(\mathbf{A}_{k/2^n} + \mathbf{A}_{(k-1)/2^n}) \cdot \Delta_k \mathbf{W}$; for D vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$, etc. and ∇ the gradient on $\mathbb{R}^D$ $\mathbf{x} \cdot (\nabla \mathbf{y}) \cdot \mathbf{z} = \sum_{\mu,\nu=1}^D x_\mu (\partial_\mu y_\nu) z_\nu$, $((\nabla \mathbf{y}) \cdot \mathbf{z})_\mu = \sum_{\nu=1}^D (\partial_\mu x_\nu) y_\nu$, etc. Also $J^n \mathbf{A}$ denotes the process $(J^n \mathbf{A})_t = \sum_{k=1}^{2^n} 1_{\Delta_k}(t) \mathbf{A}_{(k-1)/2^n}$.

We shall now show that X^n_t converges as $n \to \infty$ uniformly on [0,1] to the solution X_t of Eq. (3.3), B_s being given by $B_s = \frac{1}{2}(\nabla \cdot \mathbf{A}_s + \mathbf{A}_s^2)$. This is done in three steps. First we derive for X^n_t an integral equation of the type previously discussed. Then we show how to reduce the question of convergence of X^n_t to X_t to the question of convergence of their coefficient functions. Finally we prove convergence of the latter.

Proposition 3.3: Let X^n_t be given by (3.4). Then X^n_t is the solution of the integral equation

$$X^n_t = \mathbb{1} + \int_0^t d\mathbf{W}_s \cdot \mathbf{C}_s X^n_s + \int_0^t ds D_s X^n_s, \tag{3.5}$$

where

$$\mathbf{C}_s = \tfrac{1}{2}\{(\nabla \mathbf{A}_s) \cdot (\mathbf{W}_s - (J^n \mathbf{W})_s) + \mathbf{A}_s + (J^n \mathbf{A})_s\},$$
$$D_s = \tfrac{1}{4}\{(\Delta \mathbf{A}_s) \cdot (\mathbf{W}_s - (J^n \mathbf{W})_s) + 2(\nabla \cdot \mathbf{A})_s + \tfrac{1}{2}[(\nabla \mathbf{A}_s) \cdot (\mathbf{W}_s - (J^n \mathbf{W})_s) + \mathbf{A}_s + (J^n \mathbf{A})_s]^2 + 2\left(\frac{\partial}{\partial s}\mathbf{A}\right)_s \cdot (\mathbf{W}_s - (J^n \mathbf{W})_s)\}. \tag{3.6}$$

Proof: The proof is based on an application of Itô's lemma.[19] For $t \in \Delta_m$, $1 \leqslant m \leqslant 2^n$, we compute

$$\int_0^t d\mathbf{W}_s \cdot \mathbf{C}_s X^n_s$$
$$= \sum_{k=1}^m \left\{\frac{1}{2}\int_0^t 1_{\Delta_k}(s) d\mathbf{W}_s \cdot ((\nabla \mathbf{A}_s) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n}) + (\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n})) \exp\left[\tfrac{1}{2}(\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n}) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n})\right]\right\} \prod_{l=1}^{k-1} \exp \Delta_l \mathbf{A} \tag{3.7}$$

using the definition of X^n_t. By Itô's lemma

$$\tfrac{1}{2} d\mathbf{W}_s \cdot ((\nabla \mathbf{A}_s) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n}) + (\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n})) \times \exp\{\tfrac{1}{2}(\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n}) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n})\}$$
$$= d \exp\left[\tfrac{1}{2}(\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n}) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n})\right] - ds 1_{\Delta_k}(s) D_s \exp\left[\tfrac{1}{2}(\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n}) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n})\right]$$

for D as defined before. Inserting this into the rhs of (3.7) gives

$$\int_0^t d\mathbf{W}_s \cdot \mathbf{C}_s X^n_s = X^n_t - \mathbb{1} - \int_0^t ds D_s X^n_s$$

proving the proposition.

Proposition 4.4: Let X_t, X^n_t be as above and u be the following stopping time:

$u = \inf\{t;t>0,$

$$Q_t(\sup_{s<t}(\|X_s\|^2 + \|C_s\|^2 + \|D_s\|^2)) > L\} \wedge 1,$$

L some positive constant. Then we have

$$E\left(\sup_{t<u}\|X_t - X_t^n\|^2\right) \leqslant KE\left(\sup_{t<1}(\|\mathbf{A}_t - \mathbf{C}_t\|^2 + \|B_t - D_t\|^2)\right),$$

where the constant K depends only on L.

Proof: Write

$$\begin{aligned} X_t - X_t^n &= \int_0^t d\,\mathbf{W}_s\cdot(\mathbf{A}_s - \mathbf{C}_s)X_s \\ &\quad + \int_0^t d\,\mathbf{W}_s\cdot\mathbf{C}_S(X_s - X_s^n) \\ &\quad + \int_0^t ds(B_s - D_s)X_s \\ &\quad + \int_0^t ds D_s(X_s - X_s^n), \end{aligned}$$

and by Theorem 2.2 and the definition of u we obtain

$$\begin{aligned} E\left(\sup_{t<u}\|X_t - X_t^n\|^2\right) \leqslant 4\Big\{ & 2LE\left(\sup_{t<q}\|\mathbf{A}_t - \mathbf{C}_t\|^2\right) \\ & + 2LE\left(\int_0^u \|X_s - X_s^n\|^2 \times (1-s)^{-1/2}ds\right) \\ & + LE\left(\sup_{t<1}\|B_t - D_t\|^2\right)\Big\}. \end{aligned}$$

Hence denoting $\phi_t = \sup_{s<t}\|X_s - X_s^n\|^2$ we may bound

$$\begin{aligned} E(\phi_u) \leqslant & 8LE(\sup_{t<1}(\|\mathbf{A}_t - \mathbf{C}_t\|^2 + \|B_t - D_t\|^2)) \\ & + 8LE\left(\int_0^u \phi_s(1-s)^{-1/2}ds\right). \end{aligned} \tag{3.8}$$

The following very simple version of Gronwall's lemma (cf., e.g., Ref. 12) shows that (3.8) implies the proposition:

Let $\{t_k\}_{0<k<k_0}$ be a finite increasing sequence with $t_0 = 0$, $t_{k_0} = 1$ and

$$\int_{t_k}^{t_{k+1}} ds(1-s)^{-1/2} \leqslant (16L)^{-1}.$$

Define a sequence of stopping times $\{v_k\}$ by setting $v_k = t_k \wedge u$. Then (3.8) entails

$$\begin{aligned} E(\phi_{v_{k+1}}) \leqslant & 8LE(\sup_{t<1}(||A_t - C_t||^2 + ||B_t - D_t||^2)) \\ & + 8LE(\phi_{v_k}) + \tfrac{1}{2}E(\phi_{v_{k+1}}), \end{aligned}$$

so that by iteration for every $k \leqslant k_0$,

$$\begin{aligned} E(\phi_{v_{k+1}}) \leqslant & 16LE(\sup_{t<1}(\|\mathbf{A}_t - \mathbf{C}_t\|^2 + \|B_t - D_t\|^2)) \\ & \times \sum_{j=0}^{k_0}(16)^j \end{aligned} \tag{3.9}$$

and (3.9) holds in particular for $E(\phi_u)$.

Proposition 3.5: Let $\mathbf{A}$, B, $\mathbf{C}$, D be as above. Then

$$E(\sup_{t<1}(\|\mathbf{A}_t - \mathbf{C}_t\|^2 + \|B_t - D_t\|^2)) \leqslant \text{const} \times 2^{-n}.$$

Proof: Using the explicit expressions (3.6) and $B_s = \frac{1}{2}(\nabla\cdot\mathbf{A})_s + \mathbf{A}_s^2)$ and Taylor expansion, it suffices to show that

$$E(\sup_{t<1}|\mathbf{W}_t - (J^n\mathbf{W})_t|^2) \leqslant \text{const} \times 2^{-n}.$$

Consider

$$\begin{aligned} & E(\sup_{t<1}|\mathbf{W}_t - (J^n\mathbf{W})_t|^2) \\ & = E(\sup_{\substack{1<k<2^n \\ t\in\Delta_k}}|\mathbf{W}_t - \mathbf{W}_{(k-1)/2^n}|^2) \\ & = E(\sup_{\substack{1<k<2^n \\ t\in\Delta_k}}|\mathbf{W}_{t-(k-1)/2^n}|^2) \\ & = E(\sup_{0<t<2^{-n}}|\mathbf{W}_t|^2) \\ & = E\left(\sup_{0<t<2^{-n}}\left|\int_0^t d\,\mathbf{W}_s\right|^2\right) \\ & \leqslant 2\left(E\left(\int_0^{2^{-n}} ds\right) + E\left(\left(\int_0^{2^{-n}}|\mathbf{W}_s|(1-s)ds\right)^2\right)\right) \\ & \leqslant 6\times 2^{-n}, \end{aligned}$$

where we used (2.3) and (2.4) in the next to last inequality.

Altogether we have found that for u defined as in the hypothesis of Proposition 3.4 the following estimate holds:

$$E(\sup_{t<u}\|X_t - X_t^n\|^2) \leqslant \text{const} \times 2^{-n}.$$

Chebyshev's inequality and the Borel–Cantelli lemma imply now the convergence of X_t^n to X_t uniformly in $t \leqslant u$ as $n \to \infty$. But, for a.e. ω, we have $u = 1$. This follows from the boundedness of the coefficient functions and the continuity properties of Q_t and X_t. We formulate this result in the following

Theorem 3.6: Let $\mathbf{A}$ be a bounded C^2 function with bounded first and second derivatives; then X_t^n (3.4) converges with probability one to the solution X_t of the integral equation

$$X_t = \mathbb{1} + \int_0^t d\,\mathbf{W}_s\cdot\mathbf{A}_s X_s + \frac{1}{2}\int_0^t ds(\nabla\cdot\mathbf{A}_s + \mathbf{A}_s^2)X_s, \tag{3.10}$$

the convergence being uniform in $t\in[0,1]$. The solution of (3.10) is called stochastic product integral or stochastic path ordered exponential, denoted $\overleftarrow{\Pi}_{s<t}\exp \mathbf{A}_s\cdot d\mathbf{W}_s$.

IV. THE WIENER–ITÔ REPRESENTATION

In this last section we shall discuss an application of the results of Sec. III.

For the rest of the paper $\mathbf{A}$ will denote a D-tuple of maps from $\mathbb{R}^D$ into the Banach space of anit-Hermitian $m\times m$ matrices. (The results carry over to the case of real skew-symmetric matrices.) Define the operator $\Delta_{\mathbf{A}} = \Sigma_{\mu=1}^D(\partial/\partial x_\mu + A_\mu)^2$ on the L^2-space of functions on $\mathbb{R}^D$ taking values in $\mathbb{C}^m$ (resp. $\mathbb{R}^m$, in the skew-symmetric case). We quote the

following theorem of Schechter,[20] formulated for scalar **A**, generalized to the matrix-valued situation by Schrader.[6]

Theorem 4.1: Let **A** be such that

(i) $A_\mu \in L^4_{\text{loc}}, \quad 1 \leqslant \mu \leqslant D,$

(ii) $\nabla \cdot \mathbf{A} \in L^2_{\text{loc}},$

(iii) $$\sup_x \int_{|\mathbf{x}-\mathbf{y}|<1} \|\mathbf{A}(\mathbf{y})\| \, |\mathbf{x}-\mathbf{y}|^{-D+1} d^D y < \infty.$$

Then $\Delta_{\mathbf{A}}$ is nonpositive on $L^2(\mathbb{R}^D, \mathbb{C}^m)$ and essentially self-adjoint on $C_0^\infty(\mathbb{R}^D, \mathbb{C}^m)$.

Consider the contraction semigroup $\exp t\Delta_{\mathbf{A}}$, $t \geqslant 0$. By standard methods, e.g., Refs. 1, 21, and 22, we have

$$(\exp t\Delta_{\mathbf{A}})(\mathbf{x},\mathbf{y}) = \lim_{n\to\infty} \int dP^t_{\mathbf{xy}} X^n_T(\mathbf{Z}),$$

$$X^n_t(\mathbf{Z}) = \prod_{m=1}^{2^n} \exp\left[\tfrac{1}{2}(\mathbf{A}(\mathbf{Z}_{mt/2^n}) + \mathbf{A}(\mathbf{Z}_{(mt-t)/2^n})) \cdot (\mathbf{Z}_{mt/2^n} - \mathbf{Z}_{(mt-t)/2^n})\right] \tag{4.1}$$

as an equality of kernels of operators on $L^2(\mathbb{R}^D, \mathbb{C}^m)$, whenever the limit exists.

Using now relation (1.2) it is easy to see that the results of Sec. III carry over to $X^n_t(\mathbf{Z})$; i.e., by Theorem 3.2 $X^n_t(\mathbf{Z})$ converges as $n \to \infty$ to the solution X_t of the equation

$$X_t = \mathbb{1} + \int_0^t d\mathbf{Z}_s \cdot \mathbf{A}_s X_s + \frac{1}{2}\int_0^t ds(\nabla \cdot \mathbf{A}_s + \mathbf{A}_s^2) X_s \tag{4.2}$$

if A is a bounded C^2-function with bounded first and second derivatives.

Furthermore we have $\|X_t\| \leqslant 1$, since **A** is anti-Hermitian (resp. skew symmetric), so that by Lebesgue's dominated convergence theorem the rhs of (4.1) converges to $\int dP_{\mathbf{xy}} X_t(\mathbf{Z})$.

It is easy now to extend this representation to continuous **A** by the following standard argument[1,6]: let $\mathbf{A}_n$ be a sequence of smooth functions converging to **A** in L^p_{loc}, p as remarked after Theorem 2.2, and let X_t, X^n_t resp., denote the solution of (4.2) with the corresponding coefficients. Then $\Delta_{\mathbf{A}_n}$ converges to $\Delta_{\mathbf{A}}$ in strong resolvent sense, hence the semigroup $\exp t\Delta_{\mathbf{A}_n}$ converges strongly to $\exp t\Delta_{\mathbf{A}}$. An argument parallel to the proof of Proposition 3.4 (with coefficient functions multiplied by the characteristic function of a large ball) shows that X^n_t converges to X_t, hence $\int dP^t_{\mathbf{xy}} X^n_t(\mathbf{Z})$ converges to $\int dP^t_{\mathbf{xy}} X_t(\mathbf{Z})$ by the dominated convergence theorem.

Theorem 4.2: Let **A** be a continuous anti-Hermitian (resp. skew-symmetric) matrix, such that div **A** is continuous. Then we have the representation

$$(\exp t\Delta_{\mathbf{A}})(\mathbf{x},\mathbf{y}) = \int dP^t_{xy} \prod_{s<t} \exp \mathbf{A}_s \cdot d\mathbf{Z}_s, \tag{4.3}$$

where $\Pi_{s<t} \exp \mathbf{A} \cdot d\mathbf{Z}_s$ denotes the solution of (4.2).

This theorem has an obvious

Corollary: Denoting by Δ the Laplace operator in $\mathbb{R}^D$, then we have the inequalities

$$\|(\exp t\Delta_{\mathbf{A}})(x,y)\| \leqslant (\exp t\Delta)(x,y), \tag{4.4a}$$

$$\|(m^2 - \Delta_{\mathbf{A}})^{-1}(x,y)\| \leqslant (m^2 - \Delta)^{-1}(x,y), \tag{4.4b}$$

for nonzero, real m.

Inequalities (4.5) are called Kato's inequalities or diamagnetic inequalities; cf. also Refs. 1, 6, 20, and 23–25.

ACKNOWLEDGMENTS

It is a pleasure to thank P. A. Meyer, W. Plass, H. Rost, E. Seiler, and particularly Ph. Blanchard for helpful discussions.

[1] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).
[2] Yu. L. Daleckii, "Continual integrals associated with certain differential equations and systems," Sov. Math. Dokl. **2**, 259–269 (1961).
[3] Z. Haba, "Feynman–Kac formula for Green functions and determinants in Euclidean gauge theories," Wroclaw Preprint 519, 1980.
[4] J. Potthoff, "Euclidean φ^4_3 theory in electromagnetic potential," Bielefeld preprint, 1981 (to appear in Ann. Inst. Henri Poincaré).
[5] J. Potthoff, "Euclidean φ^4_3 theory in an external Yang–Mills field," Thesis, Freie Universität, Berlin, 1980.
[6] R. Schrader, "Towards a constructive approach of a gauge invariant, massive $P(\phi_2)$ theory," Commun. Math. Phys. **58**, 299–312 (1978).
[7] M. Ibero, "Intégrales stochastiques multiplicatives et construction des diffusion sur un groupe de Lie," Bull. Sci. Math. **100**, 175–191 (1976).
[8] H. P. McKean, "Brownian motion on the 3-dimensional rotation group," Mem. Coll. Sci. Kyôto Univ. **33**, 25–38 (1960).
[9] H. P. McKean, *Stochastic Integrals* (Academic, New York, 1969).
[10] M. Emergy, "Stabilité des solutions des équations differentielles stochastique, application aux integrales multiplicatives stochastiques," Z. Wahrscheinlichkeitstheorie **41**, 241–262 (1978).
[11] Th. Jeulin, *Semimartingales et grossisement d'une filtration, Springer Lecture Notes in Mathematics*, Vol. 833 (Springer, Berlin, 1980).
[12] H. Métivier and J. Pellaumail, *Stochastic Integration* (Academic, New York, 1980).
[13] Séminaire de Probabilités, Université de Strassbourg, *Springer Lecture Notes in Mathematics*, Vols. X–XIV (Springer-Verlag, Berlin, 1965, 1975, 1966, 1966, 1966).
[14] **W** presents a D-dimensional Wiener path starting at time zero at the origin, ending there at time one.
[15] To avoid endless repetitions, in the whole paper statements are understood to hold with probability one (a.s).
[16] A semimartingale is a process which admits a decomposition into the sum of a local martingale and a process of finite variation.
[17] $\cdot$ denotes the scalar product in d dimensions.
[18] $u \wedge v := \min(u,v)$.
[19] It is not hard to see that Itô's lemma (see, e.g., Ref. 9) holds for W too.
[20] M. Schechter, "Essential selfadjointness of the Schrödinger operator with magnetic vector potential," J. Funct. Anal. **20**, 93–104 (1975).
[21] D. G. Babitt, "Wiener integral representation for certain semigroups which have infinitesimal generators with matrix coefficients," J. Math. Mech. **19**, 1051–1067 (1970).
[22] E. B. Dynkin, *Markov Processes* (Academic, New York, 1965).
[23] D. C. Brydges, J. Fröhlich, and E. Seiler, "On the construction of quantized gauge fields," Ann. Phys. **121**, 227 (1977).
[24] H. Hess, R. Schrader, and D. Uhlenbrock, "Domination of semigroups and generalization of Kato's inequality," Duke Math. J. **44**, 833–904 (1977).
[25] B. Simon, "Abstract Kato's inequality and the comparison of positivity preserving semigroups," Indiana Math. J. **26**, 1067–1073 (1977).

Path integrals in parametrized theories: Newtonian systems

James B. Hartle
Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637

Karel V. Kuchař
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 10 March 1983; accepted for publication 13 May 1983)

Constraints in dynamical systems typically arise either from gauge or from parametrization. We study Newtonian systems moving in curved configuration spaces and parametrize them by adjoining the absolute time and energy as conjugate canonical variables to the dynamical variables of the system. The extended canonical data are restricted by the Hamiltonian constraint. The action integral of the parametrized system is given in various extended spaces: Extended configuration space or phase space and with or without the lapse multiplier. The theory is written in a geometric form which is manifestly covariant under point transformations and reparametrizations. The quantum propagator of the system is represented by path integrals over different extended spaces. All path integrals are defined by a manifestly covariant skeletonization procedure. It is emphasized that path integrals for parametrized systems characteristically differ from those for gauge theories. Implications for the general theory of relativity are discussed.

PACS numbers: 03.20. + i, 02.30. + g, 03.65. − w

1. MOTIVATION

The most straightforward way to describe an evolving classical system is to give its true dynamical degrees of freedom q^a, p_a, $a = 1,...,n$, as functions of the physical time t. The most straightforward way to describe an evolving quantum system is to give its state ψ on the physical configuration space as a function of the physical time.

The actual classical path of the system extremizes the action functional

$$s[q(t)] = \int_{t'}^{t''} dt\, l(t,q,d_t q) \tag{1.1}$$

in configuration space or the canonical action functional

$$s[q(t), p(t)] = \int_{t'}^{t''} dt\, (p_a\, d_t q^a - h(t,q,p)) \tag{1.2}$$

in phase space. In quantum theory, the state function $\psi(t',q')$ at t' is evolved into the state function $\psi(t'',q'')$ at t'' by the quantum propagator $\langle t'',q''|t',q'\rangle$,

$$\psi(t'',q'') = \int d^n q' \langle t'',q''|t',q'\rangle \psi(t',q'). \tag{1.3}$$

The connection between quantum theory and the classical theory is established when we represent the quantum propagator as an integral over all paths connecting t',q' with t'',q'' in the configuration space,[1]

$$\langle t'',q''|t',q'\rangle\, d^n q' = \int \bar{D}q\, e^{is[q(t)]}, \tag{1.4}$$

or as an integral over all paths connecting t',q' with t'',q'' in the phase space,

$$\langle t'',q''|t',q'\rangle\, d^n q' = \int Dq\, Dp\, e^{is[q(t),\, p(t)]}. \tag{1.5}$$

The transition from classical theory to quantum theory thus amounts to an interpretation of the formal expressions (1.4) or (1.5). To do that, one must explain what is meant by integrating the exponentiated classical action functionals (1.1) or (1.2) and what are the measures $\bar{D}q$ or $Dq\,Dp$ in the space of paths. Both problems can be solved by a skeletonization procedure. In configuration space, the skeletonization of the action functional is obvious: $s[q(t)]$ is replaced by a sum of Hamilton's principal functions for individual segments of the skeletonized path. However, the choice of the skeletonized measure is not obvious. One can use different measures and these measures yield different propagators.[1] This ambiguity corresponds exactly to factor ordering in Hamiltonian quantum mechanics: The Hamilton operators in Schrödinger's equation for the propagators differ by curvature terms of the order $\hbar^2$.

In the phase space path integral, the situation is reversed. The invariant Liouville measure $d^n q\, d^n p$ in the phase space induces a natural measure in the space of skeletonized paths. On the other hand, the skeletonization of the canonical action (1.2) by a sum of phase space principal functions is not unique.[2] Different principal functions yield the same classical dynamics but because nondifferentiable paths are the most significant contributors to the path integral (1.5) they do not yield equivalent quantum dynamics. The advantage of Eq. (1.5) over Eq. (1.4) is that the measure is fixed, and the ambiguity is shifted to the skeletonization of the canonical action where it can be resolved by applying geometric criteria.

The clarity achieved by formulating a physical theory in terms of its true dynamical degrees of freedom is often at the expense of obscuring its fundamental symmetries. Examples of this statement are found in gauge theories and in parametrized theories. The symmetries in these two cases are, of course, gauge invariance and parametrization invariance. For complicated gauge and parametrized systems, it is often impractical or even impossible to exhibit the true dynamical degrees of freedom explicitly. It is thus imperative to have a procedure for passing from the classical version to the quantum version of the theory in its symmetry revealing form. We shall briefly discuss one example of a gauge system

 0022-2488/84/010057-19$02.50

and one example of a parametrized system to get a feeling for the problem.

Though one can easily concoct finite-dimensional gauge theories, the best known specimen of gauge theories is a field system, namely, Maxwell's electrodynamics. The gauge invariance and the Lorentz invariance of this theory are readily seen when the field action is expressed as a functional of the 4-potential $A_\alpha(x)$, $\alpha = 0,1,2,3$. Due to gauge invariance the variables A_α are redundant. However, it is extremely cumbersome to describe the field by its two physical degrees of freedom which are the transverse components of $A_\alpha(x)$, especially in the presence of interactions. The desirability of a quantization procedure which operates at the level of unphysical variables $A_\alpha(x)$ is readily seen. For gauge theories such procedures have been extensively developed.[3] The path integrals are of the same form as Eq. (1.4) or (1.5), but the paths lie in the configuration or phase space augmented by gauge variables. The central issue of these formulations is then specifying the measure which reproduces the physical predictions of the theory. This measure is often quite complicated and difficult to guess from first principles.

As a consequence of gauge invariance, the electric field strength $E^a(x)$ cannot be freely specified, but on each spatial hypersurface it is subject to the constraint

$$C(x^a) = \partial_a E^a = 0. \tag{1.6}$$

In the Hamiltonian version of the theory, $E^a(x)$ is the momentum conjugate to the vector potential $A_a(x)$. The constraint (1.6) is the price we pay for the freedom to perform the gauge transformations.

Another important but quite distinct class of theories with internal symmetries are parametrized theories. The invariance with respect to reparametrization is achieved by adjoining the physical time to the dynamical variables of the system. An arbitrary parameter is then used to locate the system on its dynamical path. Any field theory on a given background can be cast into a parametrized form, but the best known example of a parametrized theory is a finite-dimensional system, namely, the free relativistic particle. Let us discuss the canonical version of the theory. The canonical action (1.2) of the particle is expressed as a functional of the spatial coordinates $q^a(t)$ and their conjugate momenta $p_a(t)$ considered as functions of the Minkowskian time t in a given inertial frame:

$$s[q^a(t), p_a(t)] = \int_{t'}^{t''} dt\,(p_a\, d_t q^a - (\delta^{ab} p_a\, p_b + m^2)^{1/2}). \tag{1.7}$$

In the physical variables q^a, p_a and with the fixed parametrization t, it is difficult to discuss the Lorentz invariance and the reparametrization invariance of the theory. However, if we let t be a function of a parameter τ (not necessarily the proper time) and introduce the Minkowskian time $t = q^0(\tau)$ and energy $-p_0(\tau)$ as dynamical variables, we can write the action (1.7) in the form

$$S[q^\alpha(\tau), p_a(\tau)] = \int_{\tau'}^{\tau''} d\tau\, p_\alpha\, \dot{q}^\alpha, \tag{1.8}$$

which is manifestly invariant both under Lorentz transformations and under reparametrizations of paths, $\tau \to \tau^* = \tau^*(\tau)$. The momenta p_α cannot be varied freely, but they must lie on the mass shell,

$$H = (1/2m)(p_\alpha\, p^\alpha + m^2) = 0. \tag{1.9}$$

The equations of motion follow from extremizing the action (1.8) subject to the constraint (1.9). The constraint (1.9) is the counterpart of the constraint (1.6) in electrodynamics. It is a consequence of the reparametrization invariance in the same way as the constraint (1.6) is a consequence of gauge invariance.

The most important and also the most intricate system in which both gauge and parametrization are subtly intertwined is general relativity. It may be studied as a Hamiltonian theory by foliating space-time with a family of spacelike hypersurfaces. The foliation is specified by giving the lapse function $N(x,t)$ and the shift vector $N^a(x,t)$. The lapse function determines the normal proper time separation $d\sigma = N(x,t)dt$ between two nearby spatial hypersurfaces t and $t + dt$ and the shift vector $N^a(x,t)$ tells us how to displace the point x^a on the hypersurface t so that by launching from the displaced point $x^a + N^a\, dt$ in the direction perpendicular to the hypersurface t we land at the point x^a of the deformed hypersurface $t + dt$. The canonical variables $g_{ab}(t,x)$ and $p^{ab}(t,x)$ are the intrinsic metric and the extrinsic curvature of the hypersurface t. The gauge transformations of the theory are spatial diffeomorphisms on the hypersurfaces of the foliation. The reparametrization is connected with the change of the foliation. Invariance of the theory under gauge transformations implies the supermomentum constraint

$$H_a(x) = -2p^b_{a|b} = 0, \tag{1.10}$$

on the canonical data $g_{ab}(x)$, $p^{ab}(x)$; the reparametrization invariance implies the super-Hamiltonian constraint

$$H(x) = g^{-1/2}(p_{ab}p^{ab} - \tfrac{1}{2}p^2) - g^{1/2}R = 0. \tag{1.11}$$

Here, $g(x) = \det g_{ab}(x)$, the vertical stroke denotes the covariant derivative on the hypersurface and R is the curvature scalar on the hypersurface.

The gauge and reparametrization changes together with the constraints (1.10)–(1.11) imply that the metric field has only 2^{∞^3} degrees of freedom, i.e., $2{\cdot}2^{\infty^3}$ physical field coordinates and conjugate momenta. The remaining 2^{∞^3} coordinates and momenta play the role of an internal time which distinguishes one hypersurface from another by looking at its intrinsic geometry or extrinsic curvature, and of an internal energy. Unfortunately, no one knows how to write an action for general relativity which involves only the two physical degrees of freedom expressed as functions of the physical time. The best we can do is to work with the extended variables g_{ab}, p^{ab}. General relativity comes to us directly only in the gauged and parametrized form. This is our strongest motivation for studying the relation between gauge and parametrized theories in an attempt to understand their similarities and differences.

The similarities are obvious. Both types of invariance imply constraints. In electrodynamics, we have the diver-

gence equation (1.6). In the parametrized relativistic particle theory, we have the restriction (1.9) of the 4-momentum to the mass shell, and in general relativity, we have the constraints (1.10) and (1.11). Further, the constraints generate the changes of extended canonical variables under corresponding transformations. In electrodynamics, we smear the constraint $C(x)$ by an arbitrary test function $\Lambda(x)$,

$$C_\Lambda \equiv \int d^3x\, \Lambda(x) C(x). \tag{1.12}$$

The Poisson bracket of C_Λ with the extended phase space variables $A_a(x)$, $E^a(x)$, $a = 1,2,3$, generates their gauge transformation,

$$\begin{aligned} \delta A_a(x) &= [A_a(x), C_\Lambda] = A_a(x) - \partial_a \Lambda(x), \\ \delta E^a(x) &= [E^a(x), C_\Lambda] = 0. \end{aligned} \tag{1.13}$$

Similarly, for the relativistic particle the constraint (1.9) determines the change of the canonical variables x^α, p_α under displacement $\delta\sigma$ in proper time,

$$\delta x^\alpha = [x^\alpha, H]\,\delta\sigma, \quad \delta p_\alpha = [p_\alpha, H]\,\delta\sigma. \tag{1.14}$$

Finally, in general relativity we smear the super-Hamiltonian (1.11) by the lapse function $N(x)$ and the supermomentum (1.10) by the shift vector $N^a(x)$:

$$\begin{aligned} H_N &\equiv \int d^3x\, N(x) H(x), \\ H_{\mathrm{N}} &\equiv \int d^3x\, N^a(x) H_a(x). \end{aligned} \tag{1.15}$$

The Poisson brackets

$$\delta g_{ab}(x) = [g_{ab}(x), H_{\mathrm{N}}]\,\delta t, \quad \delta p^{ab}(x) = [p^{ab}(x), H_{\mathrm{N}}]\,\delta t \tag{1.16}$$

yield the changes of the canonical variables $g_{ab}(x)$, $p^{ab}(x)$ when the point x^a is displaced by amount $\delta x^a = N^a\,\delta t$ along the hypersurface, while the Poisson brackets

$$\delta g_{ab}(x) = [g_{ab}(x), H_N]\,\delta t, \quad \delta p^{ab}(x) = [p^{ab}(x), H_N]\,\delta t \tag{1.17}$$

yield the changes of $g_{ab}(x)$, $p^{ab}(x)$ when the hypersurface is deformed by the amount $N\,\delta t$ in the normal direction.

There is, however, an important physical distinction between gauge theories and parametrized theories. For gauge theories the changes generated by the constraints do not change the physical state of the system. They change only the gauge in which it is represented. The true physical degrees of freedom do not change. So, in electrodynamics, $A_a(x)$ is changed by the transformation (1.13), but the field strengths $E^a(x)$ and $H^a(x)$ remain unaffected. By contrast, in parametrized theories the changes induced by the constraints are those associated with the dynamical evolution of the system. The true physical degrees of freedom are moved along the dynamical path. This is clearly seen in Eq. (1.14) for a free relativistic particle. In general relativity, the changes (1.16) generated by the supermomentum leave the intrinsic geometry and the extrinsic curvature of the hypersurface unaffected. The quantities like $\int d^3x\, g^{1/2} R$ or $\int d^3x\, g_{ab}\, p^{ab}$ stay the same. On the other hand, the super-Hamiltonian generates the dynamical evolution of the spatial geometry and of the extrinsic curvature under the normal deformation of the hypersurface [Eq. (1.17)].

The different roles which the constraints resulting from gauge invariance and those resulting from reparametrization invariance play in classical theory have fundamental consequences for the quantum theory. This is because in quantum theory time is clearly distinguished from all other variables and cannot be represented by a Hermitian operator. As a result, the path integral procedure developed in gauge theories to achieve the transition from classical mechanics to quantum mechanics is not directly applicable to parametrized theories. In this paper we shall build a correct procedure for a simple class of parametrized systems and show how it differs from the prescription developed for gauge theories. An understanding of where the two prescriptions differ would seem an essential prerequisite to understanding the quantization of general relativity by path integrals.

The finite-dimensional theory which we have chosen as our model is a nonrelativistic system described by the Hamiltonian

$$h = \tfrac{1}{2} g^{ab}(p_a - \phi_a)(p_b - \phi_b) + \phi. \tag{1.18}$$

The potentials ϕ and ϕ_a depend on the configuration variables q^a, $a = 1,\ldots,n$ and on absolute time t. A curved nondegenerate metric g^{ab} is also a function of these variables. We study a nonrelativistic theory because it contains an easily and uniquely identifiable time variable. We consider curved configuration spaces because the parametrized version of the theory can be expressed in terms of a degenerate curved metric in $n+1$ dimensions and so bears structural similarity to general relativity which is our ultimate theory of interest. To emphasize this similarity, we shall express our results in a manifestly covariant manner using this extended metric. We can thus clearly exhibit the geometric structure these theories possess.

Our starting point is the path integral (1.5) in the physical phase space with the canonical action (1.2) containing the Hamiltonian (1.18). We interpret this path integral by a manifestly covariant skeletonization procedure which leads to the Schrödinger equation for the quantum propagator without additional curvature term. This choice fixes the factor ordering and the quantum theory. Our ending points are path integrals for the same propagator over associated spaces. The simplest of these is the integral (1.4) over paths in the physical configuration space. More important, however, are path integrals corresponding to the parametrized version of the theory.

We parametrize the system by adjoining time and a conjugate momentum to the variables $\{q^a, p_b\}$, forming thus an enlarged configuration space $\{Q^A\}$, $A = 0,\ldots,n$ and phase space $\{Q^A, P_A\}$. The quantum propagator can be expressed as a path integral in the enlarged phase space or in the enlarged configuration space. Each case divides into two, corresponding to the classical choice of how the constraint connected with reparametrization invariance is enforced. It can be enforced either explicitly on the variations of an action or implicitly using a Lagrange multiplier. This possibility is reflected quantum-mechanically in two forms of the path integral for each space of variables: one in which the action is

free from multipliers but the measure includes a δ function of the constraint and a second in which the action contains a multiplier and the measure includes an integration over it. There are thus four forms of the path integral for parametrized theories with the basic Hamiltonian (1.18). This may seem an unnecessary proliferation of possibilities, but each of the four forms of the classical action corresponding to these choices can be actually constructed in general relativity. They are displayed in Table I. It therefore seems appropriate to consider all of them in the simple nonrelativistic systems under consideration.

Our results are thus the six forms for the path integral for the system described by the Hamiltonian (1.18)—two in terms of physical variables and four in terms of extended variables. They are specified by six actions displayed in Table II (Sec. 3) and by six measures summarized in Table III (Sec. 10). They are six equivalent ways for passing from the classical theory to the quantum theory. None of the parametrized versions of this passage correspond to the standard procedures for quantizing gauge theories. We shall discuss this in detail in Sec. 9. This only underlines once again the depth of the issues involved in quantizing gravity.

2. PARAMETRIZED NEWTONIAN SYSTEMS

Our immediate goal is to reformulate classical dynamics of a Newtonian system in an extended phase space. In this process, absolute time and energy are adjoined as conjugate canonical variables to the dynamical variables of the system. The absolute time loses thereby its privileged role in parametrizing paths, and it is replaced by an arbitrary label time. For this reason, the process is called parametrization. With absolute time lifted among the configuration variables, one can introduce arbitrary coordinates in the configuration space-time. This underscores the geometric content of the parametrized theory. To reduce the theory back to its humble physical origins, one should learn how to identify the original physical variables from the geometric structures and reinstate them as privileged variables into the action. This inverse process is summarily called a deparametrization. Our goal is thereby set: First, parametrize the physical theory and geometrize it; second, deparametrize the geometric theory and return to the physical starting point.

We assume that the Newtonian space-time is endowed by a privileged foliation of hypersurfaces whose leaves are instants of absolute time. We label the hypersurfaces by a parameter t, not necessarily coinciding with the pace of a standard clock. We assume that each hypersurface carries a positive-definite metric. We do not insist, however, that this metric be flat or time-independent. We introduce into the space-time an arbitrary congruence of world lines transversal to the time foliation and label them by three coordinates. The congruence represents a choice of reference frame.

The dynamical system which we have in mind might be a single point particle or a system of such particles subject to holonomic though in general rheonomic (time-dependent) constraints. (These constraints have nothing to do with the Hamiltonian constraint we introduce later.) Knowing the masses of the particles and the constraints to which they are subject, we can express the kinetic energy of the system in terms of the generalized coordinates q^a, $a = 1,2,...,n$, and generalized velocities $\dot{q}^a$ and deduce thus the instantaneous metric $g_{ab}(t,q)$ induced in the configuration space $\{q^a\}$ of the system. The system is also subject to forces derivable from a scalar potential $\phi(t,q)$ and a vector potential $\phi_a(t,q)$. We do not need to distinguish "true" forces from "ficticious" forces, which are already contained in the expression for the

TABLE I. Alternative forms of the action for general relativity.

	Canonical variables	Multipliers	Action	Lagrangian, Hamiltonian, constraints
Extended canonical action, conditional	g_{ab}, p^{ab}		$S[g_{ab}, p^{ab}] = \int dt \int d^3x\, p^{ab}\dot{g}_{ab}$,	$H = g^{-1/2}(p_{ab}p^{ab} - \frac{1}{2}p^2) - g^{1/2}R$
			$H(x) = 0 = H_a(x)$	$H_a = -2p^b_{a\|b}$
Extended canonical action, with lapse and shift multipliers	g_{ab}, p_{ab}	N, N^a	$S[g_{ab}, p^{ab}, N, N^a] = \int dt \int d^3x(p^{ab}\dot{g}_{ab} - H(g_{ab}, p^{ab}, N, N^a))$	$H = N(x)H(x) + N^a(x)H_a(x)$
Extended Lagrangian action, with lapse and shift multipliers	g_{ab}	N, N^a	$S[g_{ab}, N, N^a] = \int dt \int d^3x L(g_{ab}, \dot{g}_{ab}, N, N^a)$	$L = Ng^{1/2}[(K_{ab}K^{ab} - K^2) + R] = (-{}^4g)^{1/2}\,{}^4R$ + (divergence terms)
				$K_{ab} = \frac{1}{2}N^{-1}(-\dot{g}_{ab} + N_{a\|b} + N_{b\|a})$
Extended Lagrangian action, without the lapse multipliers[a]	g_{ab}	N^a	$S[g_{ab}, N^a] = \int dt \int d^3x\, L(g_{ab}, \dot{g}_{ab}, N^a)$	$L = [gR(U_{ab}U^{ab} - U^2)]^{1/2}$
				$U_{ab} = \dot{g}_{ab} - N_{a\|b} - N_{b\|a}$

[a] The elimination of N^a would lead to the homogeneous Lagrangian action without multipliers. The elimination cannot be carried out explicitly.

kinetic energy. We thus include both types of terms into the potentials ϕ, ϕ_a.

An elementary example of such a system would be a charged particle moving on an expanding curved surface placed in an external electromagnetic field. The generalized coordinates q^a might be any curvilinear coordinates on the surface. Another example, closer to actual systems studied in nonrelativistic quantum mechanics, would be a charged rigid rotator in an external electromagnetic field. The generalized coordinates q^a might be the Euler angles. The kinetic energy of the rigid rotator expressed as a quadratic form of generalized velocities indicates that the configuration space of the rotator is curved, but the metric is time-independent.

The dynamical evolution of the system takes place in the physical phase space $\{q^a, p_a\}$ which is a cotangent bundle over the physical configuration space $\{q^a\}$. The evolution of physical variables is governed by the canonical action

$$s[q,p] = \int dt\,(p_a\, d_t q^a - h(t,q,p)) \tag{2.1}$$

with the Hamiltonian

$$h(t,q,p) = \tfrac{1}{2} g^{ab}(p_a - \phi_a)(p_b - \phi_b) + \phi. \tag{2.2}$$

A new choice of the time labeling, $t^* = t^*(t)$, or a change of the reference frame changes the Hamiltonian (2.2) into another Hamiltonian of the same type. The only features of the Newtonian system which are important for our purposes are the existence of a privileged foliation of the configuration space-time by leaves of absolute time and the fact that the Hamiltonian of the system is a quadratic function of canonical momenta. There is no need to introduce other features usually associated with Newtonian physics like the presence of the Galilei group.

We now parametrize a possible path along which the system moves in the phase space $\{q^a, p_a\}$ by an arbitrary label time τ and adjoin the originally chosen absolute time $t(\tau)$ to the configuration variables $q^a(\tau)$:

$$Q^A = \{t, q^a\}, \quad Q^A = Q^A(\tau), \quad A = 0,1,2,...,n. \tag{2.3}$$

The action (2.1) takes the form

$$s[Q^A, p_a] = \int d\tau\,(p_a \dot{q}^a - h(Q,p)\dot{t}) \tag{2.4}$$

when written in the τ-parametrization. The dot denotes a derivative with respect to the label time τ. Numerically, the expression (2.4) is equal to the expression (2.1) and so variation with respect to q^a, p_a yields equivalent equations of motion. Moreover, the variation of the parametrized action (2.4) with respect to t also yields a correct equation, namely, the energy balance equation

$$\dot{h} = \partial_t h\, \dot{t}. \tag{2.5}$$

The integrand of the action functional (2.4) is linear in the velocities $\dot{Q}^A = \{\dot{t}, \dot{q}^a\}$. By introducing a momentum $p_0 = -h$ canonically conjugate to t and by putting

$$P_A = \{p_0, p_a\} \tag{2.6}$$

we cast the action (2.4) into a suggestive form

$$S[Q^A, P_A] = \int d\tau\, P_A \dot{Q}^A. \tag{2.7}$$

However, the variables P_A cannot be varied freely, because p_0 is a mere abbreviation for the function $-h(Q^A, p_a)$. To obtain correct equations of motion, we must vary the action (2.7) under the constraint

$$H^{(o)} \equiv p_0 + h(Q^A, p_a) = 0. \tag{2.8}$$

In this way, a constraint on the variables of the enlarged phase space enters into the theory. It is called the Hamiltonian constraint.

The actual path of the system extremizes the action functional (2.7) in comparison to all neighboring paths which lie on the constraint surface (2.8). In other words, the actual path is selected by the conditions

$$H^{(o)}(Q,P) = 0,$$
$$\delta S[Q,P] = 0 \quad \forall\ \delta Q,\ \delta P\colon\ \delta H^{(o)} = 0. \tag{2.9}$$

Equations (2.9) constitute a conditional variational principle.

The constraint function $H^{(o)}$ is a quadratic function of extended momenta P_A. This property is preserved if we multiply the constraint by an arbitrary function $\Lambda(Q^A) > 0$ of extended configuration variables,

$$H \equiv \Lambda(Q) H^{(o)}, \quad H = 0. \tag{2.10}$$

The constraint function $H(Q,P)$ is called a super-Hamiltonian of the system.

We shall now write the constraints (2.8) or (2.10) in a manifestly covariant notation. We introduce a covector field

$$t_A = t_{,A}(Q) = (1;0,...,0) \tag{2.11}$$

normal to the instants of absolute time and a vector field

$$u^A = (1;0,...,0) \tag{2.12}$$

tangent to the world lines $q^a = \text{const}$ of our "configuration reference frame." We collect the potentials into a space-time covector field

$$\phi_A = (-\phi;\phi_a) \tag{2.13}$$

and complete the spatial metric g^{ab} into a degenerate space-time metric

$$g^{AB} = \begin{vmatrix} 0, & 0 \\ 0, & g^{ab} \end{vmatrix}. \tag{2.14}$$

The metric g^{AB} has the signature $(0; +,..., +)$ and t_A is its degeneracy direction,

$$g^{AB} t_B = 0. \tag{2.15}$$

Of course,

$$u^A t_A = 1. \tag{2.16}$$

When g^{AB} and u^A are given, Eqs. (2.15) and (2.16) determine t_A.

The super-Hamiltonian (2.8) can now be written in a manifestly covariant form

$$H^{(o)} = u^A(P_A - \phi_A) + \tfrac{1}{2} g^{AB}(P_A - \phi_A)(P_B - \phi_B). \tag{2.17}$$

After scaling the fields u^A and g^{AB} by the factor $\Lambda(Q)$,

$$G^{AB} \equiv \Lambda g^{AB}, \quad U^A \equiv \Lambda u^A, \tag{2.18}$$
$$H = U^A(P_A - \phi_A) + \tfrac{1}{2} G^{AB}(P_A - \phi_A)(P_B - \phi_B). \tag{2.19}$$

Up to now, the absolute time variable $Q^0 = t$ was clearly separated from the configuration variables $Q^a = q^a$. At this stage, however, we can easily mix the space-time variables Q^A by an arbitrary transformation $Q^{A*}(Q^B)$, inducing thereby a transformation of the conjugate momenta:

$$Q^{A*} = Q^{A*}(Q^B), \quad P_{A*} = Q^B{}_{A*} P_B, \tag{2.20}$$

$$Q^{A*}{}_B \equiv \frac{\partial Q^{A*}}{\partial Q^B}, \quad Q^B{}_{A*} \equiv \frac{\partial Q^B}{\partial Q^{A*}}.$$

When we transform U^A (or u^A) as a vector, G^{AB} (or g^{AB}) as a tensor, and ϕ_A as a covector, the constraint (2.19) [or (2.17)] preserves its form. We shall omit the asterisks with the understanding that Eqs. (2.17)–(2.19) are written in general coordinates. The action principle (2.9) then yields the actual motion of the system in general coordinates.

In the special coordinates $Q^A = \{t, q^a\}$, the coefficients u^A, g^{AB} assume the simplified form (2.12),(2.14). This implies that the scaled coefficients U^A, G^{AB} cannot be arbitrary functions of general coordinates Q^A. In a permissable parametrized Newtonian theory, U^A and G^{AB} must be subject to two sets of restrictions which ensure that the physical theory can be recovered by deparametrization. These restrictions are:

(I) The metric G^{AB} must be degenerate, with signature $(0; +,..., +)$. The degeneracy direction T_A,

$$G^{AB} T_B = 0, \quad T_B \neq 0, \tag{2.21}$$

must be surface-forming. This happens if and only if the metric G^{AB} satisfies the integrability condition (Appendix A)

$$\delta_{AA_1 \cdots A_n} G^{A[B,C]} G^{A_1 B_1} \cdots G^{A_n B_n} = 0. \tag{2.22}$$

(II) The inner product $U^A T_A$ cannot vanish and, for a future-oriented T_A, it must be positive,

$$U^A T_A > 0. \tag{2.23}$$

Equation (2.23) implies that T_A can be normalized so that

$$U^A T_A = 1. \tag{2.24}$$

The parametrized Newtonian system is characterized by a quadratic super-Hamiltonian (2.19) whose coefficients U^A and G^{AB} satisfy our restrictions (I) and (II). We complete our demonstration that the physical and parametrized versions of the theory are equivalent by showing how to deparametrize the system. To do this, we have to find the absolute time function and return back to the physical Hamiltonian (2.2).

Notice first that the quadratic function (2.19) determines the coefficient G^{AB} uniquely, but the coefficients U^A and ϕ_A only up to a gauge transformation

$$*U^A = U^A + G^{AB} \psi_B, \tag{2.25}$$

$$*\phi_A = \phi_A + \psi_A,$$

generated by a gauge variable ψ_A which satisfies the condition

$$\tfrac{1}{2} G^{AB} \psi_A \psi_B + U^A \psi_A = 0. \tag{2.26}$$

The transformation (2.25)–(2.26) expresses an arbitrary change of the configuration reference frame. We have discussed the influence of such a gauge transformation on quantum description of a Newtonian system in an earlier paper.[4] Here, we shall simply assume that one reference field U^A is chosen within the equivalence class (2.25)–(2.26).

Return now to the problem of how to reconstruct the physical Hamiltonian. For the metric G^{AB} with signature $(0; +,..., +)$ all solutions T_A of Eq. (2.21) fill a ray. The integrability condition (2.22) ensures that at least one solution t_A within this ray is a gradient of a scalar function,

$$\exists\ t(Q):\ t_A = t_{,A}. \tag{2.27}$$

In fact, all solutions which are gradients are related to one another by the transformations $t^* = t^*(t)$. We select one which increases to the future, i.e., which satisfies the condition

$$U^A t_{,A} \equiv \Lambda(Q) > 0 \tag{2.28}$$

for our time function $t(Q)$. We then scale the super-Hamiltonian (2.19) down by the factor Λ^{-1}, scaling G^{AB} down to g^{AB} and U^A to u^A by Eq. (2.18). Equation (2.28) then implies Eq. (2.16). Of course, the scaled metric satisfies Eq. (2.15).

We can now introduce within the reference frame u^A comoving coordinates q^a as any n functionally independent solutions $q^a(Q)$ of the equations

$$u^A q^a{}_{,A} = 0. \tag{2.29}$$

We take the time function (2.27) and the comoving coordinates (2.29) as our special coordinates $Q^A = \{t, q^a\}$. Equations (2.16) and (2.29) then ensure that u^A in special coordinates has the components (2.12). Similarly, Eq. (2.15) ensures that the rescaled metric g^{AB} has the components (2.14). Therefore, the rescaled super-Hamiltonian (2.17) reduces back to the form (2.8), where h is our old Hamiltonian (2.2). When we solve the constraint (2.8) with respect to p_0, substitute this solution into the action (2.7), and parametrize paths by the absolute time t, we return back to the physical action (2.1). In this way, we regain the physical action from the parametrized action (2.7) subject to the super-Hamiltonian constraint (2.8).

3. ALTERNATIVE FORMS OF THE ACTION

We have transformed the canonical action (2.1)–(2.2) on the physical phase space into a constrained action (2.7)–(2.10), (2.17), (2.19) on the extended phase space. Besides these forms of the action, there are still others which are frequently used in dynamical considerations. In particular, one can adjoin the Hamiltonian constraint (2.10) to the extended phase space action (2.7) by a lapse multiplier, and one can cast the parametrized action into a Lagrangian form, either on the physical or on the extended configuration space, and either including or excluding the lapse multiplier.

We have argued in the Introduction that any of these forms could serve as the starting point for the transition to quantum theory by path integrals. However, only in the physical phase space do we have a universal prescription for the measure. All other path integrals should be thus derived from the path integral in the physical phase space. To proceed, we must first understand how the various forms of the action are connected to each other. We shall study this clas-

sical problem now and postpone its application to path integrals to subsequent sections.

In the beginning, we replace the conditional variational principle by a free variational principle by adjoining the constraint (2.17) to the action (2.7) by a Lagrange multiplier $N^{(o)}$,

$$S[Q,P,N^{(o)}] = \int d\tau\,(P_A \dot{Q}^A - N^{(o)} H^{(o)}), \tag{3.1}$$

or the scaled constraint (2.19) by a Lagrange multiplier N,

$$S[Q,P,N] = \int d\tau\,(P_A \dot{Q}^A - NH). \tag{3.2}$$

All the variables Q^A, P_A, $N^{(o)}$ or Q^A, P_A, N may now be varied freely.

The physical meaning of the multipliers $N^{(o)}$ or N follows from the Euler–Lagrange equations. By varying Eq. (3.2) in the momenta P_A, we get

$$\dot{Q}^A = N(U^A + G^{AB}(P_B - \phi_B)). \tag{3.3}$$

We multiply Eq. (3.3) by a degeneracy covector T_A, Eqs. (2.21), (2.23), and calculate N:

$$N = (T_B U^B)^{-1} T_A \dot{Q}^A. \tag{3.4}$$

In the special coordinates $Q^A = \{t, q^a\}$, Eq. (3.4) reduces to

$$N = \Lambda^{-1}\dot{t} \tag{3.5}$$

by virtue of Eq. (2.28). The same sequence of steps starting from the action (3.1) leads to the equation

$$N^{(o)} = (t_{,B} u^B)^{-1} t_{,A} \dot{Q}^A = \dot{t}. \tag{3.6}$$

We thus see that the multiplier $N^{(o)}$ equals the rate of change $\dot{t}$ of the absolute time t with respect to the label time τ. For this reason, it is called the lapse function. We shall loosely use this name also for the scaled multiplier (3.5).

The action (3.2) is the best starting point for further rearrangements. We group its arguments into several classes:

extended configuration variables Q^A

$= \{$physical time t, physical coordinates $q^a\}$,

extended momenta variables P_A

$= \{$physical Hamiltonian $- p_0$, physical momenta $p_a\}$,

Lagrange multiplier $= \{$lapse function $N\}$.

By eliminating one or more classes of variables from the action, we cast it into a number of alternative forms which lead to equivalent sets of equations of motion. The transition from the extended action (3.2) to the physical action (2.1) has this character: It is achieved by using the equations of motion to eliminate the lapse multiplier N and the time–energy pair t, p_0 from the action. One can proceed one step further and eliminate all momenta variables from the canonical action (2.1). One arrives then at the physical Lagrangian action

$$S[q] = \int dt\, l(t,q,d_t q) \tag{3.7}$$

by the Legendre dual transformation

$$l(t,q,d_t q) = [\,p_a\, d_t q^a - h(t,q,p)\,]_{p = p(t,q,d_t q)}, \tag{3.8}$$

$$d_t q^a = \frac{\partial h}{\partial p_a}.$$

Because the physical Hamiltonian is nondegenerate, the second equation uniquely determines the generalized momenta p_a in terms of the generalized velocities $d_t q^a$, $p_a = p_a(t,q,d_t q)$. For the Newtonian system (2.2),

$$l(t,q,d_t q) = \tfrac{1}{2} g_{ab} d_t q^a d_t q^b + \phi_a d_t q^a - \phi,$$
$$p_a = g_{ab} d_t q^b + \phi_a. \tag{3.9}$$

Start now from the parametrized canonical action (3.2) instead of from the physical canonical action (2.1). Try to eliminate the momenta P_A, but leave the lapse function N in the action. This time, however, the expression (3.3) for the velocities $\dot{Q}^A$ in terms of the momenta P_A is not invertible because the metric G^{AB} is degenerate. One can, however, go most of the way by defining the covariant metric G_{AB} by the equations

$$U^B G_{BA} = 0, \quad G^{AB} G_{BC} = \delta^A_C - U^B T_C, \tag{3.10}$$

where T_C is the normalized degeneracy covector (2.21), (2.24). The metric G_{AB} is again degenerate, with signature $(0; +,\ldots, +)$. After introducing the abbreviations

$$\phi_{\|} \equiv \phi_A U^A, \quad P_{\|} \equiv P_A U^A, \tag{3.11}$$

we express the momenta P_A in terms of the velocities $\dot{Q}^A$ and a single scalar $P_{\|}$

$$P_A = N^{-1} G_{AB} \dot{Q}^B + (\phi_A + \phi_{\|} T_A) + P_{\|} T_A. \tag{3.12}$$

After the Legendre transformation

$$L \equiv [P_A \dot{Q}^A - NH]_{P_A = P_A(Q,\dot{Q},N,P_{\|})}$$
$$= \tfrac{1}{2} N^{-1} G_{AB} \dot{Q}^A \dot{Q}^B + N\phi_{\|} + (\phi_A - \phi_{\|} T_A)\dot{Q}^A$$
$$+ P_{\|}(T_A \dot{Q}^A - N) \tag{3.13}$$

$P_{\|}$ stays in the action as another Lagrange multiplier. However, it can be eliminated by using the Euler–Lagrange equation obtained by varying the lapse multiplier N,

$$P_{\|} - \phi_{\|} + \tfrac{1}{2} N^{-2} G_{AB} \dot{Q}^A \dot{Q}^B = 0. \tag{3.14}$$

This leads to the Lagrangian

$$L(Q,\dot{Q},N) = (N^{-1} - \tfrac{1}{2} N^{-2} T_C \dot{Q}^C) G_{AB} \dot{Q}^A \dot{Q}^B + \phi_A \dot{Q}^A \tag{3.15}$$
$$= -\tfrac{1}{2}(N^{-1}(T_C \dot{Q}^C)^{1/2} - (T_C \dot{Q}^C)^{-1/2})^2 G_{AB} \dot{Q}^A \dot{Q}^B$$
$$+ \tfrac{1}{2}(T_C \dot{Q}^C)^{-1} G_{AB} \dot{Q}^A \dot{Q}^B + \phi_A \dot{Q}^A. \tag{3.16}$$

It is not difficult to check that by varying Q^A and N we obtain correct equations of motion. In special coordinates $Q^A = \{t, q^a\}$ with the lapse function $N^{(o)} = \Lambda N$ the Lagrangian (3.15) reduces to

$$L(t,q,\dot{q},N^{(o)}) = (N^{(o)\,-1} - \tfrac{1}{2} N^{(o)\,-2}\dot{t})$$
$$\times g_{ab} \dot{q}^a \dot{q}^b + \phi_a \dot{q}^a - \phi \dot{t}. \tag{3.17}$$

As a final transformation, we eliminate the lapse function N from the extended Lagrangian (3.16). The Euler–Lagrange equation obtained by varying N can be solved for N, with the result

$$N = T_A \dot{Q}^A. \tag{3.18}$$

This expression replicates Eq. (3.4) which was obtained from the canonical action. By substituting it back into the Lagrangian (3.16), we get the reduced Lagrangian

$$L(Q,\dot{Q}) = \tfrac{1}{2}(T_C \dot{Q}^C)^{-1} G_{AB} \dot{Q}^A \dot{Q}^B + \phi_A \dot{Q}^A, \tag{3.19}$$

TABLE II. Alternative forms of the action.

	Action	Lagrangian, Hamiltonian, super-Hamiltonian	
Physical canonical action	$s[q,p]=\int dt\,(p_a d_t q^a - h(t,q,p))$	$h=\frac{1}{2}g^{ab}(p_a-\phi_a)(p_b-\phi_b)+\phi$	
Physical Lagrangian action	$s[q]=\int dt\, l(t,q,d_t q)$	$l=\frac{1}{2}g_{ab}\,d_t q^a\, d_t q^b+\phi_a\, d_t q^a-\phi$	
Extended canonical action, conditional	$S[Q,P]=\int d\tau\, P_A\dot{Q}^A,\quad H=0$	General coordinates	$H=U^A(P_A-\phi_A)+\frac{1}{2}G^{AB}(P_A-\phi_A)(P_B-\phi_B)$
Extended canonical action, with lapse multiplier	$S[Q,P,N]=\int d\tau\,(P_A\dot{Q}^A-NH)$	Special coordinates, rescaled	$H^{(0)}=p_0+\frac{1}{2}g^{ab}(p_a-\phi_a)(p_b-\phi_b)+\phi$
Extended Lagrangian action, homogeneous	$S[Q]=\int d\tau\, L(Q,\dot{Q})$	General coordinates	$L(Q,\dot{Q})=\frac{1}{2}(T_C\dot{Q}^C)^{-1}G_{AB}\dot{Q}^A\dot{Q}^B+\phi_A\dot{Q}^A$
		Special coordinates	$L(t,q,\dot{t},\dot{q})=\frac{1}{2}\dot{t}^{-1}g_{ab}\dot{q}^a\dot{q}^b+\phi_a\dot{q}^a-\phi\dot{t}$
Extended Lagrangian action, with lapse multiplier	$S[Q,N]=\int d\tau\, L(Q,\dot{Q},N)$	General coordinates	$L(Q,\dot{Q},N)=-\frac{1}{2}(N^{-1}(T_C\dot{Q}^C)^{1/2}-(T_C\dot{Q}^C)^{-1/2})^2\times G_{AB}\dot{Q}^A\dot{Q}^B+L(Q,\dot{Q})$
		Special coordinates	$L(t,q,\dot{t},\dot{q},N)=-\frac{1}{2}(N^{-1}\dot{t}^{1/2}-\dot{t}^{-1/2})^2\times g_{ab}\dot{q}^a\dot{q}^b+L(t,q,\dot{t},\dot{q})$

which is a homogeneous function of the first degree in the extended velocities $\dot{Q}^A$. In special coordinates $Q^A=\{t,q^a\}$, the homogeneous Lagrangian assumes the form

$$L(t,q,\dot{t},\dot{q})=\tfrac{1}{2}\dot{t}^{-1}g_{ab}\dot{q}^a\dot{q}^b+\phi_a\dot{q}^a-\phi\dot{t}. \tag{3.20}$$

We display a summary of our results for the alternative forms of the action in Table II.

4. PATH INTEGRALS IN PHYSICAL PHASE SPACE

The canonical action (2.1)–(2.2) on physical phase space is a logical starting point for path integration because the privileged Liouville measure $d^n q\, d^n p$ in this space induces a natural measure in space of skeletonized paths. We represent the quantum propagator by a path integral on the physical phase space following the procedure of Ref. 2. In subsequent sections, we transform this path integral into equivalent path integrals corresponding to alternative forms of the action. In this process, nontrivial and often quite complicated measures are induced in alternative spaces of paths.

The Hilbert space of our dynamical system is the space of scalar state functions $\psi(q,t)$ with the scalar product

$$\langle \underset{1}{\psi}|\underset{2}{\psi}\rangle=\int d^n q\, g^{1/2}(t,q)\underset{1}{\psi}{}^*(t,q)\underset{2}{\psi}(t,q). \tag{4.1}$$

Positions q^a and momenta p_a are represented by Hermitian operators

$$\mathbf{q}^a=q^a,\quad \mathbf{p}_a=-ig^{-1/4}\partial_a g^{1/4}. \tag{4.2}$$

The classical Hamiltonian (2.2) is turned into a covariant operator

$$\begin{aligned}\mathbf{h}&=\tfrac{1}{2}g^{-1/4}(\mathbf{q})(\mathbf{p}_a-\phi_a(\mathbf{q}))g^{1/2}g^{ab}(\mathbf{q})\\&\quad\times(\mathbf{p}_b-\phi_b(\mathbf{q}))g^{-1/4}(\mathbf{q})+\phi(\mathbf{q})\\&=-\tfrac{1}{2}\Delta+i(\phi^a\partial_a+\tfrac{1}{2}\phi^a{}_{|a})+\phi+\tfrac{1}{2}\phi^a\phi_a,\end{aligned} \tag{4.3}$$

which is again Hermitian under the norm (4.1). The state function $\psi(t,q)$ is evolved in time by the Schrödinger equation

$$ig^{-1/4}\partial_t(g^{1/4}\psi)=\mathbf{h}\psi. \tag{4.4}$$

The general solution of Eq. (4.4) is provided by the quantum propagator $\langle t'',q''|t',q'\rangle$,

$$\psi(t'',q'')=\int d^n q'\,\langle t'',q''|t',q'\rangle\psi(t',q'). \tag{4.5}$$

This propagator is a scalar in q'' and a scalar density in q'. It satisfies the Schrödinger equation

$$ig''^{-1/4}\partial_{t''}(g''^{1/4}\langle t'',q''|t',q'\rangle)=\mathbf{h}''\langle t'',q''|t',q'\rangle \tag{4.6}$$

with the boundary condition

$$\langle t'',q''|t'',q'\rangle=\delta(q''|q'). \tag{4.7}$$

We represent the quantum propagator by an integral over all phase space paths $q(t)$, $p(t)$ which start in the configuration q' at t' and end in the configuration q'' at t'',

$$\langle t'',q''|t'',q'\rangle\, d^n q'=\int Dq\, Dp\, e^{is[q(t),p(t)]}. \tag{4.8}$$

Here, $s[q(t),p(t)]$ is the canonical action integral (2.1) and $Dq\,Dp$ is a measure in the space of phase space paths.

We interpret the formal expression (4.8) by a skeletonization procedure in which the time between t' and t'' is sliced into small intervals and the measure becomes the product of the Liouville phase space measures on each slice. In the integrand, we need to skeletonize the action for each path in phase space. We replace the action functional by a sum of principal functions for getting from one phase space point on the skeletonized path to the next. These principal functions cannot be the Hamilton principal functions, because Hamilton's principal functions are determined by the initial and

final configurations and do not depend on momentum. A correct construction was discussed in Ref. 2. Evaluate the canonical action (2.1) along the actual path $q^a(t)$ in configuration space and the momentum path $p_a(t)$ found by transporting an arbitrary initial momentum along the configuration space path by a specified rule. There results a principal function $s(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)})$ which depends on the initial and final configurations and on the initial momentum. By summing such principal functions for all segments of the phase space path, one arrives at an action function which is manifestly covariant under point transformations

$$q^{a^*} = q^{a^*}(t,q), \quad p_{a^*} = \frac{\partial q^b(t,q^*)}{\partial q^{a^*}} p_b. \tag{4.9}$$

There are, in fact, a variety of such skeletonization procedures, depending on which rule is used to transport the momentum along the actual classical path. Each gives a different quantum mechanical propagator. We shall use the rule of geodesic deviation transport. There are compelling reasons for such a choice: (1) *A fortiori*, the momentum vector is Lie propagated by a flow of actual configuration paths; (2) *a posteriori*, the Schrödinger equation (4.4) does not contain any curvature term.

Let us now describe this procedure in detail. The skeletonized phase space path $t_{(K)}, q_{(K)}, p_{(K)}$, $K = 0,1,...,N$, starts at the configuration q' at t' and ends in the configuration q'' at t'',

$$t_{(0)} = t', \quad q_{(0)} = q', \quad t_{(N)} = t'', \quad q_{(N)} = q''. \tag{4.10}$$

The canonical action integral $s[q(t), p(t)]$ is replaced by a chain

$$\sum_{K=0}^{N-1} s(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)}) \tag{4.11}$$

of phase space principal functions $s(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)})$. The skeletonized measure $Dq\,Dp$ is taken as the product

$$\prod_{K=0}^{N-1} (2\pi)^{-n}\, d^n q_{(K)}\, d^n p_{(K)} \tag{4.12}$$

of invariant Liouville measures on phase space. There is one such measure at each time $t_{(K)}$, $K = 0,1,...,N-1$, with the exception of the final time $t_{(N)}$. The integration is performed over all of the momenta $p_{(K)}$, $K = 0,1,...,N-1$, but only over the interpolated coordinates $q_{(I)}$, $I = 1,...,N-1$. The differential $d^n q'$ thus remains unused in the integral (4.8) and appears on both sides of the equation. The asymmetric way in which q integrations and p integrations are performed reflects the fact that the paths have fixed boundary configurations but free boundary momenta. The path integral (4.8) is defined as a limit of the described $[Nn\cdot(N-1)n]$-fold integral (q' integration omitted) as $N \to \infty$ while the skeletonization is infinitely refined. That is, if

$$\Delta t_{\mathrm{MAX}} \equiv \max_{K=0,...,N-1} |t_{(K+1)} - t_{(K)}|, \tag{4.13a}$$

then

$$\int Dq\,Dp\, e^{is[q(t),p(t)]} \equiv \lim_{\Delta t_{\mathrm{MAX}} \to 0} \int \prod_{K=0}^{N-1} d^n q_{(K)}\, d^n p_{(K)} \times C(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)}). \tag{4.13b}$$

The biscalar

$$C(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)}) \equiv (2\pi)^{-n} e^{is(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)})} \tag{4.14}$$

we call the classical propagator.

The phase space principal function $s(t'', q''|t', q', p')$ is defined as the canonical action integral (2.1) evaluated along the actual configuration path $q(t)$ between t', q' and t'', q'' given by the equations

$$\nabla_t(g_{ab}\, d_t q^b) = F_a \equiv B_{ab}\, d_t q^b + E_a,$$
$$B_{ab} = \partial_a \phi_b - \partial_b \phi_a, \quad E_a = -\partial_a \phi - \partial_t \phi_a, \tag{4.15}$$

with the momentum p_a propagated from its initial value $p_{a'}$ by the equation of geodesic deviation with a force term,

$$\nabla_t^2(p_a - \phi_a) + R_{abcd}\, d_t q^b (p^c - \phi^c)\, d_t q^d = \nabla_t F_a,$$
$$[\nabla_t(p_a - \phi_a - g_{ab}\, d_t q^b)]_{t=t'} = 0. \tag{4.16}$$

The phase space principal function $s(t'', q''|t', q', p')$ is a biscalar under point transformations (4.9). It is a quadratic function of the initial momenta.

At each step of the skeletonization procedure, the corresponding phase space principal function enters into the classical propagator (4.14). In the limit (4.13), we need to know each function only up to terms linear in the time interval $\Delta t_{(K)} = t_{(K+1)} - t_{(K)}$ and quadratic in the instantaneous geodesic separation $\sigma_{t_{(K)}}(q_{(K+1)}|q_{(K)})$.

To write such an approximate form of the phase space principal function $s(t'', q''|t', q', p')$, we introduce the configuration space Hamilton principal function $s(t'', q''|t', q')$. This function is the extremum of $s(t'', q''|t', q', p')$ with respect to p' and it satisfies the Hamilton–Jacobi equations

$$\partial_{t''} s + h(t'', q^{a''}, p_{a''} = \partial_{a''} s) = 0,$$
$$-\partial_{t'} s + h(t', q^{a'}, p_{a'} = -\partial_{a'} s) = 0. \tag{4.17}$$

From the Hamilton principal function, we can find the initial velocity $d_t q^{a'}$ on the actual path from t', q' to t'', q'':

$$d_t q^{a'} = -g^{a'b'}(\partial_{b'} s + \phi_{b'}). \tag{4.18}$$

This velocity is of the order $\sigma_{t'}/\Delta t$. The approximate form of the phase space principal function can be written in the suggestive form

$$s(t'', q''|t', q', p') \approx (p_{a'}\, d_t q^{a'} - \tfrac{1}{2}\bar{g}^{a'b'}(p_{a'} - \phi_{a'})(p_{b'} - \phi_{b'}) - \phi')\,\Delta t. \tag{4.19}$$

The coefficient

$$\bar{g}^{a'b'}(t'', q''|t', q') = g^{a'b'} - \tfrac{1}{3} R^{a'}{}_{c'}{}^{b'}{}_{d'}(\Delta t\, d_t q^{c'})(\Delta t\, d_t q^{d'}) \tag{4.20}$$

differs from the metric $g^{a'b'}(t', q')$ by a Riemann curvature term which is brought in by the geodesic deviation transport. This term is of the order $\sigma_{t'}^2$. The function (4.19) is constructed in the following way: (I) The initial value of the canonical Lagrangian $p_{a'}\, d_t q^{a'} - h(t', q', p')$ is multiplied by the time interval $\Delta t = t'' - t'$; (II) the initial velocity $d_t q^{a'}$ is expressed

as a function of the boundary data t',q' and t'',q'', Eq. (4.18); (III) the metric in the initial Hamiltonian is replaced by the tensor–scalar coefficient (4.20). We call the modified Hamiltonian $\bar{h}(t'',q''|t',q',p')$.

The description of the phase space integral is now complete. The approximate form (4.19)–(4.20) of the phase space principal function can be used in each classical propagator (4.14) and the path integral defined as the limit (4.13). One can prove[1] that the quantum propagator (4.8) represented by this path integral satisfies the Schrödinger equation (4.6) with the boundary condition (4.7). The geodesic deviation transport which induces the modification (4.20) of the metric ensures that no scalar curvature potential appears in the Schrödinger equation.

5. PATH INTEGRALS IN PHYSICAL CONFIGURATION SPACE

We pass from the phase space path integral (4.13)–(4.14) to a path integral on the physical configuration space by performing momenta integrations. The $K+1$ step in the skeletonization process starts at $t_{(K)},q_{(K)},p_{(K)}$ and ends at $t_{(K+1)},q_{(K+1)}$. Generically, we call

$$t_{(K)}=t,\quad q_{(K)}=q,\quad p_{(K)}=p$$

and (5.1)

$$t_{(K+1)}=\bar{t},\quad q_{(K+1)}=\bar{q}.$$

The phase space principal function (4.19) at each step can be completed into a square,

$$s(\bar{t},\bar{q}|t,q,p)=-\tfrac{1}{2}\bar{g}^{ab}\pi_a\pi_b\,\Delta t+l(t,q,d_tq)\Delta t. \tag{5.2}$$

Here,

$$\Delta t=\bar{t}-t, \tag{5.3}$$

$$\pi_a=p_a-g_{ab}d_tq^b-\phi_a \tag{5.4}$$

and $l(t,q,d_tq)$ is the physical Lagrangian (3.9). The initial velocity d_tq^a is still expressed through the configuration space boundary data $\bar{q},\bar{t},q,t$:

$$d_tq^a=-g^{ab}(t,q)[\partial_b s(\bar{t},\bar{q}|t,q)+\phi_b(t,q)]. \tag{5.5}$$

Let (4.14) be the phase space classical propagator from t,q,p to $\bar{t},\bar{q}$,

$$C(\bar{t},\bar{q}|t,q,p)=(2\pi)^{-n}e^{is(\bar{t},\bar{q}|t,q,p)}. \tag{5.6}$$

We define the configuration space classical propagator as an integral of Eq. (5.6) over the momenta,

$$C(\bar{t},\bar{q}|t,q)\equiv\int d^np\,C(\bar{t},\bar{q}|t,q,p). \tag{5.7}$$

The integration over p can be replaced by integration over π. This leads to the Gaussian integral

$$\int d^n\pi\,e^{-(1/2)i\Delta t\bar{g}^{ab}\pi_a\pi_b}=((2\pi)^{-1}i\Delta t)^{-n/2}\bar{g}^{1/2}, \tag{5.8}$$

where

$$\bar{g}(\bar{t},\bar{q}|t,q)\equiv\det\bar{g}_{ab}. \tag{5.9}$$

Up to the first order terms in Δt,

$$s(\bar{t},\bar{q}|t,q)=l(t,q,d_tq)\,\Delta t. \tag{5.10}$$

This leads to the configuration space classical propagator

$$C(\bar{t},\bar{q}|t,q)=(2\pi i\Delta t)^{-n/2}\bar{g}^{1/2}e^{is(\bar{t},\bar{q}|t,q)}. \tag{5.11}$$

By integrating over all the momenta $p_{(K)}$, $K=0,1,\ldots,N-1$, we transform the quantum propagator (4.8),(4.13) to a configuration space form

$$\langle t'',q''|t',q'\rangle\,d^nq'=\int\bar{D}q\,e^{is[q(t)]}$$
$$\equiv\lim_{\Delta t_{MAX}\to 0}\int\prod_{K=0}^{N-1}d^nq_{(K)}\,C(t_{(K+1)},q_{(K+1)}|t_{(K)},q_{(K)}). \tag{5.12}$$

The integration takes place over the interpolated positions $q_{(I)}$, $I=1,\ldots,N-1$.

The Lagrangian action integral $s[q(t)]$ in Eq. (5.12) gets skeletonized by a chain of Hamilton's principal functions

$$s[q(t)]\approx\sum_{K=0}^{N-1}s(t_{(K+1)},q_{(K+1)}|t_{(K)},q_{(K)})$$
$$\approx\sum_{K=0}^{N-1}l(t_{(K)},q_{(K)},d_tq_{(K)})\Delta t_{(K)}, \tag{5.13}$$

and the measure $\bar{D}q$ is skeletonized by the product

$$\bar{D}q\approx\prod_{K=0}^{N-1}d^nq_{(K)}\,(2\pi i\,\Delta t_{(K)})^{-1/2}$$
$$\times\bar{g}^{1/2}(t_{(K+1)},q_{(K+1)}|t_{(K)},q_{(K)}). \tag{5.14}$$

Each determinant $\bar{g}^{1/2}$ can be expressed as

$$\bar{g}^{1/2}=g^{1/2}(t_{(K)},q_{(K)})(1+R_{ab}(q_{(K)})$$
$$\times\Delta t_{(K)}d_tq^a{}_{(K)}\,\Delta t_{(K)}\,d_tq^b{}_{(K)}). \tag{5.15}$$

Under this measure, the quantum propagator (5.12) satisfies the Schrödinger equation without any curvature term.

6. PATH INTEGRALS IN EXTENDED PHASE SPACE

We shall now express the quantum propagator by path integrals in extended phase space. There are two ways of doing this corresponding classically to whether the constraints are enforced explicitly or implicitly through a lapse multiplier. We begin by replacing each classical propagator $C(\bar{t},\bar{q}|t,q,p)$ by an extended propagator $C(\bar{Q}|Q,P)$ such that, in special coordinates (2.3),(2.6),

$$C(\bar{t},\bar{q}|t,q,p)=\int dQ^0\,dP_0\,C(\bar{Q}|Q,P). \tag{6.1}$$

The procedure then closely follows the parametrization process of classical action. First, we take the absolute time as a prescribed function $t(\tau)$ of a label time $\tau\in[\tau',\tau'']$ respecting the boundary conditions

$$t(\tau')=t',\quad t(\tau'')=t''. \tag{6.2}$$

To first order in $\Delta\tau$,

$$\Delta t=\dot{t}\,\Delta\tau,\quad \Delta t\equiv\bar{t}-t,\quad \Delta\tau\equiv\bar{\tau}-\tau, \tag{6.3}$$

and, as a consequence of Eqs. (4.19) and (4.14),

$$C(\bar{t},\bar{q}|t,q,p)=(2\pi)^{-n}e^{i(p_a\dot{q}^a-\bar{h}(\bar{t},\bar{q}|t,q,p)\dot{t})\Delta\tau}. \tag{6.4}$$

The initial velocity $\dot{q}^a$ in Eq. (6.4) is again expressed as a function of the boundary configuration data [cf. Eq. (5.5)]:

$$\dot{q}^a=\dot{t}\,d_tq^a=-\dot{t}g^{ab}(\partial_bs+\phi_b). \tag{6.5}$$

We adjoin to it the quantity $\dot{t}$ and write

$$\dot{Q}^A = \dot{t}(\tau)\{1, -g^{ab}(\partial_b s + \phi_b)\}. \tag{6.6}$$

In the expression (6.4), the variables q and $\bar{q}$ are arbitrary, but t is considered as a given function of τ, $t(\tau)$. To remove this asymmetry, we consider both t and q as independent variables $Q^A = \{t,q\}$, but multiply the classical propagator (6.4) by a delta function $\delta(Q^0 - t(\tau))$. From now on, s and ϕ_b in Eq. (6.6) are also considered as functions of Q^A, though $\dot{t}(\tau)$ is still a prescribed function of τ.

We also extend the momenta variables by adding a variable p_0, $P_A = \{p_0, p_a\}$, and write the phase factor in Eq. (6.4) as the linear combination $P_A \dot{Q}^A \, \Delta\tau$. To ensure that p_0 is $-\bar{h}$, we multiply the classical propagator by the delta function $\delta(\overline{H}^{(o)})$ of the modified Hamiltonian constraint (2.8),

$$\overline{H}^{(o)} = p_0 + \bar{h}(\bar{t},\bar{q}|t,q,p). \tag{6.7}$$

These changes lead to the following classical propagator on extended phase space:

$$C(\overline{Q}|Q,P) = (2\pi)^{-n}\delta(Q^0 - t(\tau))\delta(\overline{H}^{(o)})e^{iP_A\dot{Q}^A\Delta\tau}. \tag{6.8}$$

Integration of this propagator with respect to the newly introduced variables Q^0 and P_0 reduces it to the old propagator, Eq. (6.1).

The new propagator (6.8) can be written in a manifestly covariant form. We introduce fields $t(Q)$ and $u^A(Q)$ by Eqs. (2.11) and (2.12) and a degenerate tensor–scalar $\bar{g}^{AB}(\overline{Q}|Q)$ related to the coefficient (4.20) by a counterpart of Eq. (2.14). The super-Hamiltonian $\overline{H}^{(o)}$ is thereby cast to the form (2.17) with $\bar{g}^{AB}$ in place of g^{AB}.

In the same vein, Eq. (6.6) assumes the form

$$\dot{Q}^A = \dot{t}(u^A - g^{AB}(\partial_B S + \phi_B)). \tag{6.9}$$

The Hamilton–Jacobi equations which determine the Hamilton principal function

$$S(\overline{Q}|Q) = s(\bar{t},\bar{q}|t,q) \tag{6.10}$$

are obtained by substituting $P_A = -\partial_A S$ and $P_{\bar{A}} = \partial_{\bar{A}} S$ into the Hamiltonian constraint at the initial and the final boundaries,

$$\begin{aligned} &-u^A(\partial_A S + \phi_A) + \tfrac{1}{2}g^{AB}(\partial_A S + \phi_A)(\partial_B S + \phi_B) = 0, \\ &u^{\bar{A}}(\partial_{\bar{A}} S - \phi_{\bar{A}}) + \tfrac{1}{2}g^{\bar{A}\bar{B}}(\partial_{\bar{A}} S - \phi_{\bar{A}})(\partial_{\bar{B}} S - \phi_{\bar{B}}) = 0. \end{aligned} \tag{6.11}$$

The classical propagator (6.8) then takes on a manifestly covariant appearance

$$C(\overline{Q}|Q,P) = (2\pi)^{-n}\delta(t(Q) - t(\tau))\delta(\overline{H}^{(o)})e^{iP_A\dot{Q}^A\Delta\tau}. \tag{6.12}$$

We can now mix the extended phase space variables Q^A, P_A by an arbitrary point transformation (2.20) and transform the classical propagator as a biscalar without changing its general form (6.12).

In a final step, we scale $\overline{H}^{(o)}$ into $\overline{H}$ by a positive scalar factor $\Lambda(Q)$ as in Eqs. (2.18)–(2.19). In terms of the scaled quantities (2.18), S again satisfies the Hamilton–Jacobi equations (6.11), but the scaling factor enters into Eq. (6.9) by which $\dot{Q}^A$ is interpreted in terms of the boundary configurations,

$$\dot{Q}^A = \Lambda^{-1}\dot{t}\,[U^A - G^{AB}(\partial_B S + \phi_B)]. \tag{6.13}$$

Because $\delta(\overline{H}^{(o)}) = \delta(\Lambda^{-1}\overline{H}) = \Lambda\,\delta(\overline{H})$, the scaling factor also explicitly appears in the modulus of the classical propagator (6.12), which becomes

$$C(\overline{Q}|Q,P) = (2\pi)^{-n}\Lambda(Q)\delta(t(Q) - t(\tau))\delta(\overline{H})e^{iP_A\dot{Q}^A\Delta\tau}. \tag{6.14}$$

The absolute time function $t(Q)$ is covariantly characterized by Eqs. (2.21) and (2.27). The scaling factor Λ in expressions (6.13) and (6.14) can then be interpreted by Eq. (2.28) or, alternatively, as the Poisson bracket

$$\Lambda(Q) = [t(Q),H] = [t(Q),\overline{H}]. \tag{6.15}$$

This completes a covariant characterization of the classical propagator (6.14).

The quantum propagator can be represented by a path integral in the extended phase space,

$$\begin{aligned} \langle Q''|Q'\rangle\delta(t(Q') - t')\,d^{n+1}Q' &= \int \overline{D}Q\,\overline{D}P\,e^{iS[Q,P]} \\ &= \lim_{\Delta\tau_{MAX}\to 0}\int\prod_{K=0}^{N-1} d^{n+1}Q_{(K)}\,d^{n+1}P_{(K)} \\ &\quad\times C(Q_{(K+1)}|Q_{(K)},P_{(K)}). \end{aligned} \tag{6.16}$$

The integrations are performed over all the extended momenta $P_{(K)}$, $K = 0,1,\ldots,N-1$, but only over the interpolated extended coordinates $Q_{(I)}$, $I = 1,\ldots,N-1$. Due to Eq. (6.1), we obtain in this way our old quantum propagator (4.8),(4.13).

The new form (6.16) of the path integral corresponds to the conditional form of the action, Table II, line 3. The skeletonized measure

$$\begin{aligned} \overline{D}Q\,\overline{D}P \approx \prod_{K=0}^{N-1} &d^{n+1}Q_{(K)}\,d^{n+1}P_{(K)} \\ &\times(2\pi)^{-n}\Lambda(Q_{(K)})\delta(t(Q_{(K)}) - t(\tau_{(K)})) \\ &\times\delta(\overline{H}(Q_{(K+1)}|Q_{(K)},P_{(K)})) \end{aligned} \tag{6.17}$$

contains a product of delta functions $\delta(\overline{H}(Q_{(K+1)}|Q_{(K)}, P_{(K)}))$ which enforce the Hamiltonian constraint at each instant $\tau_{(K)}$ of the skeletonized time. However, these constraints are not simply classical Hamiltonian constraints at $\tau_{(K)}$, but modified constraints in which the metric $G^{AB}(Q_{(K)})$ is replaced by the tensor–scalar coefficient $\overline{G}^{AB}(Q_{(K+1)}|Q_{(K)})$. This modification is necessary for the quantum propagator (6.16) to satisfy the Schrödinger equation without an additional scalar curvature potential. If the measure contained the unmodified super-Hamiltonian $H(Q_{(K)},P_{(K)})$, the Schrödinger equation would acquire the potential $\frac{1}{12}R$.

Besides the delta functions of super-Hamiltonians, the measure also contains the delta functions $\delta(t(Q_{(K)}) - t(\tau_{(K)}))$. These delta functions ensure that the instants of the label time τ correspond to the leaves of absolute time t. The configurations which the system has to select at an instant τ are thus all simultaneous in the absolute sense. The labeling of the leaves of absolute time, however, is provided by an arbitrary parameter τ. Finally, the factor $\Lambda(Q)$ in the measure takes care of an arbitrary scaling of the Hamiltonian constraint.

In addition to the choice of measure, one must also specify how to skeletonize the action functional

$S[Q,P] = \int_\tau^{\tau''} d\tau\, P_A \dot{Q}^A$. Our skeletonization says that $S[Q,P]$ is to be replaced by the sum

$$S[Q,P] \approx \sum_{K=0}^{N-1} P_{(K)A} \dot{Q}_{(K)}{}^A \, \Delta\tau_{(K)}, \tag{6.18}$$

in which $\dot{Q}_{(K)}{}^A$ is the actual extended velocity at $\tau_{(K)}$ on the actual path from $Q_{(K)}$ to $Q_{(K+1)}$. This actual velocity can be derived from the Hamilton principal function $S(Q_{(K+1)}|Q_{(K)})$ by Eq. (6.13).

It is easy to introduce the lapse multiplier and pass from the conditional form of the path integral to an unconditional one. We just interpret each $\delta(\overline{H})$ as the Fourier integral

$$\delta(\overline{H}) = \int dN\, \Delta\tau\, (2\pi)^{-1} e^{-iN\overline{H}\,\Delta\tau}. \tag{6.19}$$

In other words, we extend the classical propagator $C(\overline{Q}\,|Q,P)$ into the Q, P, N space by the prescription

$$C(\overline{Q}\,|Q,P,N) = (2\pi)^{-(n+1)}\, \Delta\tau\, \Lambda(Q)\delta(t(Q) - t(\tau)) \times e^{i(P_A \dot{Q}^A - NH)\,\Delta\tau} \tag{6.20}$$

and connect it with the old propagator by the equation

$$C(\overline{Q}\,|Q,P) = \int DN\, C(\overline{Q}\,|Q,P,N). \tag{6.21}$$

The quantum propagator (6.16) can then be represented by a path integral in the Q, P, N space,

$$\begin{aligned}\langle Q''|Q'\rangle \delta(t(Q') - t')\, d^{n+1}Q' &= \int \overline{D}Q\, \overline{D}P\, \overline{D}N\, e^{iS[Q,P,N]} \\ &\equiv \lim_{\Delta\tau_{MAX}\to 0} \int \prod_{K=0}^{N-1} d^{n+1}Q_{(K)}\, d^{n+1}P_{(K)}\, dN_{(K)} \\ &\quad \times C(Q_{(K+1)}|Q_{(K)},P_{(K)},N_{(K)}).\end{aligned} \tag{6.22}$$

The integration takes place over all $N_{(K)}$, $K = 0,1,\ldots,N-1$. This corresponds to the fact that the lapse function is a Lagrange multiplier which, like the momenta P_A, can be freely specified at the ends.

The skeletonized measure has the form

$$\begin{aligned}\overline{D}Q\, \overline{D}P\, \overline{D}N \approx \prod_{K=0}^{N-1} & d^{n+1}Q_{(K)}\, d^{n+1}P_{(K)}\, dN_{(K)} \\ & \times \Delta\tau_{(K)}\, \Lambda(Q_{(K)})(2\pi)^{-(n+1)} \\ & \times \delta(t(Q_{(K)}) - t(\tau_{(K)})).\end{aligned} \tag{6.23}$$

The product $N_{(K)}\, \Delta\tau_{(K)}\, \Lambda(Q_{(K)})$ which enters into the measure is unchanged when we use a different label time; in fact, $N\,\Delta\tau\,\Lambda$ is to be interpreted as the interval Δt of the absolute time, Eq. (3.5).

Finally, the action functional (3.2) is replaced by the sum

$$\begin{aligned}S[Q,P,N] \approx \sum_{K=0}^{N-1} & (P_{(K)A} \dot{Q}_{(K)}{}^A \\ & - N_{(K)} \overline{H}(Q_{(K+1)}|Q_{(K)},P_{(K)}))\, \Delta\tau_{(K)}.\end{aligned} \tag{6.24}$$

Here, $\dot{Q}_{(K)}{}^A$ is again given by Eq. (6.13) and $\overline{H}$ is the modified super-Hamiltonian.

We have thereby transformed the path integral in physical phase space into two equivalent forms in the extended phase space, one with and one without the lapse multiplier.

7. PATH INTEGRALS IN EXTENDED CONFIGURATION SPACE

The path integral in physical configuration space was obtained from the path integral in physical phase space by evaluating all integrals over the momenta. Similarly, by integrating the extended classical propagator (6.14) over the extended momentum variables we cast the path integral into a form corresponding to the homogeneous Lagrangian on the extended configuration space. To do this, we introduce for convenience mechanical energy and momenta

$$\Pi_A = P_A - \phi_A \tag{7.1}$$

as new variables. The extended classical propagator (6.14) assumes the form

$$\begin{aligned}C(\overline{Q}\,|Q,\Pi) = (2\pi)^{-n} \Lambda(Q)\delta(t(Q) - t(\tau))\delta(\overline{H}) e^{i\Pi_A \dot{Q}^A\,\Delta\tau} \\ \times e^{i\phi_A \dot{Q}^A\,\Delta\tau},\end{aligned} \tag{7.2}$$

with

$$\overline{H} = U^A \Pi_A + \tfrac{1}{2}\overline{G}^{AB}\Pi_A \Pi_B. \tag{7.3}$$

Let Q^a_A be n linearly independent covectors perpendicular to U^A,

$$U^A Q^a_A = 0, \quad a = 1,\ldots,n. \tag{7.4}$$

The projected coefficient

$$\overline{G}^{ab} = \overline{G}^{AB} Q^a_A Q^b_B \tag{7.5}$$

is nondegenerate. The covectors $\{T_A, Q^a_A\}$ form a basis in the cotangent space. We split Π_A into a longitudinal and transversal parts according to

$$\Pi_A = \Pi_\parallel T_A + \Pi_a Q^a_A. \tag{7.6}$$

The Jacobian $J = \det \partial\{\Pi_A\}/\partial\{\Pi_\parallel, \Pi_a\}$ of the transformation (7.6) from the variables $\{\Pi_\parallel, \Pi_a\}$ to the variables Π_A is (see Appendix B)

$$J = (1/n!)\delta^{AA_1\cdots A_n} T_A Q^{a_1}_{A_1} \cdots Q^{a_n}_{A_n} \delta_{a_1\cdots a_n}, \tag{7.7}$$

where $\delta_{a_1\cdots a_n}$ is the alternating symbol. As a consequence,

$$C(\overline{Q}\,|Q,P)\, d^{n+1}P = J(Q) C(\overline{Q}\,|Q,\Pi_\parallel,\Pi_a) d\Pi_\parallel\, d^n\Pi. \tag{7.8}$$

In the new variables,

$$\overline{H} = \Pi_\parallel + \tfrac{1}{2}\overline{G}^{ab}\Pi_a \Pi_b \tag{7.9}$$

and the integration with respect to $\Pi_\parallel$ is easily performed. We get

$$\begin{aligned}C(\overline{Q}\,|Q,\Pi_a) &= \int d\Pi_\parallel\, J(Q) C(\overline{Q}\,|Q,\Pi_\parallel,\Pi_a) \\ &= (2\pi)^{-n} J(Q) \Lambda(Q) \delta(t(Q) - t(\tau)) \\ &\quad \times e^{i\phi_A \dot{Q}^A\,\Delta\tau} e^{i[\Pi_a \dot{Q}^a - (1/2)(T_C \dot{Q}^C)\overline{G}^{ab}\Pi_a \Pi_b]\,\Delta\tau},\end{aligned} \tag{7.10}$$

where we have introduced the abbreviation

$$\dot{Q}^a \equiv Q^a_A \dot{Q}^A. \tag{7.11}$$

The terms in Π_a can be completed into a square,

$$\begin{aligned}(\Pi_a \dot{Q}^a - (T_C \dot{Q}^C)\tfrac{1}{2}\overline{G}^{ab}\Pi_a \Pi_b)\,\Delta\tau \\ = (-\tfrac{1}{2}(T_C \dot{Q}^C)\overline{G}^{ab}\widetilde{\Pi}_a \widetilde{\Pi}_b + \tfrac{1}{2}(T_C \dot{Q}^C)^{-1}\overline{G}_{ab}\dot{Q}^a \dot{Q}^b)\,\Delta\tau,\end{aligned} \tag{7.12}$$

where

$$\tilde{\Pi}_a = \Pi_a - (T_C \dot{Q}^C)^{-1} \overline{G}_{ab} \dot{Q}^b. \tag{7.13}$$

Moreover,

$$\overline{G}_{ab} \dot{Q}^a \dot{Q}^b = \overline{G}_{AB} \dot{Q}^A \dot{Q}^B = G_{AB} \dot{Q}^A \dot{Q}^B. \tag{7.14}$$

One can replace the modified coefficient $\overline{G}_{AB}$ by the metric G_{AB} because, in special coordinates $Q^A = \{t, q_a\}$, $(R_{acbd}\, \Delta\tau \dot{q}^c \Delta\tau \dot{q}^d) \dot{q}^a \dot{q}^b = 0$. As a result,

$$C(\overline{Q}\,|Q,\tilde{\Pi}_a) = (2\pi)^{-n} J(Q) \Lambda(Q) \delta(t(Q) - t(\tau)) \times e^{-(1/2)i(T_C \dot{Q}^C)\, \Delta\tau\, \overline{G}^{ab} \tilde{\Pi}_a \tilde{\Pi}_b} e^{iL(Q,\dot{Q})\, \Delta\tau}, \tag{7.15}$$

where $L(Q,\dot{Q})$ is the homogeneous Lagrangian (3.19). The Gaussian integral over $\tilde{\Pi}_a$ gives

$$(2\pi)^{-n} \int d^n \tilde{\Pi}\, e^{-(1/2)i(T_C \dot{Q}^C)^{-1}\, \Delta\tau\, \overline{G}^{ab} \tilde{\Pi}_a \tilde{\Pi}_b} = (2\pi i T_C \dot{Q}^C \Delta\tau)^{-n/2} \overline{G}^{1/2} \tag{7.16}$$

with

$$\overline{G} \equiv \det \overline{G}_{ab}. \tag{7.17}$$

The product

$$J\overline{G}^{1/2} \equiv \overline{D}^{-1/2} \tag{7.18}$$

can be written directly in terms of the degenerate coefficient $\overline{G}^{AB}$ (Appendix B):

$$\overline{D} = (1/n!) \delta_{AA_1 \cdots A_n} U^A U^B \overline{G}^{A_1 B_1} \cdots \overline{G}^{A_n B_n} \delta_{BB_1 \cdots B_n}. \tag{7.19}$$

This sequence of steps yields the classical propagator in extended configuration space,

$$\begin{aligned} C(\overline{Q}\,|Q) &= \int d^{n+1} P\, C(\overline{Q}\,|Q,P) \\ &= \int d^n \tilde{\Pi}\, C(\overline{Q}\,|Q,\tilde{\Pi}_a) \\ &= (2\pi i T_C \dot{Q}^C \Delta\tau)^{-n/2} \\ &\quad \times \overline{D}^{-1/2} (\overline{Q}\,|Q) \Lambda(Q) \delta(t(Q) - t(\tau)) e^{iL(Q,\dot{Q})\, \Delta\tau}. \end{aligned} \tag{7.20}$$

Note that by the interpretation (6.13) of $\dot{Q}^C$ we have

$$T_C \dot{Q}^C = \Lambda^{-1}(Q) \dot{t}(\tau). \tag{7.21}$$

From Eq. (6.16), we obtain a representation of quantum propagator by a path integral in the extended configuration space,

$$\langle Q'' | Q' \rangle \delta(t(Q') - t')\, d^{n+1} Q' = \int \overline{D}Q\, e^{iS[Q]} \equiv \lim_{\Delta\tau_{\mathrm{MAX}} \to 0} \int \prod_{K=0}^{N-1} d^{n+1} Q_{(K)}\, C(Q_{(K+1)}\,|Q_{(K)}). \tag{7.22}$$

The integration takes place only over the interpolated extended coordinates $Q_{(I)}$, $I = 1,\ldots,N-1$. The homogeneous Lagrangian action $S[Q] = \int_{\tau'}^{\tau''} d\tau\, L(Q,\dot{Q})$ is skeletonized by the prescription

$$S[Q] \approx \sum_{K=0}^{N-1} (\tfrac{1}{2}(T_C(Q_{(K)}) \dot{Q}^C_{(K)})^{-1} \times G_{AB}(Q_{(K)}) \dot{Q}^A_{(K)} \dot{Q}^B_{(K)} + \phi_A(Q_{(K)}) \dot{Q}^A_{(K)}\, \Delta\tau_{(K)}). \tag{7.23}$$

The velocities $\dot{Q}_{(K)}$ are interpreted in terms of the configuration data at the ends of each step in Eq. (6.13). Note that the coefficient G_{AB} in Eq. (7.23) is the ordinary degenerate metric unmodified by the curvature term. The modified metric coefficient enters only into the measure, but not into the phase of the path integral (7.22). The measure is skeletonized by the product

$$\overline{D}Q \approx \prod_{K=0}^{N-1} d^{n+1} Q\, (2\pi i T_C(Q_{(K)}) \dot{Q}^C_{(K)}\, \Delta\tau_{(K)})^{-n/2} \times D^{-1/2}(Q_{(K+1)}\,|Q_{(K)}) \Lambda(Q_{(K)}) \times \delta(t(Q_{(K)}) - t(\tau_{(K)})). \tag{7.24}$$

The modified metric coefficient appears in the determinant (7.19).

In the special coordinates (2.3) all previous expressions considerably simplify. The Jacobian (7.7) reduces to

$$J = \Lambda^{-1}, \tag{7.25}$$

the determinant (7.19) goes over to

$$\overline{D} = \Lambda^{n+2} \overline{g}^{-1}, \quad \overline{g} \equiv \det \overline{g}_{ab}, \tag{7.26}$$

and the classical propagator assumes the form

$$C(\bar{t},\bar{q}\,|t,q) = (2\pi i \dot{t}(\tau)\, \Delta\tau)^{-n/2} \times \overline{g}^{1/2}(\bar{t},\bar{q}\,|t,q) \delta(t - t(\tau)) e^{iL(t,q,\dot{t},\dot{q})\, \Delta\tau}. \tag{7.27}$$

Here, $L(t,q,\dot{t},\dot{q})$ is the homogeneous Lagrangian (3.20). In the expression (7.24), t is an independent variable, while $t(\tau)$ and $\dot{t}(\tau)$ are prescribed functions of τ. The velocity $\dot{q}^a$ is interpreted as a function of t,q, and $\bar{t},\bar{q}$ by Eq. (6.5). The measure (7.24) in path integral (7.22) reduces to

$$\overline{D}t\, \overline{D}q \approx \prod_{K=0}^{N-1} dt_{(K)}\, d^n q_{(K)}\, [2\pi i t(\tau_{(K)})\, \Delta\tau_{(K)}]^{-n/2} \times \overline{g}^{1/2}(t_{(K+1)}, q_{(K+1)}\,|t_{(K)}, q_{(K)}) \delta(t_{(K)} - t(\tau_{(K)})), \tag{7.28}$$

while the Lagrangian action $S[t,q]$ gets skeletonized by

$$S[t,q] \approx \sum_{K=0}^{N-1} [\tfrac{1}{2} \dot{t}^{-1}(\tau_{(K)}) g_{ab}(t_{(K)}, q_{(K)}) \dot{q}^a_{(K)} \dot{q}^b_{(K)} + \phi_a(t_{(K)}, q_{(K)}) \dot{q}^a_{(K)} - \phi(t_{(K)}, q_{(K)}) \dot{t}(\tau_{(K)})]\, \Delta\tau_{(K)}. \tag{7.29}$$

When we perform the integrations over $t_{(I)}$, $I = 1,\ldots,N-1$, and parametrize the paths by absolute time, $t(\tau) = \tau$, the path integral (7.22) reduces back to the path integral (5.12) in physical configuration space.

8. PATH INTEGRALS IN EXTENDED CONFIGURATION SPACE WITH LAPSE

The only form of the action remaining in Table I is the Lagrangian action on extended configuration space with the lapse multiplier, Eqs. (3.15)–(3.16). We now represent the quantum propagator by a path integral whose phase is this action.

We start from the classical propagator (7.10) in which the integration over longitudinal part of the momentum $\Pi_{\|}$ has been performed, but which still depends on the transversal momenta Π_a. Instead of integrating over all transversal momenta Π_a [which would lead us back to the classical propagator (7.20)], we decompose Π_a into a component parallel to the velocity $\dot{Q}^a$ and $n-1$ components perpendicular to $\dot{Q}^a$. We choose a basis Q^α_a, $\alpha = 1,\ldots,N-1$, in the subspace perpendicular to $\dot{Q}^a$,

$$\dot{Q}^a Q^\alpha_a = 0, \tag{8.1}$$

and write

$$\Pi_a = N^{-1}\dot{Q}_a + \Pi_\alpha Q^\alpha_a. \tag{8.2}$$

The Jacobian of this transformation is (Appendix B)

$$\det \frac{\partial\{\Pi_a\}}{\partial\{N,\Pi_\alpha\}} = \det \begin{vmatrix} -N^{-2}\dot{Q}_a \\ Q^\alpha_a \end{vmatrix} = -N^{-2}\tilde{J}, \tag{8.3}$$

with

$$\tilde{J} = (1/(n-1)!)\delta^{aa_1\cdots a_{n-1}}\dot{Q}_a Q^{\alpha_1}_{a_1}\cdots Q^{\alpha_{n-1}}_{a_{n-1}}\delta_{\alpha_1\cdots\alpha_{n-1}}. \tag{8.4}$$

The last equation is the counterpart of Eq. (7.7) in a space of lower dimension.

Expressing the phase of the propagator (7.10) in terms of our new variables, we find

$$\begin{aligned}
&[\Pi_a\dot{Q}^a - \tfrac{1}{2}(T_C\dot{Q}^C)\overline{G}^{ab}\Pi_a\Pi_b + \phi_A\dot{Q}^A]\,\Delta\tau \\
&\quad = -\tfrac{1}{2}(T_C\dot{Q}^C)\overline{G}^{\alpha\beta}\Pi_\alpha\Pi_\beta\,\Delta\tau \\
&\qquad + [(N^{-1} - \tfrac{1}{2}T_C\dot{Q}^C N^{-2})G_{ab}\dot{Q}^a\dot{Q}^b + \phi_A\dot{Q}^A]\,\Delta\tau \\
&\quad = -\tfrac{1}{2}(T_C\dot{Q}^C)\overline{G}^{\alpha\beta}\Pi_a\Pi_\beta\,\Delta\tau + L(Q,\dot{Q},N)\,\Delta\tau.
\end{aligned} \tag{8.5}$$

The metric coefficient $\overline{G}^{\alpha\beta}$ is the projection

$$\overline{G}^{\alpha\beta} = \overline{G}^{ab}Q^\alpha_a Q^\beta_b \tag{8.6}$$

and $L(Q,\dot{Q},N)$ is the Lagrangian (3.15) with the lapse function N. The propagator (7.10) thereby assumes the form

$$\begin{aligned}
&C(\overline{Q}\,|Q,N,\Pi_\alpha) \\
&\quad = -N^{-2}\tilde{J}(Q)C(\overline{Q}\,|Q,\Pi_a) \\
&\quad = (2\pi)^{-n}(-N^{-2})\tilde{J}(Q)J(Q)\Lambda(Q)\delta(t(Q)-t(\tau)) \\
&\qquad \times e^{iL(Q,\dot{Q},N)\,\Delta\tau}e^{-(1/2)i(T_C\dot{Q}^C)G^{\alpha\beta}\Pi_\alpha\Pi_\beta\Delta\tau}.
\end{aligned} \tag{8.7}$$

We now evaluate the Gaussian integral over the momenta Π_α and find

$$\begin{aligned}
&\int d^{n-1}\Pi\, e^{-(1/2)i(T_C\dot{Q}^C)\,\Delta\tau\, G^{\alpha\beta}\Pi_\alpha\Pi_\beta} \\
&\quad = (2\pi)^{(n-1)/2}(iT_C\dot{Q}^C\Delta\tau)^{-(n-1)/2}\det{}^{1/2}\overline{G}_{\alpha\beta}.
\end{aligned} \tag{8.8}$$

Taking into account Eqs. (B24) and (7.14),

$$\tilde{J}J\det{}^{1/2}\overline{G}_{\alpha\beta} = \overline{D}^{-1/2}(G_{AB}\dot{Q}^A\dot{Q}^B)^{1/2}. \tag{8.9}$$

This sequence of operations leads us to the classical propagator

$$\begin{aligned}
&C(\overline{Q}\,|Q,N) = \int d^{n-1}\Pi\, C(\overline{Q}\,|Q,N,\Pi_\alpha) \\
&(-N^{-2})(2\pi)^{-1}(2\pi iT_C\dot{Q}^C\Delta\tau)^{-(n-1)/2}\overline{D}^{-1/2}(\overline{Q}\,|Q) \\
&\quad \times(G_{AB}\dot{Q}^A\dot{Q}^B)^{1/2}\Lambda(Q)\delta(t(Q)-t(\tau))e^{iL(Q,\dot{Q},N)\,\Delta\tau}.
\end{aligned} \tag{8.10}$$

All velocities $\dot{Q}^A$ in Eq. (8.10) are expressed in terms of boundary data, Eq. (6.13).

We can now represent the quantum propagator by the path integral

$$\begin{aligned}
&\langle Q''|Q'\rangle\delta(t(Q')-t')\,d^{n+1}Q' = \int \overline{D}Q\,\overline{D}N\,e^{iS[Q,N]} \\
&\quad \equiv \lim_{\Delta\tau_{\mathrm{MAX}}\to 0}\int\prod_{K=0}^{N-1} d^{n+1}Q_{(K)}\,dN_{(K)} \\
&\qquad \times C(Q_{(K+1)}|Q_{(K)},N_{(K)}).
\end{aligned} \tag{8.11}$$

The integral in Eq. (8.11) is over all $N_{(K)}$, $K = 0,1,\ldots,N-1$, but only over the interpolated $Q_{(I)}$, $I = 1,\ldots,N$. This corresponds to the fact that the momentumlike multiplier N has free ends.

The Lagrangian action $S[Q,N]$ is skeletonized by the prescription

$$\begin{aligned}
S[Q,N] \approx \sum_{K=0}^{N-1}\{&(N^{-1}_{(K)} - \tfrac{1}{2}T_C(Q_{(K)})\dot{Q}^C_{(K)}N^{-2}_{(K)}) \\
&\times G_{AB}(Q_{(K)})\dot{Q}^A_{(K)}\dot{Q}^B_{(K)} + \phi_A(Q_{(K)})\dot{Q}^A_{(K)}\}\,\Delta\tau,
\end{aligned} \tag{8.12}$$

where $\dot{Q}_{(K)}$ are again interpreted in terms of the configuration data $Q_{(K)}, Q_{(K+1)}$ at the boundaries of each step by Eq. (6.13).

The measure is skeletonized by the product

$$\begin{aligned}
\overline{D}Q\,\overline{D}N \approx \prod_{K=0}^{N-1} &d^{n+1}Q_{(K)}\,dN_{(K)}\,(-N^{-2}_{(K)}) \\
&\times(2\pi)^{-1}[2\pi iT_C(Q_{(K)})\dot{Q}^C_{(K)}\,\Delta\tau_{(K)}]^{-(n-1)/2} \\
&\times\overline{D}^{-1/2}(Q_{(K+1)}|Q_{(K)})(G_{AB}(Q_{(K)})\dot{Q}^A_{(K)}\dot{Q}^B_{(K)})^{1/2} \\
&\times\Lambda(Q_{(K)})\delta(t(Q_{(K)}) - t(\tau_{(K)})).
\end{aligned} \tag{8.13}$$

The modified metric coefficient enters into the measure (8.13) through the determinant (7.19).

These expressions simplify considerably in the special coordinate system, but, before showing this, let us recover the path integral in the extended configuration space by performing the integrations over $N_{(K)}$. To do this, we write the phase of the classical propagator (8.10) in the form

$$\begin{aligned}
&L(Q,\dot{Q},N)\,\Delta\tau = L(Q,\dot{Q})\,\Delta\tau \\
&\quad - \tfrac{1}{2}(N^{-1}(T_C\dot{Q}^C)^{1/2} - (T_C\dot{Q}^C)^{-1/2})^2 G_{AB}\dot{Q}^A\dot{Q}^B\Delta\tau,
\end{aligned} \tag{8.14}$$

where $L(Q,\dot{Q})$ is the homogeneous Lagrangian (3.19). We replace N by a new variable

$$M = N^{-1}(T_C\dot{Q}^C)^{1/2} - (T_C\dot{Q}^C)^{-1/2} \tag{8.15}$$

and write

$$\begin{aligned}
C(\overline{Q}\,|Q,N)\,dN &= C(\overline{Q}\,|Q,N)(-N^2(T_C\dot{Q}^C)^{-1/2})\,dM \\
&= dM\,e^{-(1/2)iG_{AB}\dot{Q}^A\dot{Q}^B\Delta\tau M^2} \\
&\quad\times(2\pi)^{-1}(T_C\dot{Q}^C)^{-1/2}(G_{AB}\dot{Q}^A\dot{Q}^B)^{1/2} \\
&\quad\times(2\pi iT_C\dot{Q}^C\Delta\tau)^{-(n-1)/2}\overline{D}^{-1/2}\Lambda \\
&\quad\times\delta(t(Q)-t(\tau))e^{iL(Q,\dot{Q})\,\Delta\tau}.
\end{aligned} \tag{8.16}$$

Integration over M yields the Gaussian integral

$$\int dM\,e^{-(1/2)iG_{AB}\dot{Q}^A\dot{Q}^B\Delta\tau M^2} = \pi^2(\tfrac{1}{2}i\,\Delta\tau)^{-1/2}(G_{AB}\dot{Q}^A\dot{Q}^B)^{-1/2}, \tag{8.17}$$

and the classical propagator (8.16) reduces back to the classical propagator (7.20).

The classical propagator (8.10) again simplifies in special coordinates (2.3). Taking into account Eqs. (7.21), (7.26) and rescaling the lapse multiplier,

$$N^{(o)} = AN, \tag{8.18}$$

we get

$$C(\bar{t},\bar{q}|t,q,N^{(o)})\,dN^{(o)}$$
$$= -dN^{(o)}N^{(o)-2}(2\pi)^{-1}(2\pi i\dot{t}(\tau)\Delta\tau)^{-(n-1)/2}$$
$$\times\bar{g}^{1/2}(\bar{t},\bar{q}|t,q)[g_{ab}(q)\dot{q}^a\dot{q}^b]^2\delta(t-t(\tau))e^{iL(t,q,\dot{q},N^{(o)})\Delta\tau}, \tag{8.19}$$

where $L(t,q,\dot{q},N^{(o)})$ is the action (3.17).

The velocity $\dot{q}^a$ is again interpreted by Eq. (6.5). The measure (8.13) in the path integral (8.11) reduces thereby to

$$\bar{D}Q\,\bar{D}N^{(o)} \approx \prod_{K=0}^{N-1} dt_{(K)}\,d^nq_{(K)}\,dN^{(o)}_{(K)}$$
$$-N^{(o)-2}_{(K)}(2\pi)^{-1}(2\pi i\dot{t}(\tau_{(K)})\Delta\tau_{(K)})^{-(n-1)/2}$$
$$\times\bar{g}^{1/2}(t_{(K+1)},q_{(K+1)}|t_{(K)},q_{(K)})$$
$$\times(g_{ab}(q_{(K)})\dot{q}^a_{(K)}\dot{q}^b_{(K)})^{1/2}\delta(t_{(K)}-t(\tau_{(K)})). \tag{8.20}$$

This completes our program. We have represented the quantum propagator by path integrals corresponding to all action functionals enumerated in Table II.

9. PATH INTEGRALS: PARAMETRIZATION VERSUS GAUGE

We have now learned how to write the quantum propagator for a parametrized system as a path integral in extended phase space. Our prescription, Eqs. (6.16)–(6.17), recognizes the need to enforce the Hamiltonian constraint and to select a definite parametrization of the path. These two aims are achieved by the delta functions $\delta(\bar{H}(Q_{(K+1)}|Q_{(K)},P_{(K)}))$ and $\delta(t(Q_{(K)})-t(\tau_{(K)}))$ in the skeletonized measure.

A similar need arises in gauge theories. One must enforce the constraints generating gauge transformations, and one should fix the gauge when writing the path integral in the space of redundant variables. It is of interest to compare the algorithm which we have obtained for a parametrized theory with the standard prescription for gauge theories.

Let us first review the basic structure of gauge theories. To bring out the issues clearly, we consider again our old finite-dimensional nonrelativistic system. We can turn it into a gauge theory by adjoining an additional spurious gauge coordinate ϕ to the physical coordinates q^a. This brings us to the extended configuration space $\{q^a,\phi\}$. As q^a is kept fixed and ϕ is varied, we move along a fiber over q^a. We interpret all points in such a fiber as different descriptions of the same physical state.

As the state of the system evolves in time, the choice of the gauge variable remains arbitrary. In other words, the velocity

$$d_t\phi(t)=\lambda(t) \tag{9.1}$$

can be freely prescribed at each step of the dynamical evolution. Equation (9.1) can be obtained by varying the action

$$\sigma[\phi,\pi;\lambda] = \int_{t'}^{t''} dt\,(\pi\,d_t\phi-\lambda\pi) \tag{9.2}$$

with respect to the gauge momentum π. By varying (9.2) with respect to λ and ϕ, we learn that the momentum π is constrained to vanish,

$$\pi=0, \tag{9.3}$$

and continues to vanish in the course of time.

The evolution of the system in the extended phase space $\{q^a,\phi,p_a,\pi\}$ is then described by the action S which is the sum of the physical action (2.1) and the gauge action (9.2)

$$S[q^a,\phi,p_a,\pi;\lambda] = \int_{t'}^{t''} dt\,(p_a d_t q^a+\pi\,d_t\phi-h(t,q,p)-\lambda\pi). \tag{9.4}$$

After an arbitrary point transformation in the extended phase space,

$$Q^A=Q^A(q^a,\phi),\quad p_a=Q^A{}_{,a}P_A,\quad \pi=Q^A{}_{,\phi}P_A, \tag{9.5}$$

the action (9.5) assumes the form

$$S[Q^A,P_A;\lambda] = \int_{t'}^{t''} dt\,(P_A d_t Q^A - h(Q,P)-\lambda\pi(Q,P)). \tag{9.6}$$

The action (9.6) can be modified in two ways without changing the equations of motion. The constraint (9.3) can be scaled by an arbitrary factor $\Lambda(Q)\neq 0$,

$$\Pi=\Lambda(Q)\pi(Q,P), \tag{9.7}$$

and it can be adjoined to the physical Hamiltonian h,

$$\tilde{h}=h+[k^A(Q)P_A+k(Q)]\pi(Q,P). \tag{9.8}$$

We have chosen the coefficients $\Lambda(Q)$ and $k^A(Q)P_A+k(Q)$ so that the new constraint Π is still linear in the momenta P_A and the new Hamiltonian $\tilde{h}$ is still quadratic in the momenta P_A.

The constraints (9.3) or (9.7) generate the gauge transformation of the canonical variables Q^A,P_A. Such a transformation does not change the physical state of the system. To single out a particular representative for each physical state, one can introduce a gauge fixing condition

$$\Phi(Q^A,P_A)=0. \tag{9.9}$$

Here, Φ is any function which yields a unique value of the gauge coordinate ϕ when Eqs. (9.3) and (9.5) are taken into account.

We can write now the standard prescription for the quantum propagator as a path integral in the extended phase space $\{Q^A,P_A\}$ of the gauge theory. The propagator has the form (4.13) with the classical propagator

$$C(\bar{Q}|Q,P)=(2\pi)^{-n}\delta(\Phi)\delta(\Pi)|[\Phi,\Pi]|e^{iS(\bar{t},\bar{Q}|t,Q,P)}, \tag{9.10}$$

corresponding to the skeletonized canonical action with the Hamiltonian (9.8).

The prescription (9.10) is superficially similar in form to our result (6.14) for the classical propagator of a parametrized theory. The gauge constraint $\Pi=0$ plays the role of the super-Hamiltonian constraint $H=0$ and the gauge fixing condition (9.9) replaces the condition

$$t(Q)-t(\tau)=0, \tag{9.11}$$

which selects the parametrization of path. [Due to Eq. (6.15), the factor Λ in the measure (6.17) has the meaning of the Poisson bracket between the expression (9.11) and the super-Hamiltonian H]. However, there are two important differences:

(I) The gauge fixing condition does not need to contain any reference to time. On the other hand, the condition (9.11) selecting the parametrization must introduce a prescribed function $t(\tau)$ of τ.

(II) In gauge theories, any function (9.9) of the extended coordinates and momenta is permissible. On the other hand, in a parametrized theory $t(Q)$ is a definite function on the extended configuration space. For our Newtonian system, the time function $t(Q)$ is obtained by the reconstruction procedure discussed in Sec. 2.

To see that the distinction (I) is vital, let us blindly apply a condition (9.9) appropriate for a gauge theory to our parametrized theory. In the simplest case, this is achieved by putting $t(\tau) = 0$ and identifying $t(Q)$ with $\Phi(Q,P)$. Of course, our derivation of Eqs. (6.16)–(6.17) for the quantum propagator is no longer valid because $t(\tau) = 0$ implies $t' = 0 = t''$. When we insist that the expression (6.16)–(6.17) represents the quantum propagator from t' to $t'' > t'$ even for $t(\tau) = 0$, we predictably end with an absurd result. On the other hand, when we put $t(\tau) = 0$ and simultaneously restrict ourselves to $t' = 0 = t''$, the expression (6.16)–(6.17) for the quantum propagator equally predictably yields a correct triviality: It reduces to the delta function because the dynamics is frozen at a single instant of time.

The distinction (I) reflects the fundamental physical difference between gauge theories and parametrized theories. The constraints which follow from gauge invariance generate gauge changes of the extended phase space variables. These are unobservable; the physical state of the system is unchanged. The constraints which follow from reparametrization invariance generate the dynamics of the system. They are observable and the physical state does change. It makes sense to fix a gauge to get one representative to a physical state. It makes no sense to fix the time.

The distinction (II) is more subtle. It means that the slices of a constant label time τ coincide with the leaves of the absolute time foliation. Such a restriction follows naturally from our derivation of Eqs. (6.16)–(6.17) for the quantum propagator. There is no simple modification of this derivation which would introduce a different foliation, e.g.,

$$\Phi(Q,\tau) = 0. \tag{9.12}$$

In fact, the Schrödinger equation ceases to be a first-order equation in the foliation label when we allow the general foliation (9.12) and the Hilbert space interpretation loses thereby its meaning. We thus consider it highly unlikely that the general foliation (9.12) would yield the correct quantum propagator when used instead of the absolute time foliation (9.11) in the expressions (6.16)–(6.17) for the path integral. We emphasize yet again that the choice of the time variable is a central decision in forming quantum theories and that, once made, it cannot be easily altered without altering the theory.

10. SUMMARY

The representation of the quantum propagator by a path integral of the exponentiated canonical action on the physical phase space is a natural starting point for quantum mechanics. The measure in the space of paths is induced by the invariant Liouville measure in phase space. The geometrically privileged transport of momentum by actual classical paths of the system leads to the skeletonization of the canonical action by the chain of phase-space principal functions. This privileged skeletonization removes the ambiguity connected with the factor ordering.

Unfortunately, not all classical theories are easily formulated in terms of the true physical degrees of freedom. Both gauge theories and parametrized theories use redundant variables. The dynamical evolution of the system takes place in extended spaces of variables. General relativity is the most prominent example of a system in which the simultaneous presence of gauge and parametrization makes it extremely difficult to return back to the physical phase space. It is thus essential to represent the quantum propagator by integrals over paths in such extended spaces of variables.

We have accomplished this program for parametrized Newtonian systems moving in curved configuration spaces. Our point of departure was the path integral in the physical phase space of the system. We arrived at equivalent path integrals in alternative spaces by extending or restricting the variables.

The extension of variables was always done so that integration over the new variables yielded the integral we have started from. Typical devices for ensuring this property are delta functions introduced into the measure or representations of known functions by integrals over a parameter. The restriction of variables was always carried out by integrating over them. Typically, the integrals involved were Gaussian integrals in the momenta which can be explicitly evaluated. Such integrals lead to nontrivial measures in spaces of remaining variables.

We summarize our results in Table III, which is a continuation of our Table II for the alternative forms of the action. In the first column, we write down a symbolic expression for the path integrals. The symbolic expression is interpreted by skeletonizing the measure and skeletonizing the action. In the second column, we enter the measure associated with a segment of skeletonized path between the gate $dX = dX_{(K)}$ at $X = X_{(K)}$ and the gate $d\overline{X} = dX_{(K+1)}$ at $\overline{X} = X_{(K+1)}$ in the space $\{X\}$ of appropriate variables. The total measure is the product of such elementary measures at all gates, $K = 0,1,\ldots,N-1$. In the following column, we give the number of the equation which introduces this measure in the main text. Some of the measures are quite complicated and do not follow a clearly recognizable pattern. On the other hand, the classical action is always skeletonized in the same manner: For each step of the skeletonized path, we write the initial value $L_{(K)}$ of the appropriate Lagrangian and multiply it by the interval $\Delta\tau_{(K)} = \tau_{(K+1)} - \tau_{(K)}$ of time. The initial values of velocities which enter into the Lagrangian must be expressed in terms of the configuration data at the boundaries of each step. This is achieved by using the appropriate Hamilton principal function obeying the standard Hamilton–Jacobi equations. Moreover, in the phase space versions of the theory, the initial metric entering into the Lagrangian must be replaced by a tensor–scalar coefficient which takes into account the geodesic deviation trans-

TABLE III. Alternative forms for path integrals.

Type of action	Quantum propagator represented by the path integral	Elementary measure of a segment of path	Skeletonized measure: Eq. number	Skeletonized action: Eq. number
Physical canonical action	$\langle t'', q''\|t', q'\rangle\, d^n q' = \int Dq\, Dp\, e^{iS[q,p]}$	$d^n q\, d^n p (2\pi)^{-n}$	(4.12)	(4.11)
Physical Lagrangian action	$\langle t'', q''\|t', q'\rangle\, d^n q' = \int \bar{D}q\, e^{iS[q]}$	$d^n q\, (2\pi i\, \Delta t)^{-n/2}\, \bar{g}^{1/2}(\bar{t}, \bar{q}\|t, q)$	(5.14)	(5.13)
Extended canonical action, conditional	$\langle Q''\|Q'\rangle \delta(t(Q') - t')\, d^{n+1} Q' = \int \bar{D}Q\, \bar{D}P\, e^{iS[Q,P]}$	$d^{n+1} Q\, d^{n+1} P (2\pi)^{-n} \Lambda(Q)$ $\delta(t(Q) - t(\tau))\delta(\bar{H}(\bar{Q}\|Q, P))$	(6.17)	(6.18)
Extended Lagrangian action, with lapse multiplier	$\langle Q''\|Q'\rangle \delta(t(Q') - t')\, d^{n+1} Q'\, dN' = \int \bar{D}Q\, \bar{D}P\, \bar{D}N\, e^{iS[Q,N]}$	$d^{n+1} Q\, d^{n+1} P\, dN\, \Delta t\, \Lambda(Q)$ $(2\pi)^{-(n+1)} \delta(t(Q) - t(\tau))$	(6.23)	(6.24)
Extended Lagrangian action, homogeneous	$\langle Q''\|Q'\rangle \delta(t(Q') - t')\, d^{n+1} Q' = \int \bar{D}Q\, e^{iS[Q]}$	$d^{n+1} Q\, [2\pi i T_C(Q) \dot{Q}^C \Delta\tau]^{-n/2}$ $\bar{D}^{-1/2}(\bar{Q}\|Q)\, \Lambda(Q) \delta(t(Q) - t(\tau))$	(7.24)	(7.23)
		$dt\, d^n q [2\pi i \dot{t}(\tau) \Delta\tau]^{-n/2}\, \bar{g}(\bar{t}, \bar{q}\|t, q) \delta(t - t(\tau))$	(7.28)	(7.29)
Extended Lagrangian action, with lapse multiplier	$\langle Q''\|Q'\rangle \delta(t(Q') - t')\, d^{n+1} Q'\, dN' = \int \bar{D}Q\, \bar{D}N\, e^{iS[Q,N]}$	$D^{n+1} Q\, dN (-N^{-2}) (2\pi)^{-1}$ $(2\pi i T_C(Q) \dot{Q}^C \Delta\tau)^{-(n-1)/2}$ $\bar{D}^{-1/2}(\bar{Q}\|Q) (G_{AB}(Q) \dot{Q}^A \dot{Q}^B)^{1/2}$ $\Lambda(Q) \delta(t(Q) - t(\tau))$	(8.13)	(8.12)
		$dt\, d^n q\, dN (-N^{(0)-2}) (2\pi)^{-1}$ $(2\pi i \dot{t}(\tau) \Delta\tau)^{-(n-1)/2}\, \bar{g}^{1/2}(\bar{t}, \bar{q}\|t, q)$ $(g_{ab}(q) \dot{q}^a \dot{q}^b)^{1/2} \delta(t - t(\tau))$	(8.20)	

port of momenta. The classical action is skeletonized by the sum $\Sigma_{K=0}^{N-1} L_{(K)}\, \Delta\tau_{(K)}$ of such contributions. Because the procedure follows a well-defined algorithm, there is no need to enter the skeletonized action into our table. We refer merely to the equation where it is discussed in the paper.

While the measures are often complicated, they have one feature in common—the occurrence of $\delta(t(Q) - t(\tau))$ which fixes the integrations to the leaves of absolute time that flows from the initial instant t' to the final instant t''. The specific form of this delta function is characteristic of parametrized theories and reflects the privileged role time plays in quantum mechanics.

ACKNOWLEDGMENT

This work was supported in part by NSF Grants PHY 81-06909, PHY 80-26043, and PHY 81-07384.

APPENDIX A: INTEGRABILITY CONDITIONS ON THE DEGENERATE METRIC G^{AB}

A degenerate metric G^{AB} with signature $(0; +, \ldots, +)$ has a unique degeneracy direction, i.e., the solutions T_A to the equation

$$G^{AB} T_B = 0 \tag{A1}$$

fill a ray. The ray determines a foliation if and only if it is surface forming,

$$M_{ABC} \equiv T_A T_{[B,C]} + T_B T_{[C,A]} + T_C T_{[A,B]} = 0. \tag{A2}$$

To be so, the metric G^{AB} cannot be arbitrary, but it must satisfy certain integrability conditions which we are now going to derive.

Note that the equation

$$T_A X^A = 0 \tag{A3}$$

has n linearly independent solutions Q^A_a, $a = 1, \ldots, n$ and that the metric G^{AB} is nondegenerate on the vector subspace spanned by Q^A_a:

$$G^{AB} = G^{ab} Q^A_a Q^B_b, \quad \det G^{ab} \neq 0. \tag{A4}$$

Let U^A be an arbitrary vector linearly independent of Q^A_a, i.e.,

$$T_A U^A \neq 0. \tag{A5}$$

The vectors $\{U^A, Q^A_a\}$ form a basis. Because G^{ab} is nondegenerate, any equation $M_A = 0$ can be replaced by an equivalent set of equations

$$G^{AB} M_B = 0, \quad U^B M_B = 0. \tag{A6}$$

Handling each index of the completely antisymmetric tensor M_{ABC} in this way, we can replace Eq. (A2) by an equivalent system of equations:

$$G^{KA} G^{LB} G^{MC} M_{ABC} = 0, \tag{A7}$$

$$G^{KA} G^{LB} U^C M_{ABC} = 0. \tag{A8}$$

Due to Eq. (A1), the condition (A7) is identically satisfied.

Further, because of Eqs. (A1) and (A5), the condition (A8) reduces to

$$G^{KA}G^{LB}T_{[A,B]} = 0. \tag{A9}$$

Using Eq. (A1) again, we cast Eq. (A9) into the form

$$G^{A[K,L]}T_A = 0, \tag{A10}$$

where

$$G^{AK,L} \equiv G^{AK}{}_{,B}G^{BL}. \tag{A11}$$

From Eqs. (A3) and (A5) we see that

$$\exists\ H^{KLa}:\ G^{A[K,L]} = H^{KLa}Q^A_a. \tag{A12}$$

An alternative way of writing Eq. (A12) is

$$\delta_{AA_1\cdots A_n}G^{A[B,C]}G^{A_1B_1}\cdots G^{A_nB_n} = 0. \tag{A13}$$

Here $\delta_{AA_1\cdots A_n}$ is a completely antisymmetric tensor density of weight -1 with $\delta_{012\cdots n} = 1$. Note that in a Newtonian space-time we cannot introduce the more usual Levi-Civita pseudotensor $\epsilon_{AA_1\cdots A_n}$ because the metric G^{AB} is degenerate.

Equation (A13) is equivalent to the condition (A12) which is a necessary and sufficient condition for the degeneracy covector T_A determined by Eq. (A1) to be surface-forming.

APPENDIX B: DETERMINANTS WITH DEGENERATE METRICS

The metric G^{AB} is degenerate, and its determinant thus vanishes. However, we can project G^{AB} into the subspace orthogonal to the degeneracy direction T_A and take the determinant of the projected metric.

For a given G^{AB} and U^A, Eqs. (2.21) and (2.24) have a unique solution T_A. Furthermore, the equation

$$U^A X_A = 0 \tag{B1}$$

has n linearly independent solutions Q^a_A, $a = 1,\ldots,n$:

$$U^A Q^a_A = 0. \tag{B2}$$

The covectors $\{T_A, Q^a_A\}$ form a cobasis. Of course, Q^a_A can be changed by a transformation

$$Q^{a^*}_A = \Lambda^{a^*}_b(Q)Q^b_A. \tag{B3}$$

We introduce the alternating symbol $\delta_{a_n\cdots a_n}$ which transforms as a tensor density of weight -1 under the Λ transformations (B3). Besides it, we have at our disposal the alternating symbol $\delta^{AA_1\cdots A_n}$, which transforms as a tensor density of weight 1 under transformations of extended coordinates.

The projection

$$G^{ab} \equiv G^{AB}Q^a_A Q^b_B \tag{B4}$$

of the degenerate metric G^{AB} is nondegenerate, and we can write its determinant as

$$G^{-1} = (1/n!)\delta_{a_1\cdots a_n}G^{a_1b_1}\cdots G^{a_nb_n}\delta_{b_1\cdots b_n}. \tag{B5}$$

In terms of the original metric,

$$G^{-1} = (1/n!)\delta_{a_1\cdots a_n}Q^{a_1}_{A_1}\cdots Q^{a_n}_{A_n} \times G^{A_1B_1}\cdots G^{A_nB_n}\delta_{b_1\cdots b_n}Q^{b_1}_{B_1}\cdots Q^{b_n}_{B_n}. \tag{B6}$$

Study now the expression

$$D = (1/n!)\delta_{AA_1\cdots A_n}U^A U^B G^{A_1B_1}\cdots G^{A_nB_n}\delta_{BB_1\cdots B_n}. \tag{B7}$$

The tensor density $U^A\delta_{AA_1\cdots A_n}$ has two properties: (1) It is completely antisymmetric in $A_1,\ldots,A_n$, and (2) it is orthogonal to U^A. As a consequence, we must have

$$U^A\delta_{AA_1\cdots A_n} = J^{-1}\delta_{a_1\cdots a_n}Q^{a_1}_{A_1}\cdots Q^{a_n}_{A_n}. \tag{B8}$$

To determine the proportionality factor J^{-1}, we multiply Eq. (B8) by $\delta^{BA_1\cdots A_n}$. Because

$$\delta_{AA_1\cdots A_n}\delta^{BA_1\cdots A_n} = n!\delta^B_A, \tag{B9}$$

we get

$$n!U^B = J^{-1}\delta_{a_1\cdots a_n}Q^{a_1}_{A_1}\cdots Q^{a_n}_{A_n}\delta^{BA_1\cdots A_n}. \tag{B10}$$

Multiplication by T_B yields

$$J = (1/n!)\delta_{a_1\cdots a_n}T_A Q^{a_1}_{A_1}\cdots Q^{a_n}_{A_n}\delta^{AA_1\cdots A_n}. \tag{B11}$$

By introducing Eqs. (B8) and (B11) into the expression (B7), we learn that

$$G^{-1} = J^2 D. \tag{B12}$$

Any covector Π_A can be split into a part along T_A and a part perpendicular to U^A,

$$\Pi_A = \Pi_\parallel T_A + \Pi_A Q^a_A. \tag{B13}$$

Equation (B13) can be considered as a transformation from the variables $\Pi_\parallel, \Pi_a$ to the variables Π_A. The Jacobi matrix of this transformation is

$$\frac{\partial\{\Pi_A\}}{\partial\{\Pi_\parallel,\Pi_a\}} = \begin{vmatrix} T_A \\ Q^a_A \end{vmatrix}. \tag{B14}$$

We see that J is nothing else but the Jacobian of the transformation (B13).

We can replace the metric G^{AB} by the tensor–scalar coefficient $\overline{G}^{AB}$ and introduce appropriate quantities (B4), (B5), and (B7). We place bars over symbols denoting these quantities: $\overline{G}^{ab}, \overline{G}, \overline{D}$. The modified quantities are again connected by the equation

$$\overline{G}^{-1} = J^2\overline{D}. \tag{B15}$$

Mutatis mutandis, the same line of reasoning applies to nondegenerate metrics. Take a regular metric G^{ab}, $a = 1,\ldots,n$, and a vector $\dot{Q}^a$. Let Q^α_a, $\alpha = 1,\ldots,n-1$, be a basis in cotangent space orthogonal to $\dot{Q}^a$:

$$\dot{Q}^a Q^\alpha_a = 0. \tag{B16}$$

Project the metric G^{ab},

$$G^{\alpha\beta} \equiv G^{ab}Q^\alpha_a Q^\beta_b. \tag{B17}$$

The projected metric $G^{\alpha\beta}$ is again regular, and we can introduce its inverse $G_{\alpha\beta}$. Greek indices are raised by $G^{\alpha\beta}$ and lowered by $G_{\alpha\beta}$. Similarly, Latin indices are raised by G^{ab} and lowered by G_{ab}. With this convention,

$$G_{ab} = G_{\alpha\beta}Q^\alpha_a Q^\beta_b + \dot{Q}^{-2}\dot{Q}_a\dot{Q}_b, \tag{B18}$$

with

$$\dot{Q}^2 \equiv g_{ab}\dot{Q}^a\dot{Q}^b. \tag{B19}$$

We take the determinant of Eq. (B18). Because $G_{\alpha\beta}Q^\alpha_a Q^\beta_b$ and $\dot{Q}_a\dot{Q}_b$ are degenerate matrices,

$$G = (1/n!)\delta^{aa_1\cdots a_{n-1}} G_{ab} G_{a_1 b_1} \cdots G_{a_{n-1} b_{n-1}} \delta^{bb_1\cdots b_{n-1}}$$
$$= [1/(n-1)!]\dot{Q}^{-2}\delta^{aa_1\cdots a_{n-1}}\delta^{bb_1\cdots b_{n-1}}$$
$$\times \dot{Q}_a \delta^{aa_1\cdots a_{n-1}} Q^{\alpha_1}_{a_1}\cdots Q^{\alpha_{n-1}}_{a_{n-1}}$$
$$\times \dot{Q}_b \delta^{bb_1\cdots b_{n-1}} Q^{\beta_1}_{b_1}\cdots Q^{\beta_{n-1}}_{b_{n-1}} G_{\alpha_1\beta_1}\cdots G_{\alpha_{n-1}\beta_{n-1}}. \quad \text{(B20)}$$

As in Eqs. (B8) and (B11),

$$\dot{Q}_a \delta^{aa_1\cdots a_{n-1}} Q^{\alpha_1}_{a_1}\cdots Q^{\alpha_{n-1}}_{a_{n-1}} = \tilde{J}\delta^{\alpha_1\cdots\alpha_{n-1}}, \quad \text{(B21)}$$

with

$$\tilde{J} = [1/(n-1)!]\delta^{aa_1\cdots a_{n-1}}\dot{Q}_a Q^{\alpha_1}_{a_1}\cdots Q^{\alpha_{n-1}}_{a_{n-1}}\delta_{\alpha_1\cdots\alpha_{n-1}}. \quad \text{(B22)}$$

As a result,

$$G = \tilde{J}^2\dot{Q}^{-2}\det G_{\alpha\beta}. \quad \text{(B23)}$$

We multiply Eq. (B12) by Eq. (B23) and conclude that

$$\tilde{J}J \det{}^{1/2} G_{\alpha\beta} = (G_{AB}\dot{Q}^A\dot{Q}^B)^{1/2} D^{-1/2}. \quad \text{(B24)}$$

[1]The literature on the implementation of quantum dynamics by path integrals for nonrelativistic and relativistic systems in curved and flat configuration spaces is extensive and too large to be cited here. A useful general survey with extensive references to the original literature is L. S. Shulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1982).

[2]K. Kuchař, J. Math. Phys. **24**, 2122 (1983).

[3]L. Faddeev, Teor. Mat. Fiz. **1**, 3 (1969); L. Faddeev and V. Popov, Phys. Lett. B **25**, 30 (1967); L. Faddeev and V. Popov, Usp. Fiz. Nauk **111**, 427 (1973) [Sov. Phys. Usp. **16**, 777 (1974)]; E. S. Fradkin, and G. A. Vilkovisky "Quantization of Relativistic Systems with Constraints, Equivalence of Canonical and Covariant Formalisms in the Quantum Theory of the Gravitational Field," CERN Report TH-2332, 1977; L. Faddeev and A. Slavnov, *Gauge Fields: Introduction to Quantum Theory* (Benjamin, Reading, MA, 1980).

[4]K. Kuchař, Phys. Rev. D **22**, 1285 (1980).

Quantum energy-entropy inequalities: A new method for proving the absence of symmetry breaking

M. Fannes,[a] P. Vanheuverzwijn,[b] and A. Verbeure
Instituut voor Theoretische Fysica, Universiteit Leuven, B-3030 Leuven, Belgium

(Received 25 January 1983; accepted for publication 10 June 1983)

For quantum systems we develop a new method, based on a general energy-entropy inequality, to rule out spontaneous breaking of symmetries. The main advantage of our scheme consists in its clear-cut physical significance and its new areas of applicability; in particular we can handle discrete symmetry groups as well as continuous ones. Finally a few illustrations are discussed.

PACS numbers: 03.65. − w, 05.50. + q, 02.20. + b

I. INTRODUCTION

In the case of classical lattice systems we derived recently[1] correlation inequalities expressing the balance between energy and entropy for an equilibrium state. These inequalities were shown to reproduce easily the sharpest results concerning spontaneous magnetization in long range Ising models[2] and they gave a more direct and intuitive understanding of the underlying physics. Maybe even more important is the applicability to continuous as well as to discrete symmetry groups. In particular we proved translation invariance for one-dimensional systems under very weak conditions on the potential.[1]

Here we are concerned with the quantum-mechanical situation. The well-known method to prove absence of symmetry breaking is based on the Bogoliubov inequality. The first results along this line are the celebrated theorems of Mermin–Wagner[3] and Hohenberg.[4] Recently there was a revival of interest in the field. The best results along this line can be found in Ref. 5. It is important to remark that this method is restricted to continuous symmetry groups as the occurrence of an infinitesimal generator is essential for the method. On the contrary our method allows also for discrete symmetries. To stress this fact we will concentrate on the applications to discrete symmetries.

Our main tool is the correlation inequality [see formula (2) below] which has a clear physical significance as being an expression for the change of free energy under a dissipative perturbation of the equilibrium state.[1,6]

One should mention here also the results based on relative entropy considerations.[7] This technique as well allows for the treatment of discrete symmetries; however, our method based on the inequality seems to us more direct and intuitive.

II. ABSENCE OF SYMMETRY BREAKING

Let $(\mathscr{A},\alpha_t)$ be a C^*-dynamical system, i.e., $\mathscr{A}$ a C^*-algebra and α_t ($t\in\mathbb{R}$) is a strongly continuous one-parameter group of *-automorphisms of $\mathscr{A}$. A state ω of $\mathscr{A}$ satisfies the KMS condition for the evolution α_t at inverse temperature β, if $\omega(x\,\alpha_{i\beta}(y)) = \omega(yx)$ for all x,y in a norm dense, α_t-invariant *-subalgebra of $\mathscr{A}$. Let $\mathfrak{H}$ be the GNS representation space of the state ω and $\Omega\in\mathfrak{H}$ the cyclic vector; we denote by $\mathscr{M}$ the von Neumann algebra $\mathscr{A}''$ and by H the infinitesimal generator of the time evolution on $\mathfrak{H}$. As ω is time invariant we have $\Omega\in\mathscr{D}(H)$ (domain of H) and $H\Omega = 0$.

If

$$H = \int_{-\infty}^{\infty} \lambda\, dE(\lambda)$$

is the spectral decomposition of the Hamiltonian H, define for all $x\in\mathscr{M}$ the measures on $\mathbb{R}$

$$d\mu_x(\lambda) = (x\Omega, dE(\lambda)x\Omega),$$
$$d\nu_x(\lambda) = (x\Omega, dE(-\lambda)x^*\Omega).$$

As ω is a KMS state the measures μ_x and ν_x are equivalent with Radon–Nikodym derivative

$$\frac{d\mu_x(\lambda)}{d\nu_x(\lambda)} = e^{\beta\lambda} \tag{1}$$

(see, e.g., Ref. 8, Proposition 5.3.14).

We start with an easy derivation of an inequality for KMS states which was stated implicitly for the first time in Ref. 9.

For all $x\in\mathscr{M}$ such that $x\Omega\in\mathscr{D}(H)$,

$$\begin{aligned}\frac{\beta(x\Omega,Hx\Omega)}{(x\Omega,x\Omega)} &= \frac{\int \beta\lambda\, d\mu_x(\lambda)}{\int d\mu_x(\lambda)} \\ &= -\ln\exp -\frac{\int \beta\lambda\, d\mu_x(\lambda)}{\int d\mu_x(\lambda)} \\ &\geqslant -\ln\frac{\int e^{-\beta\lambda}\, d\mu_x(\lambda)}{\int d\mu_x(\lambda)} \\ &= -\ln\frac{\int d\nu_x(\lambda)}{\int d\mu_x(\lambda)} \\ &= \ln\frac{(x\Omega,x\Omega)}{(x^*\Omega,x^*\Omega)}\end{aligned}$$

by the Jenssen inequality. Hence

$$\beta\omega(x^*Hx\Omega) \geqslant \omega(x^*x)\ln(\omega(x^*x)/\omega(xx^*)). \tag{2}$$

Lemma II.1: Let I be a finite interval of $\mathbb{R}$. If for $0\neq x\in\mathscr{M}\cap\mathscr{D}([H,\cdot])$ $\operatorname{supp}\mu_x \subset I$ then if ω satisfies the KMS condition,

$$0 \leqslant \beta\omega(x^*Hx) - \omega(x^*x)\ln\frac{\omega(x^*x)}{\omega(xx^*)} \leqslant \beta\omega(x^*x)\Delta,$$

where Δ is the length of the interval I.

Proof: Let $I = [\lambda_1,\lambda_2]$, $\lambda_i\in\mathbb{R}$; using (1) and (2) we compute

[a] Bevoegdverklaard navorser NFWO, Belgium.
[b] Aangesteld navorser NFWO, Belgium.

 0022-2488/84/010076-03$02.50

$$0 \leqslant \beta\omega(x^*Hx) - \omega(x^*x)\ln\frac{\omega(x^*x)}{\omega(xx^*)}$$

$$= \beta\int \lambda\, d\mu_x(\lambda) - \int d\mu_x(\lambda)\ln\frac{\int d\mu_x(\lambda)}{\int e^{-\beta\lambda}d\mu_x(\lambda)}$$

$$\leqslant \beta\lambda_2\int d\mu_x(\lambda) - \int d\mu_x(\lambda)\ln\frac{1}{e^{-\beta\lambda_1}}$$

$$= \beta(\lambda_2 - \lambda_1)\int d\mu_x(\lambda). \quad \blacksquare$$

Now we proceed to our main objective, namely, the development of a theory for the absence of spontaneous symmetry breaking. We suppose that we have a symmetry represented by a *-automorphism τ of $\mathscr{A}$ satisfying the following conditions:

(a) τ is approximately inner, i.e., there exists a sequence $(u_n)_{n\geqslant 1}$ of unitaries in $\mathscr{A}$ such that for all $x\in\mathscr{A}$,

$$\lim_{n\to\infty} \|\tau(x) - u_n^* x u_n\| = 0.$$

This condition implies

$$\lim_{n\to\infty} \omega(u_u^* x u_n) = \omega(\tau(x)) \tag{3}$$

for all states ω of $\mathscr{A}$. This notion of approximately inner automorphism has been introduced in Ref. 10. As far as the physics is concerned it means that the automorphism can be approximated by local unitary transformations.

(b) As τ represents a symmetry of the system we have $[\alpha_t,\tau] = 0$ for all $t\in\mathbb{R}$. Furthermore, we suppose that the local approximations almost commute with α_t in the sense that for all m: $u_m\in\mathscr{D}([H,\cdot])$ and

$$K = \sup_m \|[H,u_m^*]\| < \infty. \tag{4}$$

This is essentially the condition used in Ref. 7.

Theorem II.2: Let ω be a KMS state with respect to the evolution α_t at inverse temperature β; let τ be a symmetry as above. Then there exists a constant C such that for all $x\in\mathscr{A}$, $\omega(xx^*)\leqslant C\omega(\tau(xx^*))$ holds.

Proof: For $f\in C_c^\infty(\mathbb{R})$ (the space of infinitely differentiable functions with compact support) and for any $x\in\mathscr{A}$ we denote

$$x(f) = \int dt\,\hat{f}(t)\alpha_t(x),$$

where $f(\lambda) = \int dt\,\hat{f}(t)e^{it\lambda}$.

For $\epsilon > 0$ one finds a decomposition of the identity by a sequence $(h_n)_{n\geqslant 1}$ of positive functions in C_c^∞ such that pointwise $\Sigma_{n\geqslant 1} h_n^2 = 1$ and such that the support of each h_n is contained in an interval of length ϵ.

By a straightforward computation one gets

$$\omega(xx^*) = \int d\nu_x(-\lambda) = \sum_n \int h_n(\lambda)^2 d\nu_x(-\lambda)$$

$$= \sum_n \omega(x(h_n)x(h_n)^*). \tag{5}$$

Substitute in the correlation inequality (2) the observable x by $u_m^* x(h_n)$ for each n such that $x(h_n)\neq 0$; adding and subtracting a term and using time invariance

$$\omega(x(h_n)^*x(h_n))\ln\frac{\omega(x(h_n)x(h_n)^*)}{\omega(u_m^*x(h_n)x(h_n)^*u_m)}$$

$$-\beta\omega(x(h_n)^*u_m[H,u_m^*]x(h_n))$$

$$\leqslant \beta\omega(x(h_n)^*Hx(h_n)) - \omega(x(h_n)^*x(h_n))\ln\frac{\omega(x(h_n)^*x(h_n))}{\omega(x(h_n)x(h_n)^*)}$$

$$\leqslant \beta\epsilon\omega(x(h_n)^*x(h_n)),$$

where the last inequality is obtained from Lemma II.1 as the support of h_n is contained in an interval of length less than ϵ. Hence by (4),

$$\omega(x(h_n)x(h_n)^*) \leqslant e^{(K+\epsilon)\beta}\omega(u_m^*x(h_n)x(h_n)^*u_m),$$

and by (3)

$$\omega(x(h_n)x(h_n)^*) \leqslant e^{(K+\epsilon)\beta}\omega(\tau(x(h_n)x(h_n)^*)).$$

As $[\tau,\alpha_t] = 0$ one has $\tau(x(f)) = (\tau x)(f)$; hence after summation over n, using (5) one gets

$$\omega(xx^*) \leqslant e^{\beta(K+\epsilon)}\omega(\tau(xx^*)). \quad \blacksquare$$

At this point it might be interesting to remark that this result of absolute continuity of states is obtained through the use of the correlation inequality. It is worthwhile to mention the work of Araki[11] and of Sakai.[12] They are interested in the problem of unicity of KMS states. Sakai is also working towards a result expressing absolute continuity of states but by explicit calculations using the Gibbs form of the state. Araki's technique is based on the notion of relative entropy and leads to quasiequivalence of states.

Finally one gets as an easy consequence the invariance of the equilibrium states under the symmetry group.

Corollary II.3: Under the conditions of Theorem II.2

$$\omega\circ\tau = \omega.$$

Proof: It is sufficient to prove the corollary for extremal KMS states. Suppose that ω is such an external state. Then, as $[\tau,\alpha_t] = 0$, $\omega\circ\tau$ is also an extremal KMS state. By Theorem II.2 and a well-known property (Ref. 8, Theorem 5.3.29) there exists $T\in\mathscr{A}''\cap\mathscr{A}'$ such that

$$\omega(\tau(x)) = \langle\Omega_\omega|Tx\Omega_\omega\rangle.$$

As ω is extremal $T = 1$ and therefore $\omega = \omega\circ\tau$. $\blacksquare$

III. ILLUSTRATION

We prove the absence of breaking of translation symmetry in one-dimensional lattice systems for long-range interactions. This result was announced in Ref. 13. The algebra of observables is the usual tensor product algebra

generated by the local algebras $\mathscr{A}_\Lambda = \bigotimes_{k\in\Lambda}\mathscr{B}(\mathfrak{H})$, where $\mathfrak{H}$ is a finite-dimensional Hilbert space.

Consider the local Hamiltonian

$$H_N = \sum_{-N\leqslant i<j\leqslant N}\sum_{rs} J_{rs}(|i-j|)\sigma_i^r\sigma_j^s + \sum_r h_r\sum_{i=-N}^{N}\sigma_i^r,$$

where $\{\sigma_i^r | r = 1,\ldots,d\}$ are the spin matrices for the lattice site i; the interaction energies $J_{rs}(k)$ satisfy

$$\sum_{k=1}^{\infty}|J_{rs}(k)| < \infty. \tag{6}$$

This condition guarantees a good thermodynamic behavior of the system.

Now we want to apply Theorem II.2. The symmetry τ is the translation over one lattice site, i.e., $\tau(\sigma_i^r) = \sigma_{i+1}^r$. Note that τ is approximately inner since it can be approximated by τ_m standing for the cyclic translation of the lattice interval $[-m, +m]$ such that

$$\tau_m(\sigma_i^r) = \sigma_{i+1}^r \quad \text{if } -m \leqslant i < m,$$
$$\tau_m(\sigma_m^r) = \sigma_{-m}^r,$$
$$\tau_m(\sigma_j^r) = \sigma_j^r \quad \text{if } |j| > m.$$

It is easy to check that there exist unitary operators u_m such that $\tau_m(x) = u_m^* x u_m$ for all elements of $\mathscr{A}$. Clearly for all $x \in \cup_\Lambda \mathscr{A}_\Lambda$ one has $\tau(x) = \tau_m(x)$ when m is large enough. Therefore formula (3) holds. Furthermore, because of condition (6) the time evolution automorphisms α_t are well defined as[8]

$$\alpha_t(x) = \lim_N e^{-itH_N} x e^{-itH_N}$$

on the C^*-algebra generated by $\cup_\Lambda \mathscr{A}_\Lambda$ and clearly $[\alpha_t, \tau] = 0$. Suppose now that for all $r,s = 1,\ldots,d$,

$$\sum_{k=1}^{\infty} k \left| J_{rs}(k) - J_{rs}(k-1) \right| < \infty; \tag{7}$$

then

$$\sup_m \| [H, u_m^*] \| = \sup_m \| u_m [H, u_m^*] \|$$
$$= \sup_m \| \lim_N (\tau_m^{-1}(H_N) - H_N) \|$$
$$\leqslant \sum_{rs} \left\{ 2 \sum_{k=1}^{\infty} k \left| J_{rs}(k) - J_{rs}(k-1) \right| + 12 \sum_{k=1}^{\infty} |J_{rs}(k)| \right\} < \infty.$$

Hence (4) is satisfied and by Theorem II.2 each KMS state ω satisfies

$$\omega(xx^*) \leqslant C\omega(\tau(xx^*)).$$

By Corollary II.3 $\omega = \omega \circ \tau$ and we proved that any equilibrium state of the system is translation invariant if the interaction energies satisfy condition (7). It is instructive to realize that in the ferromagnetic or antiferromagnetic case [i.e., the $J_{rs}(k)$ have the same sign] condition (7) follows from condition (6) if the function $k \to J_{rs}(k)$ is monotonic for large k.

Finally we remark that, although we considered here only a one-dimensional system, our method extends to higher-dimensional ones, e.g., it provides a short proof of the absence of breaking of internal symmetries in two-dimensional quantum lattice systems.[7] Furthermore, the proof of Theorem II.2 relies on an estimate for $\omega(x^* u_m [H, u_m^*] x)$ given by condition (4). Depending on the particular model under consideration more refined estimates might be obtained weakening condition (4) on the interaction and hence extending the range of applicability of the theorem.

[1] M. Fannes, P. Vanheuverzwijn, and A. Verbeure, J. Stat. Phys. **29**, 545–558 (1982).
[2] B. Simon and A. D. Sokal, J. Stat. Phys. **25**, 679 (1981).
[3] N. D. Mermin and H. Wagner, Phys. Rev. Lett. **17**, 1133 (1966).
[4] P. C. Hohenberg, Phys. Rev. **158**, 383 (1967).
[5] C. A. Bonato, J. F. Perez, and A. Klein, J. Stat. Phys. **29**, 159 (1982).
[6] M. Fannes and A. Verbeure, J. Math. Phys. **19**, 558 (1978).
[7] J. Fröhlich and C. E. Pfister, Commun. Math. Phys. **81**, 277 (1981).
[8] D. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics II* (Springer-Verlag, New York, 1981).
[9] G. Roepstorff, Commun. Math. Phys. **46**, 253 (1976).
[10] R. T. Powers and S. Sakai, Commun. Math. Phys. **39**, 273 (1975).
[11] H. Araki, Commun. Math. Phys. **44**, 1 (1975).
[12] S. Sakai, J. Functional Analysis **27**, 203 (1976).
[13] M. Fannes, P. Vanheuverzwijn, and A. Verbeure, "Quantum Energy-Entropy Balance and Breaking of Symmetries," Preprint KUL-TF-82/19.

Quantum measuring processes of continuous observables

Masanao Ozawa
Department of Information Sciences, Tokyo Institute of Technology, Oh-Okayama, Meguro-ku, Tokyo 152, Japan

(Received 3 May 1983; accepted for publication 23 June 1983)

The purpose of this paper is to provide a basis of theory of measurements of continuous observables. We generalize von Neumann's description of measuring processes of discrete quantum observables in terms of interaction between the measured system and the apparatus to continuous observables, and show how every such measuring process determines the state change caused by the measurement. We establish a one-to-one correspondence between completely positive instruments in the sense of Davies and Lewis and the state changes determined by the measuring processes. We also prove that there are no weakly repeatable completely positive instruments of nondiscrete observables in the standard formulation of quantum mechanics, so that there are no measuring processes of nondiscrete observables whose state changes satisfy the repeatability hypothesis. A proof of the Wigner–Araki–Yanase theorem on the nonexistence of repeatable measurements of observables not commuting conserved quantities is given in our framework. We also discuss the implication of these results for the recent results due to Srinivas and due to Mercer on measurements of continuous observables.

PACS numbers: 03.65.Bz, 02.50. + s

1. INTRODUCTION

In the last decade, some attempts were developed to construct a satisfactory theory of the quantum mechanical measurement of an observable with continuous spectrum.[1–9] However, we have found no satisfactory solution of the fundamental problem to determine the state changes caused by measurements of continuous observables. In spite of these difficulties in continuous spectrum, the theory for discrete spectrum has a conventionally accepted solution since the pioneering work of von Neumann.[10]

Let $A = \Sigma_i \lambda_i P_i$ be an observable with simple discrete spectrum $\lambda_1, \lambda_2, \cdots$. Then von Neumann[10] showed the following:

(1) By the repeatability hypothesis, the state change $\rho \to \rho'$ caused by the measurement of A is determined as $\rho' = \Sigma_i P_i \rho P_i$.

(2) The above state change $\rho \to \rho'$ is compatible with the Hamiltonian formalism in the description of the measuring process in terms of the time evolution of the composite system of the observed system and the measuring apparatus.

In the present paper, we shall show the following:

(1) The description of measuring processes has a satisfactory generalization to continuous observables.

(2) Every measuring process determines a state change caused by the measurement.

(3) There are no measuring processes of a nondiscrete observable whose state changes satisfy the repeatability hypothesis.

In order to clarify the present situation, we shall review some developments on the problem so far. In the early stage, Umegaki and Nakamura[11] showed that the state change $\rho \to \rho' = \Sigma_i P_i \rho P_i$ is just an example of Umegaki's noncommutative conditional expectations[12] onto the von Neumann algebra generated by A, and they conjectured that the state change caused by the measurement of a continuous observable would also be such a noncommutative conditional expectation. However, it is shown by Areveson[13] that such conditional expectations do not exist for continuous observables. In view of these results, Davies and Lewis[1] established the mathematical concept of instruments which enables us to treat statistical correlations of outcomes of successive measurements, and formulate the repeatability hypothesis for continuous observables. They conjectured the nonexistence of repeatable instruments for continuous observables and proposed the more flexible approach to measurements of continuous observables abandoning repeatability hypothesis. Recently, Srinivas[8] generalized the concept of instruments and showed the existence of such generalized instruments for continuous observables which satisfy the repeatability hypothesis. He proposed a generalized collapse postulate which determines such repeatable generalized instruments to describe the state changes caused by measurements of continuous observables. More recently, Mercer[9] considered a wider class of state transformations than conditional expectations and proposed the state change should be described by such a transformation with the locality introduced by him. It is a remarkable fact that these attempts are concerned only with the first half of von Neumann's work cited above. An operator theoretical analysis on von Neumann's second result was done by Kraus.[14] He established the complete positivity of state changes caused by the general measuring processes, but his result is concerned only with the yes–no measurements.

In this paper, we shall show that the state changes determined by measuring processes naturally correspond to completely positive instruments and vice versa. We prove Davies and Lewis's conjecture for completely positive instruments, i.e., completely positive instruments cannot be weakly repeatable unless the corresponding observable is discrete. These results show that Srinivas's generalized col-

 0022-2488/84/010079-09$02.50

lapse postulate cannot be compatible for continuous observables with the Hamiltonian description of measuring processes. We shall also show that if they can be realized by some measuring processes, Mercer's local transition maps correspond to repeatable measurements, and hence they cannot exist for continuous observables.

The nonexistence of repeatable measuring processes of continuous observables suggests that we should investigate the approximately repeatable measuring processes as models of measurements in quantum mechanics. Moreover, this direction of investigation is appropriate not only for continuous observables. Indeed, even in measurements of discrete observables, it is known that the repeatable measurement is impossible unless observed quantity commutes with conserved quantity under some conservation law (see Refs. 15 and 16, also Sec. 8). The author believes that, in future investigations on really existing approximately repeatable measurements, our framework of measuring processes will provide a nice basis. However, we shall discuss these problems elsewhere.

In Sec. 2, we give some preliminaries on semiobservables and conditional expectations. Our concept of observed quantities allows the nonorthogonal resolutions of identity, called semiobservables. In Sec. 3, we generalize von Neumann's measuring processes to continuous observables and show that every measuring process determines the state change caused by the measurement. In Sec. 4, we provide a dilation theorem and a decomposition theorem of completely positive instruments which are useful in the later sections. In Sec. 5, we shall establish the one-to-one correspondence between measuring processes and completely positive instruments. If the observed quantity is a usual one, the obtained correspondence is reduced to very simple form by the decomposition theorem, that is, measuring processes are determined by their transition $\rho\to\rho'$. In Sec. 6, we study the repeatability hypothesis and prove the nonexistence of weakly repeatable completely positive instruments for nondiscrete observables in the standard formulation of quantum mechanics. In Sec. 7, we study the local transition maps and prove the nonexistence of local transition maps corresponding to measuring processes of nondiscrete observables. In Sec. 8, we shall give a proof of the Wigner–Araki–Yanase theorem in our framework, which states the nonexistence of repeatable measuring processes of the observables which do not commute with the conserved quantity. In Sec. 9, we shall give a characterization of the measuring processes discussed in the conventional measurement theory among our general measuring processes.

2. OBSERVABLES AND CONDITIONAL EXPECTATIONS

Let $\mathscr{H}$ be a Hilbert space. Denote by $\mathscr{L}(\mathscr{H})$ the algebra of bounded operators on $\mathscr{H}$ and by $\mathscr{T}(\mathscr{H})$ the space of trace class operators on $\mathscr{H}$. A *state* ρ on $\mathscr{H}$ is a positive trace one operator on $\mathscr{H}$. Denote by $\mathscr{S}(\mathscr{H})$ the space of all states on $\mathscr{H}$. Let $(\Omega,\mathscr{B})$ be a Borel space. A *semiobservable* X on $\mathscr{H}$ with value space $(\Omega,\mathscr{B})$ is a positive operator valued measure $X:\mathscr{B}\to\mathscr{L}(\mathscr{H})$ such that $X(\Omega)=1$. An *observable* X is a semiobservable which is projection valued. Denote by $\mathscr{B}(\mathbb{R}^n)$ the Borel σ-field of $\mathbb{R}^n$. By the spectral theory, we shall identify an observable X on $\mathscr{H}$ with value space $(\mathbb{R}^n,\mathscr{B}(R^n))$ and the corresponding mutually commutable family $\{x_1,\ldots,x_n\}$ of self-adjoint operators on $\mathscr{H}$ such that

$$x_i=\int_{\mathbb{R}}\lambda X(\mathbb{R}\times\cdots\times d\hat{\lambda}\times\cdots\times\mathbb{R}). \tag{2.1}$$

An observable X with value space $(\mathbb{R},\mathscr{B}(R))$ is called *bounded* if $x=\int_{\mathbb{R}}\lambda X(d\lambda)$ is bounded. Let X be a semiobservable on $\mathscr{H}$ with value space $(\Omega,\mathscr{B})$. If the system is in the state ρ at the instant before a measurement of X, then the probability distribution $\mathrm{Prob}(X\in B;\rho)$ of the outcomes of this measurement is given by

$$\mathrm{Prob}(X\in B;\rho)=\mathrm{Tr}[\rho X(B)], \tag{2.2}$$

for any B in $\mathscr{B}$. For a semiobservable X, we shall denote by $X(\mathscr{B})$ the range of X, i.e., $X(\mathscr{B})=\{X(B);B\in\mathscr{B}\}$. A *conditional expectation* T on $\mathscr{L}(\mathscr{H})$ onto a von Neumann algebra $\mathscr{M}$ on $\mathscr{H}$ is a normal completely positive linear map T on $\mathscr{L}(\mathscr{H})$ with range $\mathscr{M}$ such that $T(axb)=aT(x)b$ for all a,b in $\mathscr{M}$, x in $\mathscr{L}(\mathscr{H})$. It is known[17] that an ultraweakly continuous linear map T on $\mathscr{L}(\mathscr{H})$ is a conditional expectation if and only if it is a projection of norm 1 onto $\mathscr{M}$.

Let $\mathscr{K}$ be another Hilbert space. Let σ be a state on $\mathscr{K}$. Then the formula

$$\mathrm{Tr}[\rho E_\sigma(x)]=\mathrm{Tr}[(\rho\otimes\sigma)x], \tag{2.3}$$

where $x\in\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$ and $\rho\in\mathscr{T}(\mathscr{H})$, defines a normal completely positive linear map $E_\sigma:\mathscr{L}(\mathscr{H}\otimes\mathscr{K})\to\mathscr{L}(\mathscr{H})$ such that $E_\sigma(a\otimes 1)=a$ for any a in $\mathscr{L}(\mathscr{H})$. Thus the formula $x\to E_\sigma(x)\otimes 1$, for x in $\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$, defines a conditional expectation on $\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$ onto $\mathscr{L}(\mathscr{H})\otimes\mathbb{C}1$. It is easily seen that the map E_σ is the adjoint of the map $\rho\to\rho\otimes\sigma$ from $\mathscr{T}(\mathscr{H})$ into $\mathscr{T}(\mathscr{H}\otimes\mathscr{K})$. The formula

$$\mathrm{Tr}[E_{\mathscr{K}}(\phi)a]=\mathrm{Tr}[\phi(a\otimes 1)], \tag{2.4}$$

where $\phi\in\mathscr{T}(\mathscr{H}\otimes\mathscr{K})$ and $a\in\mathscr{L}(\mathscr{H})$, defines a completely positive linear map $E_{\mathscr{K}}:\mathscr{T}(\mathscr{H}\otimes\mathscr{K})\to\mathscr{T}(\mathscr{H})$, which is called the *partial trace* over $\mathscr{K}$. The partial trace $E_{\mathscr{K}}$ also satisfies that for any ξ,η in $\mathscr{H}$, and any orthogonal basis $\{\psi_i\}$, we have

$$(E_{\mathscr{K}}(\rho)\xi,\eta)=\sum_i(\rho(\xi\otimes\psi_i),\eta\otimes\psi_i), \tag{2.5}$$

for any ρ in $\mathscr{T}(\mathscr{H}\otimes\mathscr{K})$. It is easily seen that the adjoint of $E_{\mathscr{K}}$ is the map $a\to a\otimes 1$ from $\mathscr{L}(\mathscr{H})$ into $\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$.

The following lemmas can be verified by easy computations.

Lemma 2.1: Let $\rho\in\mathscr{T}(\mathscr{H})$, $\sigma\in\mathscr{T}(\mathscr{H}\otimes\mathscr{K})$, and $b\in\mathscr{L}(\mathscr{K})$. If we have $\mathrm{Tr}[a\rho]=\mathrm{Tr}[(a\otimes b)\sigma]$ for any $a\in\mathscr{L}(\mathscr{H})$, then we have

$$\rho=E_{\mathscr{K}}[(1\otimes b)\sigma]. \tag{2.6}$$

Lemma 2.2: Let $T:\mathscr{T}(\mathscr{H})\to\mathscr{T}(\mathscr{H})$ be a bounded linear map, and let $U\in\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$, $b\in\mathscr{L}(\mathscr{K})$, and $\sigma\in\mathscr{T}(\mathscr{K})$. Then

$$T(\rho)=E_{\mathscr{K}}[U(\rho\otimes\sigma)U^*(1\otimes b)], \tag{2.7}$$

for any ρ in $\mathscr{T}(\mathscr{H})$ if and only if

$$T^*(a)=E_\sigma[U^*(a\otimes b)U], \tag{2.8}$$

for any a in $\mathcal{L}(\mathcal{H})$.

Lemma 2.3: Let $\sigma = \Sigma_i \lambda_i |\xi_i\rangle\langle\xi_i|$ be the spectral decomposition of σ in $\mathcal{S}(\mathcal{H})$. Then

$$E_\sigma[A] = \sum_i \lambda_i E_{|\xi_i\rangle\langle\xi_i|}[A], \tag{2.9}$$

for any A in $\mathcal{L}(\mathcal{H}\otimes\mathcal{K})$, where the sum is convergent in the weak operator topology.

3. MEASURING PROCESSES

In order to determine the possible transformations of states associated with the measurement of an observable, we shall consider the description of the measuring process in terms of the interaction between the observed system and the apparatus, which is a generalization of von Neumann's description of the measuring process for an observable with discrete spectrum (Ref. 10, Chap. IV). Our mathematical formulation of the measuring process is as follows.

Definition 3.1: Let $\mathcal{H}$ be a Hilbert space and X be a semiobservable on $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. A *measuring process* M of X is a 4-tuple $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ consisting of a Hilbert space $\mathcal{K}$, an observable $\tilde{X}$ on $\mathcal{K}$ with value space $(\Omega,\mathcal{B})$, a state σ on $\mathcal{K}$, and a unitary operator U on $\mathcal{H}\otimes\mathcal{K}$ satisfying the relation

$$X(B) = E_\sigma[U^*(1\otimes\tilde{X}(B))U] \tag{3.1}$$

for any B in $\mathcal{B}$.

Now we shall explain the physical interpretation of the measuring process $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ of a semiobservable X of a Hilbert space $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. The Hilbert space $\mathcal{H}$ and $\mathcal{K}$ describe, respectively, the measured system I and the apparatus II. The semiobservable X is to be measured by this measuring process. The observable $\tilde{X}$ is to show the value of X on a scale in the apparatus which is actually measured by the observer, i.e., $\tilde{X}$ is the position of the pointer on this scale. The state σ is the initially prepared state of the apparatus. The measurement is carried out by the interaction between the observed system and the apparatus during a finite time interval from time 0 to t. The unitary operator U describes the time evolution of the composite system, i.e.,

$$U = \exp[-it(H_{\rm I}\otimes 1 + 1\otimes H_{\rm II} + H_{\rm int})], \tag{3.2}$$

where $H_{\rm I}$ and $H_{\rm II}$ are Hamiltonians of the observed system I and the apparatus II, respectively, and $H_{\rm int}$ represents the interaction. Suppose that at the instant before the interaction the measured system is in the (unknown) state ρ. Then the composite system is in the state $\rho\otimes\sigma$ at time 0 and by the interaction it is in the state $U(\rho\otimes\sigma)U^*$ at time t. Thus the probability distribution $\mathrm{Prob}(X\in B;\rho)$ of the outcomes of this measurement must coincide with the probability distribution $\mathrm{Prob}(\tilde{X}\in B;t)$ of the observable $\tilde{X}$ at time t. Since $\mathrm{Prob}(X\in B;\rho) = \mathrm{Tr}[\rho X(B)]$ and $\mathrm{Prob}(\tilde{X}\in B;t) = \mathrm{Tr}[U(\rho\otimes\sigma)U^*\tilde{X}(B)]$, we should impose the requirement

$$\mathrm{Tr}[\rho X(B)] = \mathrm{Tr}[U(\rho\otimes\sigma)U^*\tilde{X}(B)] \tag{3.3}$$

for any B in $\mathcal{B}$, ρ in $\mathcal{S}(\mathcal{H})$. It is easy to see that the requirement (3.3) is equivalent to the requirement (3.1) in Definition 3.1.

We shall now show that the measuring process $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ determines a unique state change caused by this measurement. Suppose that a measuring process $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ of X is carried out in the initial state ρ of $\mathcal{H}$. Let $B\in\mathcal{B}$. Denote by ρ^B the state, at the instant after the measurement, of the subensemble of the measured system in which the outcomes of the measurement lie in B. In order to determine the state ρ^B, suppose that the observer were to measure the simultaneously measurable observables A in I and $\tilde{X}$ in II, where A is an arbitrary bounded observable with value space $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. Then we have the joint probability distribution of their values:

$$\mathrm{Prob}(A\in d\lambda,\tilde{X}\in d\omega) = \mathrm{Tr}[U(\rho\otimes\sigma)U^*(A(d\lambda)\otimes\tilde{X}(d\omega))]. \tag{3.4}$$

Thus, if $\mathrm{Prob}(\tilde{X}\in B)\neq 0$, we have also the conditional probability distribution of A conditioned by the value of $\tilde{X}$ lying in B,

$$\begin{aligned}\mathrm{Prob}(A\in d\lambda\,|\,\tilde{X}\in B) &= \mathrm{Prob}(A\in d\lambda,\tilde{X}\in B)/\mathrm{Prob}(\tilde{X}\in B)\\ &= \mathrm{Tr}[U(\rho\otimes\sigma)U^*(A(d\lambda)\otimes\tilde{X}(B))]/\mathrm{Tr}[\rho X(B)],\end{aligned} \tag{3.5}$$

and the conditional expectation $\mathrm{Ex}(A\,|\,\tilde{X}\in B)$ of A conditioned by the value of $\tilde{X}$ lying in B,

$$\begin{aligned}\mathrm{Ex}(A\,|\,\tilde{X}\in B) &= \int_{\mathbb{R}}\lambda\,\mathrm{Prob}(A\in d\lambda\,|\,\tilde{X}\in B)\\ &= \mathrm{Tr}[U(\rho\otimes\sigma)U^*(a\otimes\tilde{X}(B))]/\mathrm{Tr}[\rho X(B)],\end{aligned} \tag{3.6}$$

where $a = \int_{\mathbb{R}}\lambda A(d\lambda)$. On the other hand, by the probabilistic interpretation of the state ρ^B, the state ρ^B must satisfy the relation

$$\mathrm{Prob}(A\in d\lambda\,|\,\tilde{X}\in B) = \mathrm{Tr}[\rho^B A(d\lambda)] \tag{3.7}$$

or, equivalently,

$$\mathrm{Ex}(A\,|\,\tilde{X}\in B) = \mathrm{Tr}[\rho^B a]. \tag{3.8}$$

By the arbitrariness of A, we can determine the state ρ^B uniquely by Eqs. (3.6) and (3.8). That is, by Lemma 2.1, we have

$$\rho^B = \{1/\mathrm{Tr}[\rho X(B)]\}E_{\mathcal{K}}[U(\rho\otimes\sigma)U^*(1\otimes\tilde{X}(B))], \tag{3.9}$$

where $E_{\mathcal{K}}:\mathcal{T}(\mathcal{H}\otimes\mathcal{K})\to\mathcal{T}(\mathcal{H})$ is the partial trace over $\mathcal{K}$. In particular, we have

$$\rho^\Omega = E_{\mathcal{K}}[U(\rho\otimes\sigma)U^*]. \tag{3.10}$$

Therefore, we have determined the state change $\rho\to\rho^B$ caused by the measuring process $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ of the semiobservable X on $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$.

Let $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ be a measuring process of a semiobservable X. For any a in $\mathcal{L}(\mathcal{H})$, $\mathrm{Ex}^M(a|B;\rho)$ will denote the conditional expectation of the outcome of a measurement of a at that instant after the measuring process M under the condition that the measuring process M of X has been carried out in the initial state ρ on $\mathcal{H}$ and its outcome lies in $B\in\mathcal{B}$. Then from the above discussions, we have

$$\begin{aligned}\mathrm{Ex}^M(a|B;\rho) &= \mathrm{Tr}[\rho^B a]\\ &= \{1/\mathrm{Tr}[\rho X(B)]\}\\ &\quad\times\mathrm{Tr}[U(\rho\otimes\sigma)U^*(a\otimes\tilde{X}(B))].\end{aligned} \tag{3.11}$$

Conclusion: Every measuring process $M = \langle \mathscr{K}, \tilde{X}, \sigma, U \rangle$ of a semiobservable X determines a state change $\rho \to \rho^B$ caused by the measurement, where ρ^B is the state, at the instant after the measurement, of the subensemble of the measured system in which outcomes of the measurement in the initial state ρ lies in $B \in \mathscr{B}$.

4. COMPLETELY POSITIVE INSTRUMENTS

From the investigations of von Neumann's repeated measurements, Davies and Lewis[1] introduced a mathematical notion of instruments which represents statistical correlations of outcomes of successive measurements. For the theory of instruments, called operational quantum probability theory, we refer the reader to Refs. 1 and 4. In the present section, we shall provide some general results on instruments imposed complete positivity.

Our setting for operational quantum probability theory consists of a von Neumann algebra $\mathscr{M}$ on a Hilbert space $\mathscr{H}$ and a Borel space $(\Omega, \mathscr{B})$. A *state* ρ of $\mathscr{M}$ is a normal state on $\mathscr{M}$. Denote by $\mathscr{M}_*$ the predual of $\mathscr{M}$ and by $\Sigma(\mathscr{M})$ the space of all normal states on $\mathscr{M}$. A *semiobservable* X in $\mathscr{M}$ is a semiobservable on $\mathscr{H}$ whose range is contained in $\mathscr{M}$. A *subtransition map* T on $\mathscr{M}$ is a normal completely positive linear map $T: \mathscr{M} \to \mathscr{M}$ such that $0 \leqslant T(1) \leqslant 1$. A *transition map* T is a subtransition map such that $T(1) = 1$. We define the right action of a subtransition map T on $\mathscr{M}_*$ by the duality

$$\langle \rho, Ta \rangle = \langle \rho T, a \rangle, \tag{4.1}$$

for all a in $\mathscr{M}$, ρ in $\mathscr{M}_*$. A *CP instrument* $\mathscr{I}$ on $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$ is a subtransition map valued measure on $(\Omega, \mathscr{B})$ such that (i) for each countable family $\{ B_i \}$ of pairwise disjoint sets in $\mathscr{B}$,

$$\left\langle \rho, \mathscr{I}\left(\bigcup_i B_i\right) a \right\rangle = \sum_i \langle \rho, \mathscr{I}(B_i) a \rangle, \tag{4.2}$$

for all a in $\mathscr{M}$, ρ in $\mathscr{M}_*$ and that (ii) $\mathscr{I}(\Omega)1 = 1$. The condition (i) is equivalent to countable additivity of the right action in the strong operator topology on $\mathscr{L}(\mathscr{M}_*, \mathscr{M}_*)$. In what follows we shall also use the notation $\mathscr{I}(\cdot,\cdot)$ for a CP instrument $\mathscr{I}$ in such a way $\mathscr{I}(B,a) = \mathscr{I}(B)a$ for all B in $\mathscr{B}$, a in $\mathscr{M}$. By the same argument as in Ref. 1, Theorem 1, we can prove the following.

Proposition 4.1: For every CP instrument $\mathscr{I}$ on $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$ there is a unique semiobservable X in $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$ such that $X(B) = \mathscr{I}(B,1)$ for all B in $\mathscr{B}$. Every semiobservable is determined in such a way by at least one CP instrument.

Let $\mathscr{I}$ be a CP instrument. We say that a semiobservable X is the *associate semiobservable* of $\mathscr{I}$, if $X(B) = \mathscr{I}(B,1)$ for any B in $\mathscr{B}$ and that a transition map T is the *associate map* of $\mathscr{I}$ if $T(a) = \mathscr{I}(\Omega, a)$ for any a in $\mathscr{M}$. Let X be a semiobservable. A CP instrument $\mathscr{I}$ is called *X-compatible* if X is the associate semiobservable of $\mathscr{I}$. A transition map T is called *X-compatible* if the range of T is contained in $X(\mathscr{B})'$.

The following proposition is very useful in dealing with CP instruments which is a modification of the Stinespring theorem on completely positive maps.[18]

Proposition 4.2: For any CP instrument $\mathscr{I}$ of $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$ there is a Hilbert space $\mathscr{H}_0$, a spectral measure $E: \mathscr{B} \to \mathscr{L}(\mathscr{H}_0)$, a nondegenerate normal*-representation $\pi: \mathscr{M} \to \mathscr{L}(\mathscr{H}_0)$ and a linear isometry $V: \mathscr{H} \to \mathscr{H}_0$ satisfying

$$\mathscr{I}(B,a) = V^* E(B) \pi(a) V, \tag{4.3}$$

$$E(B)\pi(a) = \pi(a) E(B), \tag{4.4}$$

for any B in $\mathscr{B}$ and a in $\mathscr{M}$.

Proof: Denote by $B(\Omega)$ the space of all bounded $\mathscr{B}$-measurable functions on Ω. Consider the algebraic tensor product $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H}$. We define a sesquilinear form $(\cdot,\cdot)$ on $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H}$ as follows:

$$(\xi, \eta) = \sum_i \int_\Omega g_j(\omega)^* f_i(\omega) (\mathscr{I}(d\omega, b_j^* a_i) \xi_i, \eta_j),$$

for $\xi = \Sigma_i f_i \otimes a_i \otimes \xi_i$, $\eta = \Sigma_j g_j \otimes b_j \otimes \eta_j$ in $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H}$. Then we can prove that $(\xi, \xi) \geqslant 0$ by just a similar way as the proof of Ref. 18, Theorem 4, and thus $\xi \to \|\xi\| = (\xi, \xi)^{1/2}$ is a seminorm. Define actions π of $\mathscr{M}$ and E of $\mathscr{B}$ on $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H}$ as follows:

$$\pi(x)\xi = \sum_i f_i \otimes x a_i \otimes \xi_i,$$

$$E(B)\xi = \sum_i \chi_B f_i \otimes a_i \otimes \xi_i,$$

for x in $\mathscr{M}$, B in $\mathscr{B}$, and $\xi = \Sigma_i f_i \otimes a_i \otimes \xi_i$. Then we have that $\|\pi(x)\xi\| \leqslant \|x\| \, \|\xi\|$ and $\|E(B)\xi\| \leqslant \|\xi\|$. Thus the both actions are well defined also on the $\|\cdot\|$-completion $\mathscr{H}_0$ of the quotient space $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H} / \mathscr{N}$, where $\mathscr{N} = \{ \xi \mid \|\xi\| = 0 \}$. Define a map $V: \mathscr{H} \to \mathscr{H}_0$ as $V\phi = (1 \otimes 1 \otimes \phi) + \mathscr{N}$, for any ϕ in $\mathscr{H}$. Then the assertions can be checked in a routine manner (Ref. 18 and Ref. 19, p. 194). QED

A CP instrument $\mathscr{I}$ is called *decomposable* if it is of the form $\mathscr{I}(B,a) = X(B)T(a)$ for all B in $\mathscr{B}$, a in $\mathscr{M}$, where X is the associate semiobservable of $\mathscr{I}$ and T is the associate map of $\mathscr{I}$.

Proposition 4.3: A CP instrument $\mathscr{I}$ is decomposable if its associate semiobservable X is projection-valued or if its associate map T is homomorphic [i.e., $T(a^*a) = T(a)^*T(a)$ for all a in $\mathscr{M}$].

Proof: First suppose that T is homomorphic. We can suppose that $\mathscr{I}$ is of the form $\mathscr{I}(B,a) = V^* E(B) \pi(a) V$ as in Proposition 4.2. Since $T(a) = V^* \pi(a) V$ and $V^* V = 1$, we have

$$(\pi(a)V - VT(a))^*(\pi(a)V - VT(a)) = T(a^*a) - T(a)^*T(a) = 0.$$

Thus $\pi(a)V = VT(a)$ for all a in $\mathscr{M}$, and hence we obtain that $\mathscr{I}(B,a) = V^* E(B) \pi(a) V = V^* E(B) VT(a) = X(B)T(a)$ for any B in $\mathscr{B}$, a in $\mathscr{M}$. The proof for the case that X is projection-valued is similar. QED

Proposition 4.4: Let X be an observable in $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$. Then there is a one-to-one correspondence between X-compatible CP instruments $\mathscr{I}$ on $\mathscr{M}$ and X-compatible transition maps T on $\mathscr{M}$, which is given by $\mathscr{I}(B,a) = X(B)T(a)$ for any B in $\mathscr{B}$, a in $\mathscr{M}$.

Proof: If a CP instrument $\mathscr{I}$ is decomposable, then its associate map T is X-compatible, since $X(B)T(a) = (X(B)T(a))^* = T(a)X(B)$ for any $a \geqslant 0$ in $\mathscr{M}$, B in $\mathscr{B}$.

Conversely, if T is an X-compatible transition map then it is easy to check that the relation $\mathcal{I}(B,a) = X(B)T(a)$, where $a \in \mathcal{M}$ and $B \in \mathcal{B}$, defines an X-compatible CP instrument. Thus the assertion follows immediately from Proposition 4.3. QED

5. CLASSIFICATION OF MEASURING PROCESSES

Let $\mathcal{H}$ be a Hilbert space and X be a semiobservable on $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. We say that two measuring processes M_1 and M_2 of X are *statistically equivalent* if

$$\mathrm{Ex}^{M_1}(a|B;\rho) = \mathrm{Ex}^{M_2}(a|B;\rho), \tag{5.1}$$

for any a in $\mathcal{L}(\mathcal{H})$, B in $\mathcal{B}$, ρ in $\mathcal{S}(\mathcal{H})$. Since every two statistically equivalent measuring processes give the same state change, it is desirable to classify these equivalence classes by more tractable mathematical objects concerned only with the observed system. In this section, we shall carry out such classification.

Let $M = \langle \mathcal{K},\tilde{X},\sigma,U \rangle$ be a measuring process of X. Consider the following relation:

$$\mathcal{I}(B)a = E_\sigma[U^*(a \otimes \tilde{X}(B))U], \tag{5.2}$$

for any B in $\mathcal{B}$, a in $\mathcal{L}(\mathcal{H})$. Then it is not hard to check that Eq. (5.2) defines an X-compatible CP instrument $\mathcal{I}$ on $\mathcal{L}(\mathcal{H})$. By Lemma 2.2, Eq. (5.2) is equivalent to

$$\rho\mathcal{I}(B) = E_{\mathcal{K}}[U(\rho \otimes \sigma)U^*(1 \otimes \tilde{X}(B))], \tag{5.3}$$

for all B in $\mathcal{B}$, ρ in $\mathcal{T}(\mathcal{H})$. By Eqs. (3.1) and (3.9), we have

$$X(B) = \mathcal{I}(B,1), \tag{5.4}$$

$$\rho^B = (1/\mathrm{Tr}[\rho\mathcal{I}(B)])\rho\mathcal{I}(B), \quad \text{whenever } \mathrm{Tr}[\rho X(B)] \neq 0, \tag{5.5}$$

for all ρ in $\mathcal{S}(\mathcal{H})$, B in $\mathcal{B}$. Thus the CP instrument $\mathcal{I}$ defined by Eq. (5.2) retains the all statistical data of the measuring process M, that is, the probability distribution of outcomes of the measurement and the state change caused by the measurement. The following theorem shows that every CP instrument on $\mathcal{L}(\mathcal{H})$ arises in this way.

Theorem 5.1: Let X be a semiobservable on $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. Then there is a one-to-one correspondence between statistical equivalence classes of measuring processes M of X and X-compatible CP instruments $\mathcal{I}$ on $\mathcal{L}(\mathcal{H})$, which is given by the relation

$$\mathrm{Tr}[\rho\mathcal{I}(B)]\mathrm{Ex}^M(a|B;\rho) = \mathrm{Tr}[\rho\mathcal{I}(B)a], \tag{5.6}$$

for all B in $\mathcal{B}$, ρ in $\mathcal{S}(\mathcal{H})$, a in $\mathcal{L}(\mathcal{H})$.

Proof: Let $M = \langle \mathcal{K},\tilde{X},\sigma,U \rangle$ be a measuring process of X. Then it is easy to see that the CP instrument $\mathcal{I}$ defined by Eq. (5.2) is a unique CP instrument which satisfies Eq. (5.6). It follows that the statistically equivalent measuring processes determine the same CP instrument by Eq. (5.2). Now it suffices to construct a measuring process of X which determines by Eq. (5.2) a given X-compatible CP instrument. Let $\mathcal{I}$ be an X-compatible CP instrument on $\mathcal{L}(\mathcal{H})$ with value space $(\Omega,\mathcal{B})$. Let $\mathcal{H}_0$, E, π, and V be such as obtained in Proposition 4.2 for the CP instrument $\mathcal{I}$. Since every nondegenerate normal $*$-representation of $\mathcal{L}(\mathcal{H})$ is unitarily equivalent to the multiple of the identity representation (Ref. 4, Lemma 9.2.2), there is a Hilbert space $\mathcal{H}_1$ such that $\mathcal{H}_0 = \mathcal{H} \otimes \mathcal{H}_1$ and that $\pi(a) = a \otimes 1$ for any a in $\mathcal{L}(\mathcal{H})$. Then by Eq. (4.3) and by the commutation theorem of von Neumann algebras, for any B in $\mathcal{B}$ there is a projection $E_1(B)$ in $\mathcal{L}(\mathcal{H}_1)$ such that $E(B) = 1 \otimes E_1(B)$. Obviously, the correspondence $E_1: B \to E_1(B)$ is a projection-valued measure from $\mathcal{B}$ to $\mathcal{L}(\mathcal{H}_1)$. By Eq. (4.3), we have

$$\mathcal{I}(B,a) = V^*(a \otimes E_1(B))V,$$

for any B in $\mathcal{B}$, a in $\mathcal{L}(\mathcal{H})$. Let η_0 be a unit vector in $\mathcal{H}_0$ and η_1 be a unit vector in $\mathcal{H}_1$. Define an isometry V_0 on $\mathcal{H} \otimes [\eta_1] \otimes [\eta_0]$ into $\mathcal{H} \otimes \mathcal{H}_1 \otimes \mathcal{H}_0$ by the relation

$$V_0(\xi \otimes \eta_1 \otimes \eta_0) = V\xi \otimes \eta_0,$$

for any ξ in $\mathcal{H}$. Then, since $\dim(\mathcal{H}_0) = \dim(\mathcal{H} \otimes \mathcal{H}_1)$, by the usual computations of cardinal numbers, it is easy to show that

$$\dim(\mathcal{H} \otimes \mathcal{H}_1 \otimes \mathcal{H}_0 - \mathcal{H} \otimes [\eta_1] \otimes [\eta_0]) = \dim(\mathcal{H} \otimes \mathcal{H}_1 \otimes \mathcal{H}_0 - V_0(\mathcal{H} \otimes [\eta_1] \otimes [\eta_0])).$$

It follows that there is a unitary operator U on $\mathcal{H} \otimes \mathcal{H}_1 \otimes \mathcal{H}_0$ which is an extension of V_0. Now let $\mathcal{K}$, σ, and $\tilde{X}$ be such that

$$\mathcal{K} = \mathcal{H}_1 \otimes \mathcal{H}_0, \quad \sigma = |\eta_1 \otimes \eta_0\rangle\langle\eta_1 \otimes \eta_0|,$$
$$\text{and} \quad \tilde{X}(B) = E_1(B) \otimes 1 \quad \text{on } \mathcal{H}_1 \otimes \mathcal{H}_0,$$

for any B in $\mathcal{B}$. Then we shall claim that $\langle \mathcal{K},\tilde{X},\sigma,U \rangle$ is a measuring process which determines the CP instrument $\mathcal{I}$ by Eqs. (5.2). For any a in $\mathcal{L}(\mathcal{H})$, ξ in $\mathcal{H}$, and B in $\mathcal{B}$, we have that

$$\begin{aligned}
(\mathcal{I}(B,a)\xi,\xi) &= (V^*(a \otimes E_1(B))V\xi,\xi) \\
&= ((a \otimes E_1(B))V\xi, V\xi) \\
&= ((a \otimes E_1(B))V\xi \otimes \eta_0, V\xi \otimes \eta_0) \\
&= ((a \otimes E_1(B) \otimes 1)U(\xi \otimes \eta_1 \otimes \eta_0), U(\xi \otimes \eta_1 \otimes \eta_0)) \\
&= (U^*(a \otimes \tilde{X}(B))U(\xi \otimes \eta_1 \otimes \eta_0), \xi \otimes \eta_1 \otimes \eta_0) \\
&= \mathrm{Tr}[U^*(a \otimes \tilde{X}(B))U(|\xi\rangle\langle\xi| \otimes \sigma)] \\
&= \mathrm{Tr}[|\xi\rangle\langle\xi|E_\sigma[U^*(a \otimes \tilde{X}(B))U]] \\
&= (E_\sigma[U^*(a \otimes \tilde{X}(B))U]\xi,\xi).
\end{aligned}$$

It follows that

$$\mathcal{I}(B,a) = E_\sigma[U^*(a \otimes \tilde{X}(B))U],$$

for any a in $\mathcal{L}(\mathcal{H})$ and B in $\mathcal{B}$. Therefore, $\langle \mathcal{K},\tilde{X},\sigma,U \rangle$ is a measuring process of X which determines $\mathcal{I}$ by Eq. (5.2). QED

We say that a measuring process M is a *realization* of a CP instrument $\mathcal{I}$ if M and $\mathcal{I}$ satisfies Eq. (5.6). The above theorem asserts that every CP instrument has its realization. In the conventional theory of quantum mechanics, it is always assumed that the Hilbert space is separable and the value space is a standard Borel space, i.e., a Borel space which is Borel isomorphic to a separable complete metric space.[20] Thus it is desirable that the realization is also with a separable Hilbert space in such circumstances. We say that realization $M = \langle \mathcal{K},\tilde{X},\sigma,U \rangle$ of a CP instrument $\mathcal{I}$ is *separable* if the Hilbert space $\mathcal{K}$ is separable.

Corollary 5.2: Let $\mathcal{I}$ be a CP instrument on $\mathcal{L}(\mathcal{H})$ with value space $(\Omega,\mathcal{B})$. If $\mathcal{H}$ is separable and $(\Omega,\mathcal{B})$ is a standard Borel space, then there is a separable realization of $\mathcal{I}$.

Proof (the notations are the same as in the proof of Theorem 5.1): It is easy to see that we can assume that $\mathscr{H}_0$ in Proposition 4.2 is spanned by $\{E(B)\pi(a)V\xi;\ B\in\mathscr{B},\ a\in\mathscr{L}(\mathscr{H})$, and $\xi\in\mathscr{H}\}$. Since $\mathscr{H}$ is separable, there is a countable family $\{a_n\}$ of a_n in $\mathscr{L}(\mathscr{H})$ which is dense in $\mathscr{L}(\mathscr{H})$ in the strong operator topology. Let $\{B_n\}$ be a countable generator of $\mathscr{B}$ and $\{\xi_n\}$ be a countable dense subset of $\mathscr{H}$. Then it is easy to see that the countable family $\{E(B_i)\times\pi(a_j)V\xi_k; i,j,k = 1,2,\cdots\}$ spans $\mathscr{H}_0$, so that $\mathscr{H}_0$ is separable. Since $\mathscr{H}\otimes\mathscr{K}=\mathscr{H}_0\otimes\mathscr{H}_0$, $\mathscr{K}$ is separable. QED

We say that a measuring process $M=\langle\mathscr{K},\tilde{X},\sigma,U\rangle$ is *pure* if σ is pure state, i.e., there is a unit vector ξ in $\mathscr{K}$ such that $\sigma=|\xi\rangle\langle\xi|$. In the conventional argument of quantum measurement, the assumption that the prepared state of the apparatus is pure has been justified in some contexts. The following is one of such justification from a most general point of view.

Corollary 5.3: Every measuring process is statistically equivalent to a pure measuring process.

Proof: The assertion is immediate from the construction of the measuring process in the proof of Theorem 5.1. QED

Let $M=\langle\mathscr{K},\tilde{X},|\eta\rangle\langle\eta|,U\rangle$ be a pure measuring process. Define an isometry $V:\mathscr{H}\to\mathscr{H}\otimes\mathscr{K}$ by $V\xi=U(\xi\otimes\eta)$ for all ξ in $\mathscr{H}$. Let $\mathscr{I}$ be the corresponding CP instrument. Then it is easy to see that

$$\mathscr{I}(B,a)=E_\sigma[U^*(a\otimes\tilde{X}(B))U]=V^*(a\otimes\tilde{X}(B))V,$$

for all a in $\mathscr{L}(\mathscr{H})$, B in $\mathscr{B}$.

The following result justifies our postulate, which is tacit in Eq. (2.2), that *semiobservables can be measured.*

Corollary 5.4: For any semiobservable X, there is a measuring process of X.

Proof: By proposition 4.1, for any semiobservable X, there is an X-compatible CP instrument $\mathscr{I}$. Then any realization of $\mathscr{I}$ obtained by Theorem 5.1 is a measuring process of X. QED

Consider the case that X is an observable. In this case the classification of measuring processes is surprisingly simpler, that is, the measuring processes of X are determined by their total state changes $\rho\to\rho^\Omega$.

Theorem 5.5: Let X be an observable on $\mathscr{H}$ with value space $(\Omega,\mathscr{B})$. Then there is a one-to-one correspondence between statistical equivalence classes of measuring processes M of X and X-compatible transition maps T on $\mathscr{L}(\mathscr{H})$, which is given by the relation

$$\mathrm{Tr}[\rho X(B)]\mathrm{Ex}^M(a|B;\rho)=\mathrm{Tr}[\rho X(B)T(a)],\qquad(5.7)$$

for any a in $\mathscr{L}(\mathscr{H})$, ρ in $\Sigma(\mathscr{H})$, B in $\mathscr{B}$.

Proof: The assertion follows immediately from Proposition 4.4 and Theorem 5.1. QED

6. REPEATABILITY

Consider von Neumann's repeatability hypothesis (Ref. 10, pp. 214, 335):

(M) If the physical quantity is measured twice in succession in a system, then we get the same value each time.

Let $M=\langle\mathscr{K},\tilde{X},\sigma,U\rangle$ be a measuring process of a semiobservable X. If X is discrete, then it is easy to see that (M) is equivalent to

(M′) $\mathrm{Ex}^M(X(\{\lambda\})|\{\mu\};\rho)=\delta_{\lambda,\mu}$

for all ρ in $\Sigma(\mathscr{H})$ and all λ,μ in Ω, whenever $\mathrm{Tr}[\rho X(\{\mu\})]\neq 0$. We say that a measuring process M of X is *weakly repeatable* if

(R) $\mathrm{Ex}^M(X(C)|B;\rho)=\mathrm{Tr}[\rho X(B\cap C)]/\mathrm{Tr}[\rho X(B)]$,

for any ρ in $\Sigma(\mathscr{H})$, B,C in $\mathscr{B}$, whenever $\mathrm{Tr}[\rho X(B)]\neq 0$. Then it is easy to see that if X is discrete the condition (M′) and (R) are equivalent. The condition (R) appeared first in Ref. 1 for instruments. We say that a CP instrument $\mathscr{I}$ is *weakly repeatable* if $\mathscr{I}(B)X(C)=X(B\cap C)$ for all B,C in $\mathscr{B}$, where X is the associate semiobservable of $\mathscr{I}$. It is easily seen that a measuring process M is weakly repeatable if and only if the corresponding CP instrument $\mathscr{I}$ is weakly repeatable. In Ref. 1, p. 247, it is conjectured that the existence of repeatable instruments for continuous observables is doubtful even in the case of standard quantum theory. In the present section, we shall prove this conjecture, that is, we shall prove that there is at least one X-compatible weakly repeatable CP instrument on $\mathscr{L}(\mathscr{H})$ if and only if X is discrete.

Let $\mathscr{M}$ be a von Neumann algebra on $\mathscr{H}$ and $(\Omega,\mathscr{B})$ be a Borel space. Let $\mathscr{I}$ be a weakly repeatable CP instrument on $\mathscr{M}$ with value space $(\Omega,\mathscr{B})$, X its associate semiobservable, and T its associate map. We can assume that $\mathscr{I}$ is of the form $\mathscr{I}(B,a)=V^*E(B)\pi(a)V$ for any B in $\mathscr{B}$, a in $\mathscr{M}$, as in Proposition 4.2.

Lemma 6.1: For any B,C in $\mathscr{B}$, a in $\mathscr{M}$, we have

(1) $T(X(B)^2)=X(B)$,

(2) $\mathscr{I}(B\cap C,a)=\mathscr{I}(C,aX(B))=\mathscr{I}(C,X(B)a)$,

(3) $\mathscr{I}(B,a)=T(aX(B))=T(X(B)a)$.

Proof: Since $\mathscr{I}(B,X(B))=X(B)$ by the weak repeatability of $\mathscr{I}$, a routine computation leads that

$$(\pi(X(B))V-E(B)V)^*(\pi(X(B))V-E(B)V)=T(X(B)^2)-X(B),\qquad(6.1)$$

for any B in $\mathscr{B}$. Thus we have $T(X(B)^2)\geqslant X(B)$. On the other hand, we have $X(B)^2\leqslant X(B)$, since $0\leqslant X(B)\leqslant 1$. By weak repeatability, $T(X(B))=X(B)$, so that $X(B)=T(X(B))\geqslant T(X(B)^2)$. Thus we have the relation (1). It follows that the left-hand side of Eq. (6.1) is 0, so that we have $\pi(X(B))V=E(B)V$ and $V^*\pi(X(B))=V^*E(B)$. Thus for any B,C in $\mathscr{B}$, a in $\mathscr{M}$, we have $\mathscr{I}(B\cap C,a)=V^*E(B\cap C)\pi(a)V=V^*E(B)E(C)\pi(a)V=V^*\pi(X(B))E(C)\pi(a)V=V^*E(C)\pi(X(B)a)V=\mathscr{I}(C,X(B)a)$. By the analogous way we can show that $\mathscr{I}(B\cap C,a)=\mathscr{I}(C,aX(B))$. Thus we obtain the relation (2). The relation (3) is obtained by putting $C=\Omega$ in (2). QED

Let p be the least projection in $X(\mathscr{B})''$ such that $T(p)=1$.

Lemma 6.2: For any x in $\mathscr{M}$, $T(x)=T(xp)=T(px)=T(pxp)$.

Proof: For any ξ,η in $\mathscr{H}$, we have

$$\begin{aligned}|(T(x-px)\xi,\eta)| &= |(V^*\pi(1-p)\pi(x)V\xi,\eta)|\\ &= |(\pi(x)V\xi,\pi(1-p)V\eta)|\leqslant\|\pi(x)V\xi\|\ \|\pi(1-p)V\eta\|\\ &= \|\pi(x)V\xi\|\ \|(V^*\pi(1-p)V\eta,\eta)\|^{1/2}\\ &= \|\pi(x)V\xi\|(T(1-p)\eta,\eta)^{1/2}=0.\end{aligned}$$

Thus we have $T(x)=T(px)$. The rest of the assertions are immediate. QED

Lemma 6.3: For every x in $X(\mathscr{B})''$ with $x \geqslant 0$, if $T(x) = 0$, then $pxp = 0$.

Proof: Let e be the range projection of x. Since e is a limit of polynomials of x not containing the constant term in the strong operator topology, we have $T(e) = 0$. Thus $1 - e \geqslant p$ so that $ep = pe = 0$. It follows that $pxp = 0$. QED

Define a positive operator valued measure P: $\mathscr{B} \to X(\mathscr{B})''$ by the relation $P(B) = pX(B)p$ for all B in $\mathscr{B}$.

Lemma 6.4: P is a projection valued measure such that $P(B) = pX(B) = X(B)p$ for any B in $\mathscr{B}$.

Proof: By Lemma 6.2, we have $T(P(B)) = T(pX(B)p) = T(X(B))$. By Lemmas 6.1 and 6.2, we have $T(P(B)^2) = T(pX(B)pX(B)p) = T(X(B)pX(B)) = \mathscr{I}(B,pX(B)) = \mathscr{I}(B \cap B,p) = \mathscr{I}(B,p) = T(pX(B)) = T(X(B))$. It follows that $T(P(B) - P(B)^2) = 0$. Since $P(B) - P(B)^2$ belongs to $X(\mathscr{B})''$, we have $P(B)^2 = P(B)$ by Lemma 6.3. Thus P is a projection-valued measure. We have $T((P(B) - X(B)p)^*(P(B) - X(B)p)) = 0$, by the routine computations. Thus, by Lemma 6.3, $P(B) = X(B)p$, since $P(B) - X(B)p$ is in $X(\mathscr{B})''$. By the positivity, we have $P(B) = pX(B)$. QED

Theorem 6.5: For any weakly repeatable CP instrument $\mathscr{I}$ on $\mathscr{M}$ with value space $(\Omega,\mathscr{B})$, there is a projection-valued measure P: $\mathscr{B} \to X(\mathscr{B})''$ such that

$$\mathscr{I}(B,a) = T(aP(B)) = T(P(B)a)$$

and that

$$P(B) = P(\Omega)X(B) = X(B)P(\Omega),$$

for any B in $\mathscr{B}$, a in $\mathscr{L}(\mathscr{H})$.

Proof: The assertion follows immediately from Lemmas 6.1 and 6.4. QED

We suppose for the rest of this section that the value space $(\Omega,\mathscr{B})$ is a standard Borel space and that the Hilbert space $\mathscr{H}$ is separable. We say that a positive operator valued measure P is *discrete* if there is a countable set $\Omega_0 \subseteq \Omega$ such that $P(\Omega \setminus \Omega_0) = 0$ and that a CP instrument is *discrete* if the associate semiobservable is discrete.

Theorem 6.6: Let $(\Omega,\mathscr{B})$ be a standard Borel space, and let $\mathscr{H}$ be a separable Hilbert space. Then every weakly repeatable CP instrument $\mathscr{I}$ on $\mathscr{L}(\mathscr{H})$ with value space $(\Omega,\mathscr{B})$ is discrete.

Proof: Let P be a projection-valued measure obtained in Theorem 6.5. By the relation $X(B) = \mathscr{I}(B,1) = T(P(B))$ for every B in $\mathscr{B}$, we have only to show that P is discrete. By Ref. 4, Lemma 4.4.1, there is a countable set B_0 such that $B \to P(B \cap B_0)$ is a discrete projection-valued measure with values in $\mathscr{L}(P(B_0)\mathscr{H})$ and $B \to P(B \setminus B_0)$ is a continuous projection-valued measure with values in $\mathscr{L}(P(\Omega \setminus B_0)\mathscr{H})$. Let Q be such that $Q = P(\Omega \setminus B_0)$. Then it suffices to prove that $Q = 0$. Let T be the associate map of $\mathscr{I}$ and T_0 be such that $T_0(a) = QT(a)Q$ for all a in $\mathscr{L}(Q\mathscr{H})$. Then $T_0(Q) = QT(Q)Q = QT(X(\Omega \setminus B_0))Q = QX(\Omega \setminus B_0)Q = Q$, and hence T_0 is a transition map on $\mathscr{L}(Q\mathscr{H})$. Thus there is a trace-preserving linear map S: $\mathscr{T}(Q\mathscr{H}) \to \mathscr{T}(Q\mathscr{H})$ such that $S^* = T_0$. For any a in $\mathscr{L}(Q\mathscr{H})$, B in $\mathscr{B}$, ρ in $\mathscr{T}(Q\mathscr{H})$, we have

$$\begin{aligned}\mathrm{Tr}[aP(B \setminus B_0)S(\rho)] &= \mathrm{Tr}[T_0(aP(B \setminus B_0))\rho] \\ &= \mathrm{Tr}[QT(aP(B \setminus B_0))Q\rho] = \mathrm{Tr}[QT(P(B \setminus B_0)a)Q\rho] \\ &= \mathrm{Tr}[T_0(P(B \setminus B_0)a)\rho] = \mathrm{Tr}[P(B \setminus B_0)aS(\rho)].\end{aligned}$$

It follows that $P(B \setminus B_0)S(\rho) = S(\rho)P(B \setminus B_0)$ for any B in $\mathscr{B}$, ρ in $\mathscr{T}(Q\mathscr{H})$. Since $B \to P(B \setminus B_0)$ is a continuous projection-valued measure, we can conclude that $S = 0$ (see, Ref. 4, Theorem 4.3.3), and hence $Q = T_0(Q) = 0$. QED

7. LOCALITY

Let $\mathscr{H}$ be a Hilbert space and $\mathscr{M}$ a von Neumann algebra on $\mathscr{H}$. Let X be an observable in $\mathscr{M}$ with value space $(\Omega,\mathscr{B})$. A transition map T on $\mathscr{M}$ is called *X-local* if $T(X(B)) = X(B)$ for any B in $\mathscr{B}$. It is easy to see that T is X-local if and only if $Tx = x$ for any x in $X(\mathscr{B})''$.

Let $\{x_1,\ldots,x_n\}$ be a mutually commutable family of self-adjoint operators on $\mathscr{H}$ corresponding to a family of simultaneously measurable observables of a quantum system. Suppose that X is the joint spectral measure of $\{x_1,\ldots,x_n\}$ on $\mathscr{H}$ with value space $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$. Recently, Mercer[9] proposed that the total state change $\rho \to \rho'$ caused by a simultaneous measurement of $x_1,\ldots,x_n$ should be described by an X-local transition map T on $\mathscr{L}(\mathscr{H})$ in such a way $\rho' = \rho T$ (see Ref. 9, p. 244). However, we should notice that the X-locality is not sufficient for describing state transformations caused by measurements. In fact, the identity transformation on $\mathscr{L}(\mathscr{H})$ is obviously an X-local transition map for any observable X, in spite of the fact that we cannot measure any nontrivial quantum observable unchanging every state of the system. Thus we have to impose some further requirements for eliminating such physically irrelevant X-local transition maps in order to describe a state change caused by the measurement of X. A moderate one of such requirements seems the existence of a measuring process for observables $x_1,\ldots,x_n$, whose state change is the given X-local transition map. The following result is an easy consequence of the results obtained in the previous sections, but shows that such requirement cannot be fulfilled unless all observables $x_1,\ldots,x_n$ are discrete.

Proposition 7.1: Let $\mathscr{M}$ be a von Neumann algebra on a Hilbert space $\mathscr{H}$ and X an observable in $\mathscr{M}$ with value space $(\Omega,\mathscr{B})$. There is a one-to-one correspondence between X-compatible X-local transition maps T on $\mathscr{M}$ and X-compatible weakly repeatable CP-instruments $\mathscr{I}$ on $\mathscr{M}$, which is given by

$$\mathscr{I}(B,a) = X(B)T(a), \tag{7.1}$$

for any B in $\mathscr{B}$, a in $\mathscr{M}$.

Proof: It is known in the proof of Ref. 1, Theorem 7, that a decomposable CP instrument $\mathscr{I}$ is weakly repeatable if and only if

$$T(X(B)) = X(B) \quad \text{and} \quad X(B \cap C) = X(B)X(C),$$

for any B,C in $\mathscr{B}$. Since every X-compatible CP instrument is decomposable, the assertion follows immediately from Proposition 4.4 QED

Theorem 7.2: Let X be an observable on a separable Hilbert space $\mathscr{H}$ whose value space is a standard Borel space and T be an X-local transition map on $\mathscr{L}(\mathscr{H})$. If there is a measuring process $M = \langle \mathscr{K},\tilde{X},\sigma,U \rangle$ of X such that $\rho^{\Omega} = \rho T$ for any ρ in $\mathscr{S}(\mathscr{H})$ [see Eq. (3.10)], then X is discrete.

Proof: It is obvious that T is the associate map of the CP instrument $\mathscr{I}$ determined by the measuring process M.

Thus, by Proposition 7.1, the CP instrument $\mathcal{I}$ is weakly repeatable and hence by Theorem 6.6 the corresponding observable X is discrete. QED

8. THE WIGNER–ARAKI–YANASE THEOREM

It was pointed out by Wigner[15] that the presence of a conservation law puts a limitation of the measurement of an operator which does not commute with the observed quantity. A proof of the above assertion was given by Araki and Yanase[16] in the conventional framework of measurement theory. In this section, we shall give another proof in our framework and under somewhat general assumptions. Our assertion is the following.

Theorem 8.1: Let X be an observable on a Hilbert space $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. Let $M=\langle\mathcal{K},\tilde{X},\sigma,U\rangle$ be a weakly repeatable measuring process of X. Suppose that there is L_1 in $\mathcal{L}(\mathcal{H})$ and L_2 in $\mathcal{L}(\mathcal{K})$ such that $[U,L]=0$, where $L=L_1\otimes 1+1\otimes L_2$. Then $L_1\in X(\mathcal{B})'$.

For the proof we use the following.

Lemma 8.2. Let $M=\langle\mathcal{K},\tilde{X},\sigma,U\rangle$ be a measuring process of an observable X on $\mathcal{H}$, and $\sigma=\Sigma_i\lambda_i|\eta_i\rangle\langle\eta_i|$ be the spectral decomposition of σ. Then for any i, $M_i=\langle\mathcal{K},\tilde{X},|\eta_i\rangle\langle\eta_i|,U\rangle$ is a pure measuring process of X such that

$$E_\sigma[U^*AU]=\sum_i\lambda_i E_{|\eta_i\rangle\langle\eta_i|}[U^*AU], \tag{8.1}$$

for any A in $\mathcal{L}(\mathcal{H}\otimes\mathcal{K})$. If M is weakly repeatable, then M_i is also weakly repeatable for every i.

Proof: The formula (8.1) is obtained from Lemma 2.3. Let $B\in\mathcal{B}$. Then

$$\begin{aligned}X(B)&=E_\sigma[U^*(1\otimes\tilde{X}(B))U]\\&=\sum_i\lambda_i E_{|\eta_i\rangle\langle\eta_i|}[U^*(1\otimes\tilde{X}(B))U].\end{aligned}$$

Since any projection is an extreme point of the positive part of the unit sphere of $\mathcal{L}(\mathcal{H})$, we have that

$$X(B)=E_{|\eta_i\rangle\langle\eta_i|}[U^*(1\otimes\tilde{X}(B))U],$$

for any i. Thus M_i is a measuring process of X. Since M_i is weakly repeatable if $X(B)=E_{|\eta_i\rangle\langle\eta_i|}[U^*(X(B)\otimes 1)U]$ for any B in $\mathcal{B}$ by Proposition 7.1, the assertion for the weak repeatability follows from the same reasoning. QED

Proof of Theorem 8.1: By Theorem 5.5, there is an X compatible transition map T such that $E_\sigma[U^*(a\otimes\tilde{X}(B))U]=X(B)T(a)$ for any B in $\mathcal{B}$, a in $\mathcal{L}(\mathcal{H})$. Then we have

$$\begin{aligned}&T(L_1)+E_\sigma[U^*(1\otimes L_2)U]\\&\quad=E_\sigma[U^*(L_1\otimes 1+1\otimes L_2)U]\\&\quad=E_\sigma[L_1\otimes 1+1\otimes L_2]\\&\quad=L_1+[\mathrm{Tr}(\sigma L_2)]1.\end{aligned}$$

Since T is X-compatible, $T(L_1)\in X(\mathcal{B})'$. Thus we have only to show that $E_\sigma[U^*(1\otimes L_2)U]\in X(\mathcal{B})'$. By Lemma 8.2, we can assume without any loss of generality that there is a unit vector η in $\mathcal{K}$ such that $\sigma=|\eta\rangle\langle\eta|$, so that there is an isometry $V:\mathcal{H}\to\mathcal{H}\otimes\mathcal{K}$ such that $E_\sigma[U^*AU]=V^*AV$ for all A in $\mathcal{L}(\mathcal{H}\otimes\mathcal{K})$, where $V\xi=U(\xi\otimes\eta)$ for any ξ in $\mathcal{H}$. Let $B\in\mathcal{B}$. Since the CP instrument $\mathcal{I}$ such that $\mathcal{I}(B,a)=V^*(a\otimes\tilde{X}(B))V$ is weakly repeatable, we have $V^*(X(B)\otimes 1)V=\mathcal{I}(\Omega,X(B))=X(B)$. Thus by the simple computations we have

$$((X(B)\otimes 1)V-VX(B))^*((X(B)\otimes 1)V-VX(B))=0,$$

and hence $(X(B)\otimes 1)V=VX(B)$ and $V^*(X(B)\otimes 1)=X(B)V^*$. It follows that

$$V^*(1\otimes L_2)VX(B)=V^*(X(B)\otimes L_2)V=X(B)V^*(1\otimes L_2)V.$$

Thus we conclude that $E_\sigma[U^*(1\otimes L_2)U]\in X(\mathcal{B})'$. QED

9. CONVENTIONAL MEASURING PROCESSES

In the conventional theory of quantum measurement, the only class of measuring processes studied at all seriously is as follows. Let $\mathcal{H}$ be a separable Hilbert space and X be a discrete observable on $\mathcal{H}$ with value space $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. Let $\{\xi_{ij}\}$ be a complete orthonormal set of eigenvectors of X where the eigenvalue of ξ_{ij} is λ_i. Let $\mathcal{K}$ be another separable Hilbert space with complete orthonormal vectors $\{\eta_i\}$. Let η be a unit vector of $\mathcal{K}$ and U be a unitary operator on $\mathcal{H}\otimes\mathcal{K}$ satisfying

$$U(\xi_{ij}\otimes\eta)=\xi_{ij}\otimes\eta_i \tag{9.1}$$

for any i,j. Then $M=\langle\mathcal{K},\tilde{X},|\eta\rangle\langle\eta|,U\rangle$ is a measuring process of X, where $\tilde{X}=\Sigma_i\lambda_i|\eta_i\rangle\langle\eta_i|$. In the sequel, we call this measuring process a *conventional* measuring process of X. The total state change corresponding to M is of the form

$$\rho\to\rho'=\sum_i P_i\rho P_i, \tag{9.2}$$

where $P_i=X(\{\lambda_i\})$, i.e., $P_i=\Sigma_j|\xi_{ij}\rangle\langle\xi_{ij}|$. In fact, for $\rho=\Sigma_{ijkl}\mu_{ijkl}|\xi_{ij}\rangle\langle\xi_{kl}|$ in $\mathcal{L}(\mathcal{H})$, we have

$$\begin{aligned}\rho'&=E_{\mathcal{K}}[U(\rho\otimes|\eta\rangle\langle\eta|)U^*]\\&=\sum_{ijkl}\mu_{ijkl}E_{\mathcal{K}}[|\xi_{ij}\otimes\eta_i\rangle\langle\xi_{kl}\otimes\eta_k|]\\&=\sum_{ijkl}\mu_{ijkl}(\eta_i,\eta_k)|\xi_{ij}\rangle\langle\xi_{kl}|\\&=\sum_i P_i\rho P_i\end{aligned}$$

[see Eq. (3.10)]. Conversely, every state change given by Eq. (9.2) is realized as the above measuring process M as shown by von Neumann (see Ref. 10, p. 442). By Eq. (9.2) the corresponding CP instrument $\mathcal{I}$ is of the form

$$\mathcal{I}(B,a)=\sum_{\lambda_i\in B}P_iaP_i, \tag{9.3}$$

for any B in $\mathcal{B}(\mathbb{R})$, a in $\mathcal{L}(\mathcal{H})$, and the corresponding transition map T is a conditional expectation onto $X(\mathcal{B}(\mathbb{R}))'$.

In the present section, we shall give a characterization of the above conventional measuring processes up to statistical equivalence. A similar problem is considered in Refs. 1 and 21 in different methods.

Definition 9.1: Let X be a semiobservable on $\mathcal{H}$ with value space $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. A measuring process M of X is called *standard* if it satisfies the following three conditions.

(WR) (Weak repeatability) M is weakly repeatable.

(MD) (Minimal disturbance condition) The set $\{\rho\in\mathcal{S}(\mathcal{H});\rho^{\mathrm{R}}\neq\rho\}$ is minimal in the set inclusion among all measuring process of X.

(ND) (Nondegeneracy condition) For any B in $\mathscr{B}(\mathbb{R})$ with $X(B)\neq 0$, there is some ρ in $\mathfrak{S}(\mathscr{H})$ such that $\mathrm{Tr}[\rho^{R}X(B)]\neq 0$.

Let M be a measuring process of X. Denote by $F(M)$ the set of all nondisturbed states, i.e., $F(M)=\{\rho\in\sigma(\mathscr{H});\rho^{R}=\rho\}$. Obviously, M satisfies (MD) if and only if for any measuring process M' of X, $F(M)\subseteq F(M')$ implies $F(M')\subseteq F(M)$.

Proposition 9.2: Let M be a conventional measuring process of a discrete observable X. Then M is standard.

Proof: It is well known that M is weakly repeatable. The condition (ND) is easy to check. Thus we shall prove that M satisfies the condition (MD). Let M' be a measuring process of X such that $F(M)\subseteq F(M')$. Denote by T and S the transition maps corresponding to M and M', respectively. Let $\rho\in\mathscr{T}(\mathscr{H})$ be such that $\rho S=\rho$. Then it suffices to show that $\rho T=\rho$. Since T is a conditional expectation onto $X(\mathscr{B}(\mathbf{R}))'$ and by the X-compatibility of S the range of S is contained in $X(\mathscr{B}(\mathbb{R}))'$, we have $TS=S$. Since $T^{2}=T$, we have $(\rho T)T=\rho T$ so that $\rho T\in F(M)$. Thus by the assumption that $F(M)\subseteq F(M')$, $\rho T\in F(M')$. It follows that $\rho T=\rho TS=\rho S=\rho$. This concludes the proof. QED

Theorem 9.3: Let $\mathscr{H}$ be a separable Hilbert space and X be a semiobservable on $\mathscr{H}$ with value space $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. Let M be a standard measuring process of X. Then X is a discrete observable, and M is statistically equivalent to a conventional measuring process of X.

Proof: Let $\mathscr{I}$ be the CP instrument corresponding to a standard measuring process M of X. Since $\mathscr{I}$ is weakly repeatable, by Theorem 6.6, X is discrete and, by Theorem 6.5, there is an orthogonal family $\{P_{\lambda};\lambda\in R\}$ of projections in $X(\mathscr{B}(R))''$ such that

$$\mathscr{I}(B,a)=T\Big(\sum_{\lambda\in B}P_{\lambda}aP_{\lambda}\Big),\tag{9.4}$$

for all B in $\mathscr{B}$, a in $\mathscr{L}(\mathscr{H})$. Let Q be a projection in $X(\mathscr{B}(\mathbb{R}))''$ such that $Q=1-\Sigma_{\lambda\in R}P_{\lambda}$. Then we have $T(Q)=0$. It follows from the condition (ND) that $Q=0$ so that $\Sigma_{\lambda\in R}P_{\lambda}=1$. Thus by Lemma 6.4 we have $X(B)=\Sigma_{\lambda\in B}P_{\lambda}$ for any B in $\mathscr{B}(\mathbb{R})$. It follows that X is an observable. Let M' be a conventional measuring process of X and $\mathscr{I}'$ be the corresponding CP instrument. Then

$$\mathscr{I}'(B,a)=\sum_{\lambda\in B}P_{\lambda}aP_{\lambda},\tag{9.5}$$

for any B in $\mathscr{B}$, a in $\mathscr{L}(\mathscr{H})$. Denote by T and S the corresponding transition maps of M and M', respectively. Since T is X-compatible, we have $T(a)=\Sigma_{\lambda\in R}P_{\lambda}T(a)=\Sigma_{\lambda\in R}P_{\lambda}T(a)P_{\lambda}=ST(a)$, for any a in $\mathscr{L}(\mathscr{H})$. On the other hand, by Eq. (9.4) we have $T=TS$. It follows that $T=ST=TS$. For any ρ in $\mathscr{T}(\mathscr{H})$, if $\rho T=\rho$, then $\rho S=\rho TS=\rho T=\rho$ and hence $F(M)\subseteq F(M')$. Thus by the condition (MD), $F(M')=F(M)$. Let $\rho\in\mathscr{T}(\mathscr{H})$. Then since $S^{2}=S,\rho S\in F(M')$, so that $\rho ST=\rho S$. It follows that $S=ST$. Thus we have $T=S$. Therefore, by Theorem 5.5, M is statistically equivalent to a conventional measuring process M' of X. QED

ACKNOWLEDGMENTS

The author wishes to thank Professor H. Umegaki for his useful comments and encouragement. He is also indebted to Professor M. M. Yanase and Professor H. Araki for the reading of the manuscript and enlightening comments, and he is grateful to Professor A. S. Holevo and Professor N. N. Čencov for the stimulating discussions.

[1]E. B. Davies and J. T. Lewis, "An operational approach to quantum probability," Commun. Math. Phys. **17**, 239–260 (1970).
[2]E. B. Davies, "Quantum stochastic processes," Commun. Math. Phys. **15**, 277–304 (1969).
[3]E. B. Davies, "On the repeated measurement of continuous observables in quantum mechanics," J. Funct. Anal. **6**, 318–346 (1970).
[4]E. B. Davies, *Quantum Theory of Open Systems* (Academic, London, 1976).
[5]A. S. Holevo, "Statistical decision theory for quantum systems," J. Multivar. Anal. **3**, 337–394 (1973).
[6]H. Cycon and K. -E. Hellwig, "Conditional expectations in generalized probability theory," J. Math. Phys. **18**, 1154–1161 (1977).
[7]M. D. Srinivas, "Foundations of quantum probability theory," J. Math. Phys. **16**, 1672–1685 (1975).
[8]M. D. Srinivas, "Collapse postulate for observables with continuous spectra," Commun. Math. Phys. **71**, 131–158 (1980).
[9]R. Mercer, "General quantum measurements: Local transition maps," Commun. Math. Phys. **84**, 239–250 (1982).
[10]J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton, U. P., Princeton, NJ, 1955).
[11]M. Nakamura and H. Umegaki, "On von Neumann's theory of measurements in quantum statistics," Math. Jpn. **7**, 151–157 (1962).
[12]H. Umegaki, "Conditional expectation in an operator algebra, I–II," Tohoku Math. J. **6**, 177–181 (1954); **8**, 86–100 (1956).
[13]W. B. Areveson, "Analyticity in operator algebras," Am. J. Math. **89**, 578–642 (1967).
[14]K. Kraus, "General state changes in quantum theory," Ann. Phys. **64**, 311–335 (1971).
[15]E. P. Wigner, "Die Messung Quantenmechanischer Operatoren," Z. Phys. **133**, 101–108 (1952).
[16]H. Araki and M. M. Yanase, "Measurements of quantum mechanical operators," Phys. Rev. **120**, 622–626 (1960).
[17]J. Tomiyama, "On the projection of norm one in W^{*}-algebra," Proc. Jpn. Acad. **33**, 608–612 (1957).
[18]W. F. Stinespring, "Positive functions on C^{*}-algebras," Proc. Am. Math. Soc. **6**, 211–216 (1955).
[19]M. Takesaki, *Theory of Operator Algebras I* (Springer-Verlag, New York, 1979).
[20]G. W. Mackey, "Borel structure in groups and their duals," Trans. Am. Math. Soc. **85**, 134–165 (1957).
[21]G. Lüders, "Über die Zustandsanderung durch den Messprozess," Ann. Physik **8**(6), 322–328 (1951).

The Darboux transformation and solvable double-well potential models for Schrödinger equations[a)]

W. M. Zheng[b)]
Center for Studies in Statistical Mechanics, University of Texas at Austin, Austin, Texas 78712

(Received 8 September 1982; accepted for publication 10 December 1982)

The Darboux transformation, a method used to transform a Schrödinger-type equation to a Schrödinger equation with a new potential, is discussed. An exactly solvable double-well potential model for the one-dimensional Schrödinger equation is obtained.

PACS numbers: 03.65.Ge, 02.30.Em, 31.90. + s

I. INTRODUCTION

Much effort has been made to look for exactly solvable models for the one-dimensional Schrödinger equations. Double-well potential problems occur in the quantum theory of molecules. Because the Fokker–Planck equation is closely related to the Schrödinger equation,[1] the solution of the above problem can be directly applied to the problem of diffusion in a bistable potential field. Recently, great attention has been put on constructing exactly solvable bistable models.

In general, there are four ways to devise potentials. The first is to use piecewise potentials[1,2] such as the double square well and the double oscillator, this being the most common method. Its main difficulty, however, is that to obtain eigenvalues from the matching conditions, one needs to solve transcendental equations, for which analytic solutions of eigenvalues are not available unless in some limiting cases expansion formulas can be applied to find approximate solutions for the low-lying eigenvalues. The second[3] is to construct potentials from the wave functions which are solutions to two or more Schrödinger equations with simple potentials at the same eigenvalue. Since the potentials are made to fit given wave functions, a set of different eigenfunctions cannot be obtained in this way. The third[4] is to solve the Schrödinger equation directly for specially chosen potentials; for example, the potential with three parameters, β, ξ, and a positive integer n:

$$V(x) = (\hbar^2\beta^2/2m)\left[\tfrac{1}{8}\xi^2 \cosh 4\beta x - (n+1)\xi \cosh 2\beta x - \tfrac{1}{8}\xi^2\right]. \tag{1}$$

For this potential, the low-lying eigenfunctions can be found analytically in a form of finite-term summation of simple functions. The fourth method is to transform the Schrödinger equations to be solved to known solvable differential equations. There are many examples of this given in textbooks[5]; so far it appears that no example dealing with a double-well potential has been solved in this way.

In this paper two systematic methods, the Darboux transformation[6] and a new one, will be presented for transforming known solvable Schrödinger-type equations to Schrödinger equations with new different potentials. As an example, a double-well potential model will be obtained from the Weber equation,[7] and other interesting applications of the transformations will be given.

[a)] Supported in part by the Robert A. Welch Foundation.

[b)] On leave of absence from the Institute of Theoretical Physics, Academia Sinica, Beijing, China.

II. THE DARBOUX TRANSFORMATION[8]

The Darboux theorem: If the general solution $\varphi = \varphi(x)$ of the equation

$$\frac{d^2\varphi}{dx^2} + [\epsilon - U(x)]\varphi = 0 \tag{2}$$

is known for all values of ϵ and for a particular value of $\epsilon = \epsilon_0$, the particular solution is $\varphi = \varphi_0(x)$. Then the general solution of the equation

$$\frac{d^2\psi}{dx^2} + [E - V(x)]\psi = 0 \tag{3a}$$

with

$$V(x) = \varphi_0(x)\frac{d^2}{dx^2}\left(\frac{1}{\varphi_0(x)}\right), \tag{3b}$$

$$E = \epsilon - \epsilon_0 \tag{3c}$$

for $E \neq 0$ is

$$\psi(x) = \varphi_0(x)(\varphi(x)/\varphi_0(x))' \tag{4a}$$

$$= \varphi'(x) - \frac{\varphi_0'(x)}{\varphi_0(x)}\varphi(x). \tag{4b}$$

The Darboux transformation (4a) was previously used to transform the Schrödinger equations with the potentials given by Eq. (1).[4] It should be emphasized that the Darboux transformation is very general in the sense that the original equation (2) need not be a physical Schrödinger equation.

As an example, we consider the Weber equation[7]

$$\frac{d^2y}{dx^2} - \left(\frac{x^2}{4} + a\right)y = 0. \tag{5}$$

For any given parameter a, this equation has the particular solution

$$y_1(a,x) = e^{-x^2/4}\,{}_1F_1(a/2 + \tfrac{1}{4};\tfrac{1}{2};x^2/2), \tag{6}$$

where ${}_1F_1(\alpha;\beta;x)$ is a Kummer function. Here we have

$$\epsilon_0 = 0, \quad U(x) = x^2/4 + a. \tag{7}$$

From the asymptotic behavior of the Kummer function, we know that the positive definite function $y_1(a,x)$ does not satisfy the natural boundary conditions, i.e., it does not vanish at infinity, so it is not a "physical" solution. Equation (5) is only a Schrödinger-type equation. According to Eq. (3b), the transformed potential is

$$V_a(x) = y_1(a,x)\frac{d^2}{dx^2}\left(\frac{1}{y_1(a,x)}\right)$$

 0022-2488/84/010088-03$02.50

$$= 2\left[\frac{y_1'(a,x)}{y_1(a,x)}\right]^2 - \left(\frac{x^2}{4} + a\right). \tag{8}$$

This potential has been considered in the discussion of the Fokker–Planck equation for diffusion of a Brownian particle with a particular initial δ-function distribution peaked at $x = 0$.[8]

The curvature of $V_a(x)$ at $x = 0$ is

$$V_a''(x)|_{x=0} = 4a - \tfrac{1}{2} \times \begin{cases} >0, & a > 1/2\sqrt{2} \quad \text{or} \quad a < -1/2\sqrt{2}, \\ =0, & a = \pm 1/2\sqrt{2}, \\ <0, & -1/2\sqrt{2} < a < 1/2\sqrt{2}. \end{cases} \tag{9}$$

The shapes of the symmetric functions $V_a(x)$ are shown for different values of a in Fig. 1. One can see that for $|a| = 0.5$, $V_a(x)$ is a single well; for $a = -0.25$, $V_a(x)$ has a double-well structure; for $a = 0.25$, the shape of the curve is relatively complex.

The generated Schrödinger equation for the transformed potential $V_a(x)$ is

$$\frac{d^2\psi}{dx^2} + \left\{E - 2\left[\frac{y_1'(a,x)}{y_1(a,x)}\right]^2 + \left(\frac{x^2}{4} + a\right)\right\}\psi = 0. \tag{10}$$

It is easy to verify that function $[y_1(a,x)]^{-1}$ satisfies Eq. (10) for $E = 0$. The function $[y_1(a,x)]^{-1}$ has no node (as long as a is not less than -0.5) and is a square-integrable function (see Sec. IV) satisfying the natural boundary conditions, so it is the ground state of Eq. (10). The higher eigenvalues and eigenfunctions can be obtained from Eqs. (3c), (4a), and (5):

$$E_n = n + a + \tfrac{1}{2}, \quad n = 0,1,2,\cdots, \tag{11a}$$

$$\psi_n(x) = y_1(a,x)\frac{d}{dx}\left(\frac{D_n(x)}{y_1(a,x)}\right), \tag{11b}$$

where the Weber function $D_n(x)$ can be expressed in terms of the Hermite polynomial $H_n(x)$[7]

$$D_n(x) = 2^{-n/2}e^{-x^2/4}H_n(x/\sqrt{2}). \tag{12}$$

We have thus found *all* the eigenvalues and eigenfunctions of Eq. (10). Furthermore, the normalization factor for $\psi_n(x)$ can be obtained analytically (see Sec. IV).

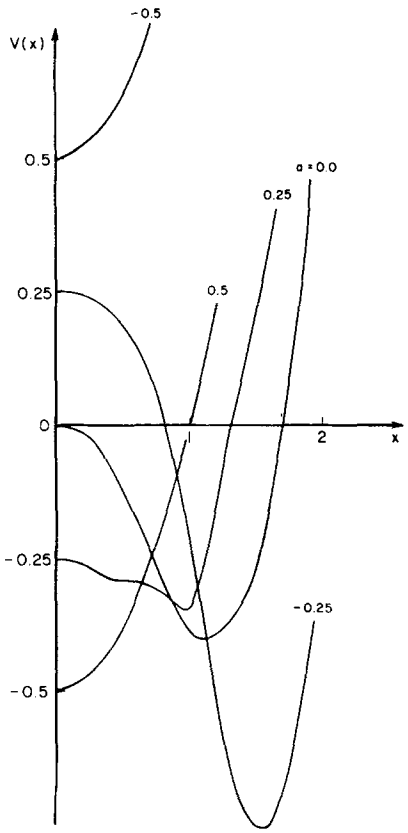

FIG. 1. Shape of $V(x)$.

III. A NEW TRANSFORMATION

The solutions to Eq. (3a) can also be given in another form:

$$\psi(x) = \frac{1}{\varphi_0(x)}\int^x \varphi(x)\varphi_0(x)\,dx. \tag{13}$$

By substituting $1/\varphi_0$ into Eq. (3a), one can directly verify that it is the solution to Eq. (3a) for eigenvalue $E = 0$. If we reinterpret Eq. (3a) as the original untransformed equation, then from the Darboux theorem we have the transformed potential

$$\tilde{V}(x) = \frac{1}{\varphi_0}\frac{d^2\varphi_0}{dx^2} = U(x) - \epsilon_0 \tag{14}$$

and the transformed equation

$$\frac{d^2\tilde{\psi}}{dx^2} + [(E + \epsilon_0) - U(x)]\tilde{\psi} = 0,$$

which is the same as Eq. (2) if one notices $E = \epsilon - \epsilon_0$. Thus, from Eq. (4a), we obtain

$$\tilde{\psi} = \varphi = \frac{1}{\varphi_0}(\psi\varphi_0)' \equiv \mathscr{W}\psi \tag{15a}$$

or

$$\psi = \mathscr{W}^{-1}\varphi = \frac{1}{\varphi_0}\int^x \varphi(x)\varphi_0(x)\,dx. \tag{15b}$$

To guarantee

$$\mathscr{W}^{-1}\mathscr{W} = \mathscr{W}\mathscr{W}^{-1} = \mathscr{I},$$

where $\mathscr{I}$ is the identity operator, we should choose the lower limit x_0 for integration in Eq. (15b) such that $\varphi(x_0) = 0$. However, the undetermined constant in the indefinite integral (13) is of no importance because the eigenvalue corresponding to $1/\varphi_0$ equals zero.

IV. DISCUSSION

(1) From the two forms of $\psi(x)$, Eqs. (4a) and (13), we have the general relation

$$\int_0^x \varphi_n(x)\varphi_0(x)\,dx = c\left[\varphi_0(x)\left(\frac{\varphi_n(x)}{\varphi_0(x)}\right)' - d\right]\varphi_0(x), \tag{16}$$

where

$$d=\varphi_0(0)\left(\frac{\varphi_n(x)}{\varphi_0(x)}\right)'_{x=0}$$

and c is a constant independent of x. By differentiating both sides of Eq. (16), we reobtain Eq. (2); thus the constant c is determined as $c=-1/E_n$.

(2) Calculation of the normalization factor for ψ_n defined by Eq. (4a) is as follows:

$$\begin{aligned}I_n &\equiv\int_{-\infty}^{\infty}\psi_n^2(x)\,dx\\ &=-2E_n\int_0^{\infty}\left[\varphi_0\left(\frac{\varphi_n}{\varphi_0}\right)'\right]\\ &\quad\times\left[\frac{1}{\varphi_0}\left(\int_0^x\varphi_n\varphi_0dx'+d\right)\right]dx\\ &=-2E_n\left\{\left[\left[\frac{\varphi_n}{\varphi_0}\left(\int_0^x\varphi_n\varphi_0dx'+d\right)\right]\right]_0^{\infty}-\int_0^{\infty}\varphi_n^2dx\right\}\\ &=2E_n\left[\frac{\varphi_n(0)}{\varphi_0(0)}d+\int_0^{\infty}\varphi_n^2dx\right].\end{aligned}\tag{17}$$

Thus for the example given in Sec. II we have

$$\begin{aligned}d&=y_1(a,0)\left(\frac{D_n(x)}{y_1(a,x)}\right)'_{x=0}\\ &=D'_n(x)|_{x=0}.\end{aligned}$$

From

$$D_n(0)\cdot D'_n(0)=0,$$

$$y_1(a,0)=1,\quad y'_1(a,0)=0,$$

and

$$\int_0^{\infty}D_n^2(x)dx=\tfrac{1}{2}n!(2\pi)^{1/2},$$

we obtain finally

$$I_n=(n+a+\tfrac{1}{2})n!(2\pi)^{1/2}.\tag{18}$$

(3) Calculation of the normalization factor for the ground state is as follows: For the ground state $E=0$, from Eq. (11), we have

$$n=-a-\tfrac{1}{2},$$

$$\frac{1}{y_1(a,x)}=ky_1(a,x)\frac{d}{dx}\left(\frac{D_{-a-1/2}(x)-D_{-a-1/2}(-x)}{2y_1(a,x)}\right),\tag{19}$$

where k is a constant. Because

$$y_1(a,0)=1,\quad D'_{\nu}(0)=-2^{\nu/2+1/2}\frac{\sqrt{\pi}}{\Gamma(-2/2)},\tag{20}$$

we have, from Eq. (19),

$$k=-2^{a/2-1/4}\Gamma(a/2+\tfrac{1}{4})/\sqrt{\pi}.\tag{21}$$

Therefore

$$\begin{aligned}\int_{-\infty}^{\infty}\frac{dx}{y_1^2(a,x)}&=2k\int_0^{\infty}d\left(\frac{D_{-a-1/2}(x)-D_{-a-1/2}(-x)}{2y_1(a,x)}\right)\\ &=2k\frac{D_{-a-1/2}(x)-D_{-a-1/2}(-x)}{2y_1(a,x)}\Bigg|_0^{\infty}\\ &=-k\lim_{x\to\infty}\left(\frac{D_{-a-1/2}(-x)}{y_1(a,x)}\right)\\ &=\frac{\sqrt{2}\Gamma(a/2+\frac{1}{4})}{\Gamma(a/2+\frac{3}{4})}.\end{aligned}\tag{22}$$

This result was derived previously in a quite different way.[9] To our knowledge, this integral is not found in tables.

(4) The exact solutions obtained by means of the transformations can be used to test approximate methods of solutions. For example, applying the WKB approximation to the energy levels below the top of the barrier in a symmetric double well,[10] one can find that at low transmission the energy levels appear in close pairs. The spectrum of our model is one in which all the energy levels higher than the lowest two are equally spaced. Thus the model is an example where the WKB approximation fails.

(5) Choosing a linear combination of $\alpha y_1(a,x)+\beta y_2(a,x)$ instead of $y_1(a,x)$ for $\varphi_0(x)$, one can construct an asymmetric potential similarly. The discussion will be made elsewhere.

(6) The methods can be applied to solve the Fokker–Planck equation[11] and other problems. In addition, the exactly solvable double-well potential model has some pedagogic value.

ACKNOWLEDGMENTS

The author would like to express deep gratitude to Max O. Hongler for introducing Ref. 4 and providing preprints of Refs. 8 and 9, as well as for the fruitful discussions which stimulated the present work. Thanks must be expressed to Professor I. Prigogine, Professor L. Reichl, and Professor W. Schieve for their hospitality at the Center for Studies in Statistical Mechanics of the University of Texas at Austin.

[1] N. G. van Kampen, J. Stat. Phys. **17**, 71 (1977).
[2] E. Merzbacher, *Quantum Mechanics* (Wiley, New York, 1970); E. W. Gettys and H. W. Gruber, Am. J. Phys. **43**, 625 (1975); M. Prakash, J. Phys. A **9**, 1847 (1976); J. Thomchick, J. P. McKelvey, and C. F. Elliott, Phys. Lett. A **66**, 86 (1978).
[3] R. G. Winter, Am. J. Phys. **45**, 569 (1977).
[4] M. Razavy, Am. J. Phys. **48**, 285 (1980).
[5] For instance, S. Flügge and H. Marschall, *Rechenmethoden der Quantentheorie* (Springer, Berlin, 1952).
[6] J. G. Darboux, C. R. Acad. Sci. Paris **94**, 1456 (1882); E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956), p. 132.
[7] M. Abramovitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
[8] M. O. Hongler, preprint.
[9] M. O. Hongler and N. D. Quach, preprint.
[10] D. ter Haar, *Problems in Quantum Mechanics* (Academic, New York, 1974), p. 144.
[11] M. O. Hongler and W. M. Zheng, J. Stat. Phys. **29**, 317 (1982).

Transmutation as a minimizing procedure

Robert Carroll
Mathematics Department, University of Illinois, Urbana, Illinois 61801

Stanley Dolzycki
Mathematics Department, Eastern Illinois University, Charleston, Illinois 61920

(Received 22 July 1983; accepted for publication 2 September 1983)

It is shown formally how transmutation kernels can be characterized via a minimizing procedure. The technique then can be extended to more general operators and transmutations.

PACS numbers: 03.65.Nk

1. INTRODUCTION

In Ref. 1 it is shown how Gel'fand–Levitan (GL) equations can be obtained by minimizing a certain quadratic functional $Q(t,K)$. The motiviation to consider $Q(t,K)$ came from a problem in optics[2] involving a feedback mechanism and statistical averaging but no motivation could be provided within scattering theory to consider $Q(t,K)$. Thus the process producing GL equations appeared to simply involve a mathematical trick which was naturally considered to be "unsatisfactory" in Ref. 1 and the "meaning" of such procedures seemed to be worth pursuing further. In the present article we will provide an interpretation of such minimizing processes in the context of transmutation theory which leads us eventually to minimize a quadratic functional essentially the same as $Q(t,K)$. This involves a characterization of transmutation kernels themselves in terms of a minimization procedure, and we sketch the development for a classical situation (a more extensive and general treatment for operators and transmutations as in Ref. 3 is clearly indicated and will appear later). Let us remark that there is a discrete version (which does not directly generalize) of a related minimization in the context of orthogonal polynomials, but without a connection to $Q(t,K)$ nor any explicit link to transmutation.[4] Although our characterization of transmutation kernels via a minimization is of interest in itself, and moreover provides "motivation" for some constructions in Ref. 1, there seem to be some deeper relations still beneath the surface. In particular, one knows that various connections between spectral measures, transmutation, autocorrelation functions, stochastic analysis, least squares optimization, etc., are all involved here.[4–8] Thus, hopefully this article will provide a contribution toward unifying some of this material as well.

2. BASIC CONSTRUCTIONS

In classical (half-line) inverse scattering theory in quantum mechanics,[9,10] one connects eigenfunctions of the Schrödinger operator $Q = D^2 - q(x)$ (q real here) with eigenfunctions of D^2 via certain (triangular) transmutation kernels of the form $\beta(y,x) = \delta(x-y) + K(y,x)$, where $K(y,x) = 0$ for $x > y$ (such K will be called causal here). Thus let $\varphi^Q_\lambda(x)$ [resp. $\theta^Q_\lambda(x)$] be solutions of

$$Qu = -\lambda^2 u \tag{2.1}$$

satisfying $\varphi^Q_\lambda(0) = 1$ and $D_x\varphi^Q_\lambda(0) = 0$ [resp. $\theta^Q_\lambda(0) = 0$ and $D_x\theta^Q_\lambda(0) = 1$]. We will write $s(\lambda,x)$ for φ^Q_λ or θ^Q_λ and think of connecting $s(\lambda,x)$ to $a(\lambda,x) = \cos\lambda x$ or $a(\lambda,x) = \sin\lambda x/\lambda$ by a formula

$$s(\lambda,y) = (1+K)a = a(\lambda,y) + \int_0^y K(y,x)a(\lambda,x)dx, \tag{2.2}$$

which we know to be valid for the GL kernel $K = K_0$. We can assume K_0 exists here and our procedure is designed to characterize it via minimization. For simplicity now let us think of $s = \theta^Q_\lambda$ and $a = \sin\lambda x/\lambda$, and remark that a systematic theory of transmutation operators $B: P \to Q$ can be developed for much more general differential operators P and Q (Ref. 3); the techniques of this article will be correspondingly extended at another time. Now one knows that associated to Q and the eigenfunctions $\theta^Q_\lambda = s$ is a spectral measure $d\omega = d\omega_Q$ which we assume here for convenience to be of the form $d\omega = \omega\, d\lambda$ (no bound states). Thus one can suppose, e.g.,

$$\int_0^\infty \theta^Q_\lambda(x)\theta^Q_\lambda(y)d\omega(\lambda) = \delta(x-y) \tag{2.3}$$

(acting on suitable functions) and we write $d\omega = d\sigma + 2\lambda^2 d\lambda/\pi$ with $\int_0^\infty a(\lambda,x)a(\lambda,y)d\sigma = \Omega(x,y)$. Thus

$$\begin{aligned}\mathfrak{A}(x,y) &= \int_0^\infty a(\lambda,x)a(\lambda,y)d\omega \\ &= \delta(x-y) + \Omega(x,y) = (1+\Omega)(x,y),\end{aligned} \tag{2.4}$$

where $a = \sin\lambda x/\lambda$ [we write 1 for the identity operator with kernel $\delta(x-y)$]. Now consider the expression (T arbitrary and fixed)

$$\Xi(T,K) = \int_0^T \int_0^\infty \{[(1+K)a(\lambda,\cdot)](y) - s(\lambda,y)\}^2 d\omega(\lambda)dy. \tag{2.5}$$

Note that when K is the GL kernel K_0 [which makes (2.2) correct], then formally $\Xi(T,K) = 0$. We can think here of Q, s, a, and $d\omega$ as given and the (causal) kernel $K(y,x)$ in (2.5) as unknown. It will be shown formally that:

Theorem 2.1: The kernel K is obtained by minimizing $\Xi(T,K)$ over a suitable class of admissible causal kernels satisfies the GL equation and represents the transmutation kernel K_0 connecting s and a via (2.2).

3. FORMAL ARGUMENTS

We proceed formally and refer to standard sources[3,9,10] for information about natural properties of $K(y,x)$, etc. Thus, from (2.5), for causal K,

 0022-2488/84/010091-03$02.50

$$\Xi(T,K)=\int_0^T\int_0^\infty\Big\{[a(\lambda,y)-s(\lambda,y)]^2+2a(\lambda,y)\int_0^y K(y,x)a(\lambda,x)dx-2s(\lambda,y)\int_0^y K(y,x)a(\lambda,x)dx+\int_0^y K(y,x)a(\lambda,x)dx\times\int_0^y K(y,\xi)a(\lambda,\xi)d\xi\Big\}d\omega(\lambda)dy. \tag{3.1}$$

Now one integrates in λ, using (2.4), and the convention $\int_0^T\Omega(y,y)dy=\operatorname{Tr}\Omega$, for example, to obtain (note that the trace Tr depends on T)

$$\Xi(T,K)=\hat{\Xi}(T)+2\operatorname{Tr}K+2\int_0^T\int_0^y K(y,x)\Omega(x,y)dx\,dy-2\int_0^T\int_0^y K(y,x)\tilde{\beta}(y,x)dx\,dy+\int_0^T\int_0^y\int_0^y K(y,x)K(y,\xi)\{\delta(x-\xi)+\Omega(x,\xi)\}d\xi\,dx\,dy, \tag{3.2}$$

where we have written $\hat{\Xi}(T)=\int_0^T\{a(\lambda,y)-s(\lambda,y)\}^2d\omega\,dy$ which we know makes sense [in fact $\hat{\Xi}(T)=\int_0^T\int_0^\infty(K_0a)^2d\omega\,dy=\int_0^T\int_0^y\int_0^y K_0(y,x)K_0(y,\xi)\{\delta(x-\xi)+\Omega(x,\xi)\}d\xi\,dx\,dy=\operatorname{Tr}\{K_0(1+\Omega)K_0^*\}$—see calculations below]. Here the term $\tilde{\beta}(y,x)=\langle s(\lambda,y),a(\lambda,x)\rangle_\omega$ is a standard object in general transmutation theory[3] which appears in extended GL equations [e.g., $\langle\beta(y,t),\mathfrak{A}(t,x)\rangle=\tilde{\beta}(y,x)$] and in particular $\tilde{\beta}(y,x)=0$ for $x<y$ (i.e., it is anticausal) with a $\delta(x-y)$ term arising along the diagonal.[11] Thus the $\tilde{\beta}$ term contributes $-2\int_0^T K(y,y)dy=-2\operatorname{Tr}K$ to (3.2). We can write now

$$K\Omega g(y)=\int_0^y K(y,x)\int_0^\infty\Omega(x,s)g(s)ds\,dx=\int_0^\infty g(s)\Big\{\int_0^y K(y,x)\Omega(x,s)dx\Big\}ds \tag{3.3}$$

(for suitable g) so that $\operatorname{Tr}K\Omega=\int_0^T\{\int_0^y K(y,x)\Omega(x,y)dx\}dy$. Similarly $\ker K^*=K(\cdot,x)$ on $[x,\infty)$ since $\int_0^\infty g(y)\int_0^y K(y,x)h(x)dx\,dy=\int_0^\infty h(x)\int_x^\infty g(y)K(y,x)dy\,dx$, and hence

$$KK^*g(y)=\int_0^y K(y,x)\int_x^\infty K(\xi,x)g(\xi)d\xi\,dx=\int_0^\infty g(\xi)\int_0^{\min(y,\xi)}K(y,x)K(\xi,x)dx\,d\xi. \tag{3.4}$$

Consequently $\operatorname{Tr}KK^*=\int_0^T\{\int_0^y K(y,x)K(y,x)dx\}dy$. Finally we have

$$K\Omega K^*g(y)=\int_0^y K(y,x)\int_0^\infty\Omega(x,s)\times\int_s^\infty K(\xi,s)g(\xi)d\xi\,ds\,dx=\int_0^y K(y,x)\int_0^\infty g(\xi)\int_0^\xi\Omega(x,s)K(\xi,s)ds\,d\xi\,dx=\int_0^\infty g(\xi)\Big\{\int_0^y K(y,x)\int_0^\xi\Omega(x,s)K(\xi,s)ds\,dx\Big\}d\xi. \tag{3.5}$$

It follows that $\operatorname{Tr}K\Omega K^*=\int_0^T\{\int_0^y K(y,x)\int_0^y\Omega(x,s)K(y,s)ds\,dx\}dy$. Now go back to (3.2) and insert the information just derived from Eqs. (3.3)–(3.5) plus the $\tilde{\beta}$ contribution, to obtain

Lemma 3.1: The expression $\Xi(T,K)$ defined in (2.5) can be written

$$\Xi(T,K)=\hat{\Xi}(T)+\operatorname{Tr}\{K(1+\Omega)K^*+K\Omega+\Omega K^*\}. \tag{3.6}$$

Proof: One obtains from (3.2), $\Xi(T,K)=\hat{\Xi}(T)+\operatorname{Tr}\{2K\Omega+KK^*+K\Omega K^*\}$. But $K(1+\Omega)K^*=KK^*+K\Omega K^*$ with $\operatorname{Tr}K\Omega=\operatorname{Tr}\Omega K^*$ (note $\Omega^*=\Omega$). Q.E.D.

Written in the form (3.6), $\Xi(T,K)$ is essentially in the same form as the expression $Q(t,K)$ (or D) in Refs. 1 and 2. We now formally examine a variational argument to minimize $\Xi=\Xi(T,K)$. Thus [note $\Xi\geqslant 0$ from (2.5)] we know there is a minimizing $K=K_0$ in some additive class $\mathfrak{K}$ of admissible (causal) kernels. Then consider $K=K_0+\epsilon L$ in $\Xi(T,K)=\hat{\Xi}(T)+\Xi_K(T)$ [$\hat{\Xi}(T)$ is independent of K] for $L\in\mathfrak{K}$ and ϵ a real number. Formally we set $D_\epsilon\Xi_k(T)|_{\epsilon=0}=0$. This leads to $\operatorname{Tr}\{L(1+\Omega)K_0^*\}+\operatorname{Tr}\{K_0(1+\Omega)L^*\}+\operatorname{Tr}L\Omega+\operatorname{Tr}\Omega L^*=2\operatorname{Tr}\{[K_0(1+\Omega)+\Omega]L^*\}=0$ for $L\in\mathfrak{K}$. If we write now $\Lambda=K_0(1+\Omega)+\Omega$ with kernel $\Lambda(y,x)$, then evidently $\ker\Lambda L^*=\int_0^{\min(s,y)}\Lambda(y,x)L(s,x)dx$ [cf. (3.4)] and $\operatorname{Tr}\Lambda L^*=\int_0^T\{\int_0^y\Lambda(y,x)L(y,x)dx\}dy$. The statement that $\operatorname{Tr}\Lambda L^*=0$ for all $L\in\mathfrak{K}$ will be true if $\Lambda(y,x)=0$ for $x<y$, and heuristically we conclude here the converse.

Theorem 3.2: The (unique) minimizing kernel K_0 satisfies the GL equation $K_0(y,x)+\Omega(y,x)+\int_0^y K_0(y,\xi)\Omega(\xi,x)d\xi=0$ for $x<y$.

One knows that the GL equation has a unique solution and this is the transmutation kernel of (2.2).[3] Hence Theorem 2.1 is verified formally.

Remark 3.3: Let us note also the following calculation which will specify (again) the minimum Ξ_0 of $\Xi(T,K)$ achieved at the GL kernel K_0. Thus given the GL equation in Theorem 3.2 we can say $K_0+\Omega+K_0\Omega=B^*$, where B is a causal operator. It follows easily that $1+B^*=(1+K_0)(1+\Omega)$, and thus

$$(1+B^*)(1+K_0^*)=(1+K_0)(1+\Omega)(1+K_0^*) \tag{3.7}$$

which is formally self-adjoint. But the left side of (3.7) is $1+$ an anticausal operator so both sides of (3.7) must be 1 (cf. Ref. 1). Hence [recall $\hat{\Xi}(T)=\operatorname{Tr}\{K_0(1+\Omega)K_0^*\}$], $\Xi_0=\min\Xi(T,K)=\hat{\Xi}(T)+\min\Xi_K(T)=\operatorname{Tr}\{2K_0(1+\Omega)K_0^*+K_0\Omega+\Omega K_0^*\}=\operatorname{Tr}\{2(1+K_0)(1+\Omega)(1+K_0^*)-2(1+\Omega)-2K_0-2K_0^*-K_0\Omega-\Omega K_0^*\}=-\operatorname{Tr}\{2\Omega+2K_0+2K_0^*+K_0\Omega+\Omega K_0^*\}=-\operatorname{Tr}\{B+B^*+K_0+K_0^*\}=\operatorname{Tr}\{B^*K_0^*+K_0B\}=0$ (since K_0 and B are causal—cf. Ref. 1). This is the desired conclusion.[12]

[1] F. Dyson, in *Studies in Math. Physics* (Princeton, U.P., Princeton, NJ, 1976), pp. 151–167.
[2] F. Dyson, J. Opt. Soc. Am. **65**, 551–558 (1975).

[3]R. Carroll, *Transmutation, Scattering Theory, and Special Functions* (North-Holland, Amsterdam, 1982).
[4]K. Case, Advances in Math. Suppl. Studies, **3**, 25–43 (1978).
[5]H. Dym and H. McKean, *Gaussian Processes, Function Theory, and the Inverse Spectral Problem* (Academic, New York, 1976).
[6]R. Carroll and F. Santosa, in *Proceedings of the Conference on Inverse Scattering*, University of Tulsa, 1983 (in press).
[7]R. Carroll and F. Santosa, "Spectral measures and autocorrelation via transmutation," C. R. Roy. Soc. Canada (in press).
[8]T. Kailath, IEEE Trans. Inf. Theory, **IT-20**, 145–181 (1974).
[9]K. Chadan and P. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer, New York, 1977).
[10]L. Faddeev, Uspehi Mat. Nauk **14**, 57–119 (1959).
[11]V. Marcenko, *Sturm-Liouville Operators and Their Applications*, (Izd. Nauk. Dumka, Kiev, 1977)—see also Ref. 3.
[12]This calculation suggests (as is indeed the case) that the characterization of K_0 by minimization does not require the trace argument [i.e., the last integral in (2.3)]; the details will appear elsewhere.

Inverse scattering in dimension two[a)]

Margaret Cheney
Department of Mathematics, Stanford University, Stanford, California 94305

(Received 17 May 1983; accepted for publication 5 August 1983)

The inverse scattering problem is solved for the two-dimensional time-independent Schrödinger equation. That is, the potential is reconstructed from the scattering amplitude, which is assumed to be known for all energies and angles.

PACS numbers: 03.65.Nk

INTRODUCTION

Our goal here is to solve the inverse scattering problem for the Schrödinger equation in two dimensions. That is, we recover the potential from the scattering data, which we take to be the entire scattering amplitude as a function of the energy and two angles.

Actually, there are a number of aspects to the inverse scattering problem: uniqueness, reconstruction, construction, and characterization. The uniqueness problem deals with the question, "Does the scattering amplitude uniquely determine the potential?" The reconstruction problem is the problem of constructing a potential from scattering data that are known to come from an underlying potential, whereas the construction problem is to construct the potential without this knowledge. And finally, the characterization problem is to determine what scattering data actually arise and to correlate properties of the potential with properties of the data.

In the one-dimensional case, solutions to all these questions via the Gel'fand–Levitan and Marchenko methods are well known. Moreover, in the 25 years since their discovery, one-dimensional inverse scattering techniques have been found to have important applications not only to particle physics but also to geophysics and to certain classes of nonlinear differential equations, the so-called soliton equations, which themselves describe a wide range of phenomena.

The popularity of one-dimensional inverse scattering has inspired much interest in the construction of higher-dimensional inversion theories; nevertheless, the uniqueness question was for many years the only one of the higher-dimensional inversion questions that was answered satisfactorily: although in the one-dimensional case additional bound state information is needed for uniqueness, in three dimensions the scattering data alone do indeed determine the potential uniquely. The other three inversion questions, however, are so much more difficult than their one-dimensional counterparts that for 25 years attempts to solve even the simplest one, the reconstruction problem, met with only partial success.

The first of these reconstruction attempts was made by Kay and Moses,[1,2] whose generalization of the Gel'fand–Levitan method accomplished inversion in a class of potentials which includes those that are nonlocal (i.e., are not multiplication operators) in the angular variables. This class, however, was never shown to include the local potentials. Another attempt, made by Faddeev[3] and Newton,[4] depended on a new, direction-dependent Green's function which had been constructed by Faddeev.[5,6] This Faddeev–Newton method, however, was awkward and cumbersome, and was hampered by a number of unanswered questions concerning exceptional points. A third attempt at multidimensional inverse scattering was made by Prosser,[7–9] who attacked all three of the remaining inversion problems using essentially an iterative scheme that applies only to weak potentials and to scattering data that are small in a certain norm. Recently, Morawetz[10] has found a generalization to higher dimensions of the Deift–Trubowitz one-dimensional method.[11] Her scheme, which is also iterative, has yet to be shown to converge for any specific class of potentials. Then, beginning in 1980, Newton published a series of papers[12–14] containing successful and elegant generalizations of both the Gel'fand–Levitan and Marchenko methods to three dimensions. Both his methods solve the reconstruction problem; his Marchenko method, in addition, solves the construction problem and gives a partial solution to the characterization problem. In this paper, we shall adapt Newton's generalized Marchenko method to dimension two.

Newton's ideas could, in fact, be applied to inverse scattering in any dimension provided that the relevant estimates hold; the success of Newton's inverse scattering techniques in two dimensions thus depends on estimates that can be considered part of the direct scattering problem.

The first five sections therefore contain the necessary results concerning direct scattering. Section 1 sets up the problem and contains basic facts and definitions for scattering in two dimensions. Also contained in Sec. 1 is a result on the behavior of the wave function for large energy. Section 2 contains the investigation of the wave function's small energy behavior.

Knowledge of the energy dependence of the wave function is crucial to our method of inverse scattering. In fact, the behavior at zero and infinity, which is heavily dimension-dependent, is the reason that later estimates must have proofs quite different from those of the corresponding estimates of Newton.[12,13] The behavior in two dimensions differs from that in three dimensions in its faster decay at infinity and in the presence of zero-energy singularities that appear in the derivatives.

The properties of symmetry and analytic continuation, however, are exactly the same as in three dimensions. These properties are recorded in Sec. 3.

[a)] This is based on the author's Ph.D. thesis, "Quantum Mechanical Scattering and Inverse Scattering in Two Dimensions," Indiana University, 1982.

 0022-2488/84/010094-14$02.50

Another ingredient essential to inverse scattering is a good deal of spectral theory. Fortunately many of the needed results have already been proved by Agmon[15] and are merely quoted in Sec. 4. These include not only the unitarity of the S matrix but also the eigenfunction expansion theorem, which is used in Sec. 5 to prove that the scattering operator maps incoming to outgoing wave functions. This relation, when combined with the analyticity properties of the wave function, forms a Riemann–Hilbert problem or a Wiener–Hopf factorization problem. This is the key to the Marchenko method of inversion.

We arrive at Sec. 6 having proved all the estimates necessary for the generalized Marchenko method of inverse scattering. The inverse scattering results, therefore, are all contained in this section; in fact the reader interested only in the results might read just Sec. 6, referring to Sec. 1 for notation. Section 6 is intended merely to give the reader a taste of the inverse scattering theory that is more fully developed in Newton's series of papers and which is generally dimension-independent. Nevertheless, in Sec. 6, the uniqueness theorem is proved, the Marchenko equation is derived, and the potential is extracted from the solution of the Marchenko equation. Thus the results of Sec. 6 solve only the reconstruction problem; the reader interested in construction should refer to Newton's work.[13,14]

Notation

In what follows we denote by $\|\cdot\|_p$ the usual L^p norm; if confusion is possible, we will add as a superscript the variable with respect to which the L^p norm is being taken.

The symbol $\|\cdot\|_{m,p}$ denotes the norm of the Sobolev space $W^{m,p}$, the space of functions with m derivatives in L^p. We shall write $H^2 = W^{2,2}$.

The symbols $\theta,\theta',\theta'',\phi$, etc., in most places denote unit vectors, although occasionally they will be used as simple angles in carrying out integrations. Where confusion is possible, the unit vectors will be decorated with hats, e.g., $\hat{\theta}$.

1. PRELIMINARIES

Two-particle scattering in the center of mass system is governed by the time-independent Schrödinger equation

$$-\Delta\psi(k,x) + V(x)\psi(k,x) = k^2\psi(k,x).$$

Here $x\in R^2$, the potential $V(x)$ is real-valued, and k is a positive scalar.

Scattering solutions are defined by the Lippmann–Schwinger equation

$$\psi(k,\theta,x) = \exp(ik\theta\cdot x) + \int G(k,|x-y|)V(y)\psi(k,\theta,y)\,d^2y, \tag{1.1}$$

where θ denotes a unit vector in R^2 and the function G is a fundamental solution of $\Delta + k^2$. We take G to be

$$G(k,r) = -(i/4)H_0^{(1)}(kr),$$

where H_0 is the zero-order Hankel function and $r = |x|$.

In order to apply Fredholm theory, we multiply the Lippman–Schwinger equation by $|V(x)|^{1/2}$ and make the following definitions:

$$\xi(k,\theta,x) = |V(x)|^{1/2}\psi(k,\theta,x),$$
$$\xi^0(k,\theta,x) = |V(x)|^{1/2}\exp(ik\theta\cdot x),$$
$$V_{1/2}(y) = V(y)|V(y)|^{-1/2},$$
$$K(k)f(x) = \int |V(x)|^{1/2}G(k,|x-y|)V_{1/2}(y)f(y)\,d^2y.$$

With this notation, the Lippmann–Schwinger equation becomes

$$\xi(k,\theta,x) = \xi^0(k,\theta,x) + K(k)\xi(k,\theta,x). \tag{1.2}$$

For k bounded away from zero, we recall[16] the following result concerning the operator $K(k)$:

Proposition 1.1: Suppose $V\in L^2$ with $\int\int |V(x)V(y)|\,|x-y|^{-1}\,d^2x\,d^2y = M < \infty$. Then for each $k_0 > 0$ the estimate $\|K(k)\|_{\text{H.S.}} \leqslant ck^{-1/2}$ holds for $k > k_0$, where c depends only on k_0 and on V.

Henceforth we will usually assume that V belongs to $L^1\cap L^2$ because[16] this assumption allows us to apply Fredholm theory to (1.2); we obtain a unique solution $\xi(k,\theta,x)$ provided the operator K does not have the eigenvalue 1. Note that for k large enough, the operator norm of $K(k)$ is less than 1, which certainly implies that (1.2) is uniquely solvable (by iteration, in fact).

We recall[16] that for V belonging to $L^1\cap L^2$ with $\int |V(x)|\,|x|^4\,d^2x < \infty$, the large x behavior of scattering states is given by

$$\psi(k,\theta,x) = \exp(ik\theta\cdot x) + \exp(-3\pi i/4)(8\pi)^{-1/2}A(k,\hat{x},\theta)\times\exp(ik|x|)(k|x|)^{-1/2} + h(k,\theta,x), \tag{1.3}$$

where $\hat{x} = x/|x|$,

$$A(k,\theta,\theta') = \int \exp(-ik\theta\cdot x)V(x)\psi(k,\theta',x)\,d^2x, \tag{1.4}$$

and

$h(k,\theta,x)\in L^2(x)$ uniformly in θ.

The quantity $A(k,\theta,\theta')$ is called the *scattering amplitude*; it essentially gives us the large x behavior of the wave function. We let the scattering amplitude act on $L^2(S^1)$ via $(A(k)f)(\theta') = \int_{S^1} A(k,\theta,\theta')f(\theta')\,d\theta$; the operator $A(k)$ is then bounded[16] and linear on $L^2(S^1)$. We also define the *scattering operator* or *S matrix* $S(k)$ on $L^2(S^1)$ by

$$S(k) = I - i(4\pi)^{-1}(\text{sgn } k)A(k).$$

In later sections we will also need the following information on the large k behavior of ψ.

Lemma 1.2: Let $V\in L^2\cap W^{2,1}$ and suppose that for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with

$$\int_0^\infty F(r)r\,dr \leqslant c\|V\|_{2,1},$$

and for some $\epsilon > 0$, $F(r) < Mr^{-1+\epsilon}$ near $r = 0$. Let k_0 be so large that, for $k > k_0$, $\|K(k)\| \leqslant a < 1$. Then for $k > k_0$, we have the estimate

$$|\psi(k,\theta,x) - \exp(ik\theta\cdot x)| \leqslant c\,k^{-(1+\epsilon/2)/2},$$

where c depends only on k_0 and V.

Proof: See Appendix A.

2. BEHAVIOR AT $k = 0$

Since the kernal of the operator $K(k)$ contains a Hankel function with a logarithmic divergence at the origin, one might expect the operator $K(k)$ and the wave function $\psi(k)$ to diverge logarithmically in some sense at the origin. However, as we shall see, the logarithmic divergence of $K(k)$ is due entirely to a rank-1 piece, and this prevents $\psi(k)$ from diverging at $k = 0$.

We recall[16] that properties of the Hankel function allow us to write $K(k) = L(k) + P \log k$, where

$$L(k)f(x) = \frac{-i}{4}\int |V(x)|^{1/2} \times\left(H_0^{(1)}(k|x-y|) - \frac{2i}{\pi}\log k\right) \times V_{1/2}(y)f(y)\,d^2y,$$

$$Pf(x) = (2\pi)^{-1}|V(x)|^{1/2}(V_{1/2},f).$$

If V is in L^1 with $\int|V(x)|\,|x|\,d^2x$ and $\int\int|V(x)V(y)|\,|\log|x-y|\,|^2\,d^2x\,d^2y$ finite, then L is a Hilbert–Schmidt operator and is well behaved at $k = 0$.

To investigate the behavior of ψ for k near zero, we will need the followimg lemma and its corollary:

Lemma 2.1: Suppose $V \in L^1$ with $\int|x|^{2\alpha}|V(x)|\,d^2x$ finite for some $0 < \alpha < 1$. Then $\|(\exp(ik\theta\cdot x) - 1)|V|^{1/2}\|_2 < ck^\alpha$ for k near zero.

Proof: Note that $\exp(ik\theta\cdot x) - 1 = (k\theta\cdot x)^\alpha h_\alpha(k\theta\cdot x)$, where $h_\alpha(it) = (\exp it - 1)t^{-\alpha}$. The function h is bounded because it is continuous and decays to zero for both large and small t. Thus

$$\|(\exp(ik\theta\cdot x) - 1)|V|^{1/2}\|_2^2 = \int(k\theta\cdot x)^{2\alpha}h_\alpha^2(ik\theta\cdot x)|V(x)|\,d^2x \leqslant k^{2\alpha}c\int|x|^{2\alpha}|V(x)|\,d^2x. \quad \text{QED}$$

Corollary 2.2: Suppose $V \in L^1 \cap L^2$ with $\int|x|^{2\alpha}|V(x)|\,d^2x$ finite for some $0 < \alpha < 1$, and suppose $(I - L(0))^{-1}$ exists. Then for $\xi_0(k) = \exp(ik\theta\cdot x)|V(x)|^{1/2}$ and for k near zero,

$$\|(I - L(k))^{-1}\xi^0(k) - (I - L(k))^{-1}|V|^{1/2}\|_2 \leqslant ck^\alpha.$$

Proposition 2.3: Let $V \in L^1 \cap L^2$ with $\int|x|\,|V(x)|\,d^2x < \infty$, and suppose $(I - L(0))^{-1}$ exists. Then $\xi(k)$ satisfies

$$\xi(k) = (I - L(k))^{-1}\xi^0(k) + \frac{(V_{1/2},(I - L(k))^{-1}\xi^0(k))\log k}{2\pi - (V_{1/2},(I - L(k))^{-1}|V|^{1/2})\log k} \times (I - L(k))^{-1}|V|^{1/2}. \tag{2.1}$$

The L^2 norm is bounded by

$$\|\xi(k)\|_2 \leqslant \begin{cases} c(\log k)^{-1} & \text{if } a_0 \neq 0, \\ c & \text{if } a_0 = 0, \end{cases}$$

where $a_0 = (V_{1/2},(I - L(0))^{-1}|V|^{1/2})$.

Proof: We shall solve the equation $(I - K(k))\xi = \xi^0$ assuming that $(I - L(k))^{-1}$ exists in a neighborhood of $k = 0$. Rewriting the equation in terms of the operators P and L, we have

$$(I - L(k))\xi - (\log k)P\xi = \xi^0. \tag{2.2}$$

Since P is a rank-1 operator, it will turn out that $P\xi = a|V(x)|^{1/2}$, where the constant a is given by

$$a = (2\pi)^{-1}(V_{1/2},\xi). \tag{2.3}$$

We will determine a at the end of our calculation. In the meantime, writing $P\xi = a|V|^{1/2}$, we can solve Eq. (2.2):

$$\xi = (I - L(k))^{-1}[\xi^0 + a(\log k)|V|^{1/2}]. \tag{2.4}$$

It now remains to determine the value of a. To do this, we use (2.3) and (2.4):

$$\begin{aligned} a &= (2\pi)^{-1}(V_{1/2},(I - L(k))^{-1}[\xi^0 + a(\log k)|V|^{1/2}]) \\ &= (2\pi)^{-1}(V_{1/2},(I - L(k)^{-1}\xi^0) \\ &\quad + a(2\pi)^{-1}(\log k)(V_{1/2},(I - L(k))^{-1}|V|^{1/2}). \end{aligned}$$

Solving this linear equation for a gives

$$a = \frac{(V_{1/2},(I - L(k))^{-1}\xi^0)}{2\pi - (\log k)(V_{1/2},(I - L(k))^{-1}|V|^{1/2})}.$$

Substitution of this value for a back into our expression for ξ, (2.4), gives us (2.1).

We now compute the limit as $k \to 0$ of (2.1). We write $\xi^0 = |V|^{1/2} + (\xi^0 - |V|^{1/2})$; by Corollary 2.2, the inner product in the numerator of (2.1) is $(V_{1/2},(I - L(k))^{-1}|V|^{1/2})$ plus something that decays like k^α as $k \to 0$. In the limit as $k \to 0$, we may therefore replace the ξ^0 by $|V|^{1/2}$. We recall[16] that differentiability of $(I - L(k))^{-1}$ allows us to write $(V_{1/2},(I - L(k))^{-1}|V|^{1/2}) = a_0 + a_1k$, where

$$a_0 = (V_{1/2},(I - L(0))^{-1}|V|^{1/2})$$

and a_1 is bounded for small k. With this in mind we compute

$$\xi(0) = \begin{cases} (I - L(0))^{-1}|V|^{1/2} & \text{if } a_0 = 0, \\ 0 & \text{if } a_0 \neq 0. \end{cases}$$

We will also need a bound for $\|\xi(k)\|_2$. Equation (2.1) yields

$$\xi(k) = (I - L(k))^{-1}(\xi^0(k) - |V|^{1/2}) + (I - L(k))^{-1}|V|^{1/2} \times\left(1 + \frac{[(V_{1/2},(I - L(k))^{-1}|V|^{1/2}) + (V_{1/2},(I - L(k))^{-1}(\xi^0 - |V|^{1/2}))]\log k}{2\pi - (V_{1/2},(I - L(k))^{-1}|V|^{1/2})\log k}\right).$$

Corollary 2.2 then gives

$$\|\xi(k)\|_2 \leqslant ck + \|(I - L(k))^{-1}\|\,\|V\|_1\frac{2\pi + ck^\alpha\log k}{2\pi + (a_0 + a_1k)\log k} \leqslant \begin{cases} c(\log k)^{-1} & \text{if } a_0 \neq 0 \\ c & \text{if } a_0 = 0. \end{cases} \quad \text{QED}$$

3. SYMMETRY AND ANALYTIC CONTINUATION

So far the wave function ψ has been defined only for positive k—the speed of the incoming particle. However, the Lippman–Schwinger equation makes sense for other values of k as well.

Invariance of the fundamental solution and of the plane wave under simultaneous complex conjugation and substitution of $-k$ for k shows that the wave function satisfies

$$\overline{\psi(-k,\theta,x)} = \psi(k,\theta,x).$$

This equation defines the wave function for negative k.

Similarly there is a relation between the incoming and outgoing waves

$$\psi^-(k,\theta,x) = \psi(-k,-\theta,x).$$

We will also need the reciprocity theorem, which is an expression of time reversal invariance of the scattering process.

Proposition 3.1 (Reciprocity Theorem): Let $V \in L^1 \cap L^2$. Then $A(k,\theta,\theta') = A(k,-\theta',-\theta)$.

Proof: See Appendix B.

Next we turn to the analyticity properties of the wave function as a function of k, which we now consider as a complex variable. Since we obtain the wave function only by means of the Lippman–Schwinger equation, we must analytically continue the integral equation.

First we note that the operator GV is Hilbert–Schmidt in the open upper half k-plane.

Proposition 3.2: Let $V \in L^2$. Then for $\operatorname{Im} k > 0$ the operator $G(k)V$ given by $G(k)Vf(x) = (-i/4)\int H_0(k|x-y|)V(y) f(y)\, d^2y$ is Hilbert–Schmidt, and $\|G(k)V\|_{\text{H.S.}} \leqslant c|k|^{-1}$.

Proof:

$$\begin{aligned}\|GV\|^2_{\text{H.S.}} &= c\int\int |H_0(k|x-y|)V(y)|^2\, d^2x\, d^2y \\ &= c\|V\|_2^2 \int |H_0(k|z|)|^2\, d^2z \\ &= c\|V\|_2^2 |k|^{-2} \int |H_0(|z'|k/|k|)|^2\, d^2z' \\ &< c|k|^{-2}. \qquad \text{QED}\end{aligned}$$

Similarly we have:

Proposition 3.3: Let $V \in L^2$. Then the operator $K(k)$ (defined in Sec. 1) is Hilbert–Schmidt for $\operatorname{Im} k \geqslant 0$, $k \neq 0$.

Proof: The proof is similar to that of Proposition 3.2. QED

However, the inhomogeneity in Eq. (1.2) is not in L^2 for $\operatorname{Im} k > 0$; we multiply the equation by $\exp(-ik\theta\cdot x)$ to obtain

$$\chi(k,\theta,x) = |V(x)|^{1/2} + \mathscr{K}(k)\chi(k,\theta,x),$$

where

$$\chi(k,\theta,x) = |V(x)|^{1/2}\psi(k,\theta,x)\exp(ik\theta\cdot x)$$

and where $\mathscr{K}(k)$ depends on θ:

$$\mathscr{K}(k)f(x) = \frac{-i}{4}\int |V(x)|^{1/2} H_0^{(1)}(k|x-y|)V_{1/2}(y) \times \exp(-ik\theta\cdot(x-y))f(y)\, d^2y.$$

Proposition 3.4: Let $V \in L^2$ with

$$\int\int \frac{|V(x)V(y)|}{|x-y|}\, d^2x\, d^2y < \infty.$$

Then the operator $\mathscr{K}(k)$ defined above is Hilbert–Schmidt for $\operatorname{Im} k \geqslant 0$, $k \neq 0$, and satisfies $\|\mathscr{K}(k)\|_{\text{H.S.}} \leqslant c|k|^{-1/2}$.

Proof: We apply the definition of the Hilbert–Schmidt norm to the operator $\mathscr{K}$:

$$\|\mathscr{K}(k)\|^2_{\text{H.S.}} = c\int\int |V(x)V(y)|\, |H_0^{(1)}(k|x-y|)|^2 \times \exp(2\operatorname{Im} k\theta\cdot(x-y))\, d^2x\, d^2y = c(I_1 + I_2),$$

where I_1 and I_2 are the integrals over the sets $|k|\,|x-y| < 1$ and $|k|\,|x-y| > 1$, respectively.

First we consider I_1. We use the small-argument behavior of the Hankel function to bound I_1 by

$$I_1 \leqslant \int\int_{|k|\,|x-y|<1} |V(x)V(y)||\log k|x-y||^2 \times \exp(2\operatorname{Im} k\theta\cdot(x-y))\, d^2x\, d^2y.$$

Next we let $z = x - y$ and note that for $|kz| < 1$, $\operatorname{Im} k\theta\cdot z \leqslant |kz| < 1$. Thus we have

$$\begin{aligned}I_1 &\leqslant c\int\int_{|kz|<1} |V(z+y)V(y)|\, |\log k|z||^2\, d^2z\, d^2y \\ &< c\|V\|_2^2 \int_{|kz|<1} |\log k|z||^2\, d^2z < c\|V\|_2^2 |k|^{-2}.\end{aligned}$$

We now turn to I_2. We use the large-argument behavior of the Hankel function to bound I_2 by

$$I_2 \leqslant \int\int_{|k|\,|x-y|>1} |V(x)V(y)|\exp(-2\operatorname{Im} k(|x-y| + \theta\cdot(x-y)))(|k|x-y||)^{-1}\, d^2x\, d^2y.$$

Note that the coefficient of $-2\operatorname{Im} k$ in the exponent is always positive; thus the exponential is bounded by 1. Use of this fact gives us

$$I_2 \leqslant |k|^{-1}\int\int \frac{|V(x)V(y)|}{|x-y|}\, d^2x\, d^2y < c|k|^{-1}. \qquad \text{QED}$$

Corollary 3.5: Let $V \in L^1 \cap L^2$. Then, for each θ, $\chi(k,\theta,x) = |V(x)|^{1/2}\psi(k,\theta,x)\exp(ik\theta\cdot x)$ is a meromorphic L^2-valued function of k for $\operatorname{Im} k > 0$.

Remark 3.6: A similar argument shows that, for $V \in L^1 \cap L^2$ and for each θ, $\chi^-(k,\theta,x) = |V(x)|^{1/2}\psi^-(k,\theta,x)\exp(-ik\theta\cdot x)$ is a meromorphic L^2-valued function of k for $\operatorname{Im} k < 0$.

4. AGMON'S SPECTRAL THEORY RESULTS

In this section we shall quote various results of Agmon[15] that will be used in the next section.

Let $L^{2,s}(R^2)$ denote the space of complex-valued functions $u(x)$ on R^2 with $(1+|x|^2)^{s/2}u(x) \in L^2(R^2)$, and let the weighted Sobolev spaces $H^{m,s}$ consist of $L^{2,s}$ functions with the first m derivatives also in $L^{2,s}$.

Agmon proves the following three theorems:

Theorem 4.1: Let $H = -\Delta + V$, where $V \in L^2_{\text{loc}}$ with $V(x) = O(|x|^{-1-\epsilon})$ as $|x| \to \infty$. Consider the resolvent $(H-E)^{-1}$ as an analytic operator-valued function on

$\mathbf{C}\backslash\sigma(\mathrm{H})$ with values in $B(L^{2,s},H^{2,-s})$ for any $S>\frac{1}{2}$. Then for real $E\neq 0$, the limits

$$\lim_{\epsilon\to 0}(H-E\pm i\epsilon)^{-1}$$

exist in the uniform operator topology of $B(L^{2,s},H^{2,-s})$.

Theorem 4.2: Let $H=-\Delta+V$, where $V\in L^2_{\mathrm{loc}}$ with $V(x)=O(|x|^{-3/2-\epsilon})$ as $|x|\to\infty$. Then there exist two families $\phi_\pm(k,\theta,x)$ of generalized eigenfunctions of H such that for every fixed k and θ, $\phi_\pm(k,\theta,x)$ as a function of x belongs to $C(R^2)\cap H^2_{\mathrm{loc}}(R^2)$ and satisfies the Schrödinger equation. Furthermore, for almost all $\theta\in S^1$, $\phi_\pm$ satisfies

$$\phi_\pm(k,\theta,x)-\exp(ik\theta\cdot x)$$

$$=-\lim_{\epsilon\to 0}(H-k^2\pm i\epsilon)^{-1}(V(x)\exp(ik\theta\cdot x)). \tag{4.1}$$

The eigenfunctions ϕ are continuous in k,θ, and x.

Theorem 4.3: Let $H=-\Delta+V$ where $V\in L^2_{\mathrm{loc}}$ with $V(x)=0(|x|^{-3/2-\epsilon})$ as $|x|\to\infty$, and let $\phi_\pm$ be the above family of generalized eigenfunctions. Let $P_{(a^2,b^2)}$ for $a>0$ denote the usual spectral projection. Then for any $f\in L^2$,

$$(P_{(a^2,b^2)}f)(x)=(2\pi)^{-2}\int_a^b\int_{S^1}\phi_\pm(k,\theta,x)$$
$$\times\int\overline{\phi_\pm(k,\theta,y)}f(y)\,d^2y\,d\theta\,k\,dk.$$

We must now relate Agmon's generalized eigenfunctions $\phi_\pm$ to our wave functions $\psi_\pm$. We first obtain a relation between the full and free resolvents by multiplying the relation $-\Delta+V+E=(-\Delta+E)+V$ on the left by $(-\Delta+E)^{-1}$ and on the right by $(-\Delta+V+E)^{-1}$. This gives us the relation

$$(-\Delta+E)^{-1}=(-\Delta+V+E)^{-1}+(-\Delta+E)^{-1}$$
$$\times V(-\Delta+V+E)^{-1}.$$

Multiplication on the left by $(I+(-\Delta+E)^{-1}V)^{-1}$ gives us

$$(-\Delta+V+E)^{-1}=(I+(-\Delta+E)^{-1}V)^{-1}$$
$$\times(-\Delta+E)^{-1}.$$

In Agmon's notation this is

$$-(H-k^2\pm i\epsilon^2)^{-1}=(I-G(\mp k+i\epsilon)V)^{-1}$$
$$\times G(\mp k+i\epsilon).$$

Upon composition with the multiplication operator $V_{1/2}$, this is

$$-(H-k^2\pm i\epsilon^2)^{-1}V_{1/2}$$
$$=(I-GV)^{-1}|V|^{-1/2}|V|^{1/2}GV_{1/2}$$
$$=|V|^{-1/2}(I-K(\mp k+i\epsilon))^{-1}K(\mp k+i\epsilon). \tag{4.2}$$

The formula (4.1) can then be expressed as (4.2) applied to ξ^0,

$$\phi_\pm(k,\theta,x)$$
$$=\exp(ik\theta\cdot x)-\lim_{\epsilon\to 0}(H-k^2\pm i\epsilon^2)^{-1}(V(x)\exp(ik\theta\cdot x))$$
$$=(|V(x)|^{-1/2}I+|V(x)|^{-1/2}(I-K^\mp)^{-1}K^\mp)\xi^0$$
$$=|V(x)|^{-1/2}(I+(I-K^\mp)^{-1}K^\mp)\xi^0$$
$$=|V|^{1/2}(I-K^\mp)^{-1}\xi^0$$
$$=|V|^{1/2}\xi^\mp(k,\theta,x)$$
$$=\psi^\mp(k,\theta,x).$$

5. THE SCATTERING OPERATOR

In this section we investigate some of the properties of the scattering operator.

The Marchenko method of inverse scattering rests on the following theorem (see Ref. 17).

Theorem 5.1: Let $V\in L^2_{\mathrm{loc}}$ with $V(x)=O(|x|^{-2-\epsilon})$ as $|x|\to\infty$. Let the scattering amplitude act on $L^2(S^1)$ via

$$A(k)f(\theta)=\int_{S^1}A(k,\theta',\theta)f(\theta')\,d\theta',$$

and let the scattering operator $S(k)$ be defined as an operator on $L^2(S^1)$ by

$$S(k)=I-i(4\pi)^{-1}\operatorname{sgn}k\,A(k). \tag{5.1}$$

Then

$$S(k)\psi^-(k,\theta,x)=\psi^+(k,\theta,x). \tag{5.2}$$

Remarks: The factor sgn k in (5.1) is needed to make (5.2) hold for negative k.

We shall show that the equality in (5.2) holds in the sense of $H^{2,-s}$ for some $s>\frac{1}{2}$; however, we recall (Theorem 4.2) that ψ^+ and ψ^- are continuous in x. Equation (5.2) therefore holds for each x.

Proof: We use Theorem 4.2 to write out the expression

$$\psi^+(k,\theta,x)-\psi^-(k,\theta,x)=\lim_{\epsilon\to 0}((H-k^2+i\epsilon)^{-1}$$
$$-(H-k^2-i\epsilon)^{-1})(V(x)\exp(ik\theta\cdot x)),$$

where the limit is in the $H^{2,-s}$ norm for some $x>\frac{1}{2}$. By Stone's formula,[18] the jump in the resolvent is the spectral projection, which in turn is given by Theorem 4.3:

$$\lim_{\epsilon\to 0}(2\pi i)^{-1}\int_{k_0}^{k_0+\delta}((H-k^2+i\epsilon)^{-1}$$
$$-(H-k^2-i\epsilon)^{-1})2k\,dk\,(V(x)\exp(ik\theta\cdot x))$$
$$=-P_{(k_0^2,(k_0+\delta)^2)}(V(x)\exp(ik\theta\cdot x))$$
$$=-(2\pi)^{-2}\int_{k_0}^{k_0+\delta}\int_{S^1}\psi^-(k,\theta',x)\int\overline{\psi^-(k,\theta',y)}V(y)$$
$$\times\exp(ik\theta\cdot y)\,d^2y\,d\theta'\,k\,dk.$$

Because of the symmetry properties of the wave function and scattering amplitude set forth in Sec. 3, the y integral is precisely $A(k,\theta',\theta)$. To remove the integration over k, we next multiply by $1/\delta$ and take the limit as δ approaches zero. Provided that the δ and ϵ limits are interchangeable, continuity in k gives us

$$\psi^+(k,\theta,x)-\psi^-(k,\theta,x)$$
$$=-i(4\pi)^{-1}\int_{S^1}\psi^-(k,\theta',x)A(k,\theta',\theta)\,d\theta'.$$

This proves the theorem, provided we can show that we may interchange the δ and ϵ limits. In other words, we must show that the following expression approaches zero as δ goes to zero:

$$I(\delta)$$
$$= \left\| \lim_{\epsilon\to 0}\left(\frac{2}{\delta}\right)\int_{k_0}^{k_0+\delta} ((H-k^2-i\epsilon)^{-1}-(H-k^2+i\epsilon)^{-1}) \times (V(x)\exp(ik\theta\cdot x))k\,dk \right.$$
$$\left. - 2k\lim_{\epsilon\to 0}((H-k^2-i\epsilon)^{-1}-(H-k^2+i\epsilon)^{-1}) \times (V(x)\exp(ik\theta\cdot x))\right\|_{2,2,-s},$$

where the norm is the $H^{2,-s}$ norm. We write the second term as the integral of a constant vector times $1/\delta$. Then continuity of the norm allows us to bring the ϵ limit outside; since the k integral is a limit of sums, the triangle inequality tells us that we can only increase $I(\delta)$ by bringing the norm inside the integral. We shall consider only I_1, the $-\epsilon$ term; the $+\epsilon$ term is similar. We have

$$\begin{aligned} I_1(\delta) &\leqslant \frac{2}{\delta}\lim_{\epsilon\to 0}\int \{\|((H-k^2-i\epsilon)^{-1} \\ &\quad -(H-k_0^2-i\epsilon)^{-1})(V(x)\exp(ik\theta\cdot x))\|_{2,2,-s} \\ &\quad +\|(H-k_0^2-i\epsilon)^{-1}(V(x)(\exp(ik\theta\cdot x) \\ &\quad -\exp(ik_0\theta\cdot x)))\|_{2,2,-s}\}k\,dk \\ &\leqslant 2\lim_{\epsilon\to 0}\max_{(k_0,k_0+\delta)}(\|(H-k^2-i\epsilon)^{-1} \\ &\quad -(H-k_0^2-i\epsilon)^{-1}\|\,\|V\|_{0,2,s} \\ &\quad +\|(H-k_0^2-i\epsilon)^{-1}\|\,\|V(x)(\exp(ik\theta\cdot x) \\ &\quad -\exp(ik_0\theta\cdot x))\|_{0,2,-s}), \end{aligned}$$

which goes to zero by continuity of the resolvent. [The hypothesis $V(x)=0(|x|^{-2-\epsilon})$ insures that $\|V(x)(\exp(ik\theta\cdot x) - \exp(ik_0\theta\cdot x))\|_{0,2,s}\to 0$.] QED

The following estimate on the scattering operator will allow us to define a Fourier transform in the next section:

Proposition 5.2: Let $V\in W^{2,1}$, and suppose that, for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with $\int_0^\infty F(r)r\,dr\leqslant c\|V\|_{2,1}$ and $F(r)<Mr^{-1+\epsilon}$ near 0 for some $\frac{1}{2}>\epsilon>0$. Then

$$\int_{-\infty}^{\infty}(\|(S(k)-I)f\|_2^{(\theta)})^2\,dk\leqslant c(\|f\|_2^{(\theta)})^2.$$

Proof: Turn to Appendix C.

6. INVERSE SCATTERING

This section is devoted to methods discovered by Newton[12] of extracting the potential $V(x)$ from the scattering amplitude $A(k,\theta,\theta')$.

We note first that the scattering amplitude does indeed uniquely determine the potential:

Theorem 6.1 (Uniqueness): Suppose the scattering amplitude $A(k,\theta,\theta')$ is constructed from a potential $V(x)$ belonging to $L^1\cap L^2$. Then the Fourier transform $\hat{V}$ can be recovered by means of the formula

$$\hat{V}(x)=\lim_{\substack{k\to\infty\\ k(\theta-\theta')=x}} A(k,\theta,\theta'). \tag{6.1}$$

This limit, an ordinary pointwise limit, is uniform in the sense that the difference

$$\hat{V}(k(\theta-\theta'))-A(k,\theta,\theta') \tag{6.2}$$

goes to zero uniformly in both angles as k goes to infinity.

Proof: We write out the definition of each term of (6.2) and apply the Schwarz inequality:

$$\begin{aligned} &|\hat{V}(k(\theta-\theta'))-A(k,\theta,\theta')| \\ &\quad=\left|\int \exp(ik\theta\cdot x)V(x)(\exp(ik\theta'\cdot x)-\psi(k,\theta',x))\,d^2x\right| \\ &\quad\leqslant\|V\|_1\|\,|V|^{1/2}(\psi(k,\theta',x)-\exp(ik\theta'\cdot x))\|_2. \end{aligned}$$

In the notation of Eq. (1.2), the second factor of this last expression is $\xi-\xi^0$. We write (1.2) as $\xi=(I-K)^{-1}\xi^0 = \xi^0+K(I-K)^{-1}\xi^0$. This allows us to bound (6.2) by

$$\begin{aligned} &|\hat{V}(k(\theta-\theta'))-A(k,\theta,\theta')| \\ &\quad\leqslant\|V\|_1\,\|K(k)\|\,\|(I-K(k))^{-1}\|\,\|V\|_1. \end{aligned} \tag{6.3}$$

By Proposition 1.1, the right side of (6.3) goes to zero as k becomes infinite. QED

Remark: Formula (6.1) is the well-known *Born approximation*. It gives a simple solution of the inverse scattering problem provided that the scattering amplitude is known for all k. In fact, this method of inversion depends exclusively on the high energy scattering data, which in practice may be known only approximately. There is, therefore, reason to investigate other inversion techniques, especially those whose dependence on high energy data might be less severe. One such technique is given in the following theorem.

Remark 6.2 (Notation): We shall use the following notation. We define the operator Q: $L^2(S^1)\to L^2(S^1)$ by $Qf(\theta)=f(-\theta)$. We use $\mathcal{F}$ for the vector-valued Fourier transform in k,

$$\mathcal{F}_k f(\alpha)=(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp(-ik\alpha)f(k)\,dk,$$

where for f belonging to $L^2(S^1)$, the limit inherent in the integral is taken in the norm topology.

We write $\beta(k,\theta,x)=\psi(k,\theta,x)\exp(-ik\theta\cdot x)$ and $\eta(\alpha,\theta,x)=\mathcal{F}_k(\beta(k,\theta,x)-1)$; we note that Lemma 1.2 implies that $\beta-1$ is a square-integrable $L^2(S^1)$-valued function of k, and that therefore η is a square-integrable $L^2(S^1)$-valued function of α.

We define the operator $\mathcal{S}(k)$: $L^2(S^1)\to L^2(S^1)$ by $\mathcal{S}(k)=\exp(ik\theta\cdot x)\,\overline{S(k)}\exp(-ik\theta\cdot x)$, where the exponentials act as multiplication operators and $\overline{S(k)}$ denotes the integral operator whose kernal is the complex conjugate of that of $S(k)$. For any f belonging to $L^2(S^1)$, we write $G(\alpha)f=\mathcal{F}_k^{-1}((\mathcal{S}(k)-I)f)$; by Proposition 5.2, $G(\alpha)f$ is a square-integrable $L^2(S^1)$-valued function of α. We note that G depends on x; this dependence will, however, be suppressed in what follows. Explicitly, $G(\alpha)$ is given by

$$\begin{aligned} &G(\alpha)f(\theta) \\ &\quad=(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp(ik(\alpha+\theta\cdot x))i(4\pi)^{-1}(\operatorname{sgn}k) \\ &\qquad\times\int_{S^1}\overline{A(k,\theta',\theta)}\exp(-ik\theta'\cdot x)f(\theta')\,d\theta'\,dk. \end{aligned} \tag{6.4}$$

Theorem 6.3 (The Marchenko Equation): Suppose $V\in W^{2,1}$ with $V(x)=O(|x|^{-2-\epsilon})$ as $|x|\to\infty$, and suppose that, for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all

bounded by a decreasing positive radial function $F(|x|)$ with $\int_0^\infty F(r)r\,dr<c\|V\|_{2,1}$ and $F(r)<Mr^{-1+\epsilon}$ near zero for some $0<\epsilon<\frac{1}{2}$. Suppose further that $-\Delta+V$ has no bound or half-bound states. Then in the notation of the above remark, the following equation in $L^2(R^+\times S^1)$ holds for positive α and for fixed x

$$\eta(\alpha,\theta,x)=\int_0^\infty G(\alpha+\beta)Q\eta(\beta,\theta,x)\,d\beta+G(\alpha)1. \quad (6.5)$$

Remark: The above hypotheses on V allow, for example, radial potentials with logarithmic singularities at x_0. Absence of bound states can be ascertained by means of the two-dimensional Levinson theorem.[16] Inversion in the presence of bound states can be accomplished with the use of further dimension-independent techniques developed by Newton.[12–14]

Proof: Theorem 5.1 gives us the relation

$$S(k)\psi^-(k,\theta,x)=\psi^+(k,\theta,x); \quad (6.6)$$

Sec. 3 allows us to eliminate ψ^- from (6.6):

$$S(k)Q\psi^+(-k,\theta,x)=\psi^+(k,\theta,x). \quad (6.7)$$

Note that $\psi^+(k)$ is analytic in the upper half-plane in k while $\psi^+(-k)$ is analytic in the lower one; (6.7) is therefore a Wiener–Hopf factorization problem or a Riemann–Hilbert problem. We shall solve the problem by using the Fourier transform to convert it into an integral equation.

In (6.7), we first put $-k$ in place of k and use the fact that $S(-k)=\overline{S(k)}$:

$$\overline{S(k)}Q\psi(k,\theta,x)=\psi(-k,\theta,x);$$

then multiplication by $\exp(ik\theta\cdot x)$ gives, in the notation of Remark 6.2,

$$\mathscr{S}(k)Q\beta(k,\theta,x)=\beta(-k,\theta,x). \quad (6.8)$$

In order to apply the Fourier transform to (6.8), we must subtract off the asymptotic values:

$$\beta(-k)-1=(\mathscr{S}(k)-I)Q(\beta(k)-1)+Q(\beta(k)-1)+(\mathscr{S}(k)-I)Q1. \quad (6.9)$$

Application of the inverse Fourier transform to (6.9) now gives

$$\eta(\alpha,\theta,x)=\int_{-\infty}^{\infty}G(\alpha-\beta)Q\eta(-\beta,\theta,x)\,d\beta+Q\eta(-\alpha,\theta,x)+G(\alpha)1. \quad (6.10)$$

Note that analyticity of $\beta(k)-1$ in the upper half k-plane implies that $\eta(\alpha,\theta,x)$ is zero for negative α. Consideration of positive α only in (6.10) gives us the Marchenko equation (6.5).

Theorem 6.4 (Compactness): Let $V\in W^{3,1}$ with $\int|x|^i|V(x)|\,d^2x<\infty$ for $i=1,2,3,4$, and suppose that for some x_0, $|V(x-x_0)|$, $|\nabla\nabla(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with $\int F(r)\,dr$ and $\int F(r)r^{3/2}\,dr$ finite. Suppose also that $(I-L(0))^{-1}$ exists. Then the integral operator $\mathscr{G}$ occuring in the Marchenko equation is a Hilbert–Schmidt operator on $L^2(R^+\times S^1)$.

Proof: The proof will be given in a later paper.

Remark 5.6: Newton has shown[13] that the spectrum of $\mathscr{G}$ is in fact contained in the interval $[-1,1]$. Thus, if $\mathscr{G}$ has neither the eigenvalue 1 nor -1, then $\mathscr{G}$ is a contraction and the Marchenko equation can be solved by iteration.

The above theorem allows us to apply Fredholm theory to the Marchenko equation, and, if the spectrum of $\mathscr{G}$ does not contain the point one, to obtain a solution $\eta(\alpha,\theta,x)$ belonging to $L^2(R^+\times S^1)$ for each x. We could then invert the Fourier transform to obtain the wave function, which could then be used in the formula

$$V(x)=[(\Delta+k^2)\psi(k,\theta,x)]/\psi(k,\theta,x).$$

However, the following formal calculation gives a simpler method of recovering the potential.

We use $\psi(k,\theta,x)=\beta(k,\theta,x)\exp(ik\theta\cdot x)$ in the Schrödinger equation, and find that the function $\beta(k,\theta,x)$ satisfies the equation

$$(\Delta+2ik\theta\cdot\nabla)\beta=V. \quad (6.11)$$

Into (6.11) we substitute

$$\beta(k,\theta,x)=1+\mathscr{F}_\alpha^{-1}(\eta(\alpha,\theta,x)),$$

obtaining

$$\int_0^\infty(\Delta-V(x)+2ik\theta\cdot\nabla)\eta(\alpha,\theta,x)\exp(ik\alpha)\,d\alpha-V(x)=0. \quad (6.12)$$

Formal integration by parts of the third term of (6.12) leads to

$$V(x)+2\theta\cdot\nabla_x\eta(0,\theta,x)+\int_0^\infty\exp(ik\alpha)\left[\Delta-V(x)-\frac{\partial}{\partial\alpha}\theta\cdot\nabla\right]\eta(\alpha,\theta,x)\,d\alpha=0.$$

For smooth η, the integral will go to zero for large k and leave us with

$$[\Delta_x-V(x)-(\partial/\partial\alpha)\theta\cdot\nabla_x]\eta=0,$$
$$V(x)=-2\theta\cdot\nabla_x\eta(0,\theta,x). \quad (6.13)$$

Equation (6.13) is known as the *miracle*. It is related to the characterization problem as follows. If the scattering amplitude with which we begin is known to come from a potential satisfying the hypotheses of Theorem 6.3, then the right side of (6.13) is guaranteed to be independent of θ. However, if we begin with an inadmissible scattering amplitude (i.e., one that does not correspond to a potential), then the miracle will not be satisfied (i.e., the right side of (6.13) will depend on θ). By counting variables, it is easy to see that most randomly chosen scattering amplitudes will not lead to a miraculous solution of (6.5). This is because the scattering amplitude, a function of three variables, is being used to determine the potential, which is a function of only two variables. At present, this miracle is the only known characterization of admissible scattering amplitudes.

ACKNOWLEDGMENTS

I am grateful to my thesis advisor, Roger G. Newton, for suggesting the problem and for discussing it with me on many occasions. I would also like to thank Joe Keller for reading the manuscript and making a number of helpful

comments. The work was supported in part by the Air Force Office of Scientific Research, the National Science Foundation, the Office of Naval Research, and the Army Research Office.

APPENDIX A: LARGE k BEHAVIOR OF ψ

Lemma 1.2: Let $V\in W^{2,1}\cap L^2$, and suppose that for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with $\int_0^\infty F(r)r\,dr\leqslant c\|V\|_{2,1}$ and $F(r)<Mr^{-1+\epsilon}$ near $r=0$, for some $\epsilon>0$. Let $k_0>0$ be so large that for $k>k_0$, $\|K(k)\|\leqslant a<1$. Then, for $k>k_0$, $|\psi(k,\theta,x)-\exp(ik\theta\cdot x)|\leqslant ck^{-(1+\epsilon/2)/2}$, where c depends only on V.

Proof: The wave function ψ is defined by Eq. (1.1). Provided that k is not an exceptional point, this equation has a solution with $\xi=|V|^{1/2}\psi\in L^2$. We split the integral in (1.1) into pieces corresponding to small and large arguments of the Hankel function:

$$\int H_0^{(1)}(k|x-y|)V(y)\psi(k,\theta,y)\,d^2y=I_1+I_2+I_3+I_4,$$

where

$$I_1=\frac{-2i}{\pi}\int_{|x-y|<k^{-1}}\log(k|x-y|)V(y)\psi(k,\theta,y)\,d^2y,$$

$$I_2=\int_{|x-y|<k^{-1}}\left[H_0^{(1)}(k|x-y|)+\frac{2i}{\pi}\log(k|x-y|)\right]V(y)\psi(k,\theta,y)\,d^2y,$$

$$I_3=\int_{|x-y|>k^{-1}}2^{1/2}(\pi k|x-y|)^{-1/2}\times\exp[ik|x-y|-\tfrac{1}{4}i\pi]V(y)\psi(k,\theta,y)\,d^2y,$$

$$I_4=\int_{|x-y|>k^{-1}}[H_0^{(1)}(k|x-y|)-2^{1/2}(\pi k|x-y|)^{-1/2}\times\exp(ik|x-y|-\tfrac{1}{4}i\pi)]V(y)\psi(k,\theta,y)\,d^2y.$$

Application of the Schwarz inequality to I_1 gives

$$|I_1|\leqslant\frac{2}{\pi}\left(\int_{|x-y|<k^{-1}}|\log k|x-y||^2|V(y)|d^2y\right)^{1/2}\|\xi\|_2. \tag{A1}$$

For $k>k_0$, the second factor of (A1) is bounded by

$$\|\xi\|_2\leqslant(1+\|K\|+\|K\|^2+\cdots)\|\xi^0\|_2\leqslant(1-a)^{-1}\|V\|_1, \tag{A2}$$

where the notation is as in Eq. (1.2). In the first factor of (A1), we let $x-y=r\phi$ with $\phi=(x-y)/|x-y|$ and $r=x-y$:

$$\int_{S^1}\int_0^{k^{-1}}|\log kr|^2|V(x-r\phi)|f\,dr\,d\phi\leqslant 2\pi\int_0^{k^{-1}}(kr)^{-1-\epsilon/2}F(|r-|x+x_0||)r\,dr\leqslant 2\pi k^{-1-\epsilon/2}\int_0^{k^{-1}}F(|r-|x+x_0||)r^{-\epsilon/2}\,dr. \tag{A3}$$

The integral converges if it converges when the singularities coincide; therefore (A3) is bounded by

$$ck^{-1-\epsilon/2}\int_0^{k^{-1}}r^{-1+\epsilon/2}\,dr\leqslant ck^{-1-\epsilon}.$$

Thus we have $|I_1|\leqslant c\|V\|_1k^{-(1+\epsilon)/2}$.

We treat I_2 the same way and obtain the same bound: $|I_2|\leqslant c\|V\|_1k^{-(1+\epsilon)/2}$.

Next we consider I_3. We replace $|V|^{1/2}\psi$ by

$$|V(y)|^{1/2}\psi(k,\theta,y)=|V(y)|^{1/2}\exp(ik\theta\cdot y)+K(k)[|V(y)|^{1/2}\psi(k,\theta,y)].$$

This splits I_3 into $I_3=I_5+I_6$, where

$$I_5=\int_{|x-y|>k^{-1}}2^{1/2}(\pi k|x-y|)^{-1/2}\times\exp[ik|x-y|-\tfrac{1}{4}i\pi+ik\theta\cdot y]V(y)d^2y, \tag{A4}$$

$$I_6=\int_{|x-y|>k^{-1}}2^{1/2}(\pi k|x-y|)^{-1/2}\exp[ik|x-y|-i\pi/4]\times V_{1/2}(y)K(k)[|V(y)|^{1/2}\psi(k,\theta,y)]\,d^2y. \tag{A5}$$

First we consider I_5. Letting $z=x-y$ in (A4) gives

$$I_5=\left(\frac{2}{\pi}\right)^{1/2}\exp\left(\frac{-i\pi}{4}\right)\times\int_{|z|>k^{-1}}\exp[ik|z|+ik\theta\cdot(x-z)]\times(k|z|)^{-1/2}V(x-z)\,d^2z.$$

With z written in polar coordinates as $z=r\hat{\phi}$, I_5 becomes

$$I_5=ck^{-1/2}\exp\left(\frac{-i\pi}{4}+ik\hat{\theta}\cdot x\right)\int_{k^{-1}}^\infty r^{1/2}\times\exp(ikr)\int_{S^1}\exp(-ikr\cos\phi)V(x-r\hat{\phi})\,d\hat{\phi}\,dr, \tag{A6}$$

where the unit vectors are now adorned with hats and ϕ is the angle between the vectors $\hat{\phi}$ and $\hat{\theta}$. We can now apply the stationary phase approximation (Appendix D) to the angular integral:

$$\int_{S^1}\exp(-ikr\cos\phi)V(x-r\hat{\phi})\,d\hat{\phi}=M(kr)^{-1/2}(aV(x-r\hat{\theta})+bV(x+r\hat{\theta}))+R,$$

where

$$|R|\leqslant M(kr)^{-1}\max_{\hat{\phi}\in S^1}\{|V(x-\hat{\phi})|,|\nabla V(x-r\hat{\phi})|,|\Delta V(x-r\hat{\phi})|\}.$$

We note that over the range of integration in (A6), we have $(kr)^{-1}<1$. This allows us to combine the leading term and remainder term:

$$|I_5|\leqslant ck^{-1}\int_{k^{-1}}^\infty F(|r-|x+x_0||)\,dr\leqslant ck^{-1}\|V\|_{2,1}.$$

Next we consider I_6. We apply the Schwarz inequality

$$|I_6|\leqslant\left(\frac{c}{k}\int_{|x-y|>k^{-1}}\frac{|V(y)|}{|x-y|}d^2y\right)^{1/2}\|K(k)(|V|^{1/2}\psi)\|_2\leqslant ck^{-1/2}\left(\int\frac{|V(x-z)|}{|z|}d^2z\right)^{1/2}\|K(k)\|\,\|\xi\|_2\leqslant ck^{-1},$$

where we have used the estimate $\|K(k)\|\leqslant ck^{-1/2}$.

Finally we consider I_4. Application of the Schwarz inequality and use of information about the asymptotic behavior of $H_0^{(1)}$ gives

$$|I_4| \leqslant \left(\int_{|x-y|>k^{-1}} c(k|x-y|)^{-3} |V(y)|\, d^2y \right)^{1/2} \| \, |V|^{1/2}\psi \|_2. \tag{A7}$$

Since $k|x-y|>1$, some of the factors of $k|x-y|$ in the denominator can be replaced by 1; we also use inequality (A2) to estimate the second factor of (A7).

$$\begin{aligned} |I_4| &< \Big(ck^{-1-\epsilon/2} \\ &\quad \times \int_{S^1} \int_{k^{-1}} |V(x+r\phi)| r^{-1-\epsilon/2} r\, dr\, d\phi \Big)^{1/2} \frac{\|V\|_1}{1-a} \\ &< ck^{-(1+\epsilon/2)/2} \left(\int_{k^{-1}}^{\infty} F(|r-|x+x_0||) r^{-\epsilon/2}\, dr \right)^{1/2} \|V\|_1 \\ &< c\|V\|_1 k^{-(1+\epsilon/2)/2}. \end{aligned}$$

APPENDIX B: THE RECIPROCITY THEOREM

Proposition 3.1: Let V belong to $L^1 \cap L^2$. Then $A(k,\theta,\theta') = A(k,-\theta',-\theta)$.

Proof: We recall that the scattering amplitude is given by $A(k,\theta,\theta') = \int \overline{\exp(ik\theta\cdot x)} V(x) \psi(k,\theta',x)\, d^2x$. We now use the Lippman–Schwinger equation (1.1) to write the exponential in terms of the wave functions:

$$\begin{aligned} A(k,\theta,\theta') &= \int \overline{\psi^-(k,\theta,x)} V(x) \psi^+(k,\theta',x)\, d^2x \\ &\quad - \iint \overline{G^-(k,|x-y|)} V(y) \overline{\psi^-(k,\theta,y)}\, d^2y \\ &\quad \times V(x)\psi^+(k,\theta',x)\, d^2x. \end{aligned}$$

Next we use the symmetry properties mentioned at the beginning of Sec. 3:

$$\begin{aligned} A(k,\theta,\theta') &= \int \psi^+(k,-\theta,x) V(x) \psi^+(k,\theta',x)\, d^2x \\ &\quad - \int \psi^+(k,-\theta,y) V(y) \int G^+(k,|x-y|) \\ &\quad \times V(x)\psi^+(k,\theta',x)\, d^2x\, d^2y. \end{aligned}$$

Again we use the Lippmann–Schwinger equation to obtain an exponential: $A(k,\theta,\theta') = \int \psi^+(k,-\theta,x) V(x)$ $\times \exp(ik\theta'\cdot x)\, d^2x = A(k,-\theta',-\theta)$. The interchange of x and y integration in the third step is justified by absolute convergence of the iterated integral:

$$\begin{aligned} &\iint |G^-(k,|x-y|) V(y) \psi^-(k,\theta,y)|\, d^2y \\ &\quad \times |V(x)\psi^+(k,\theta',x)|\, d^2x \\ &< \int |(K\xi)(k,\theta,x)|\, |\xi(k,\theta',x)|\, d^2x \\ &< \|K\|\, \|\xi\|_2^2. \end{aligned}$$

QED

APPENDIX C: STRONG SQUARE INTEGRABILITY OF $S-I$

Proposition 5.2: Let $V \in W^{2,1}$, and suppose that for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with

$$\int_0^{\infty} F(r) r\, dr < c\|V\|_{2,1}$$

and

$$F(r) < Mr^{-1+\epsilon} \quad \text{near zero,}$$

where $0<\epsilon<\frac{1}{2}$. Then

$$\int_{-\infty}^{\infty} (\|(S(k)-I)f\|_2^{(\theta)})^2\, dk < c(\|f\|_2^{(\theta)})^2. \tag{C1}$$

[The superscript θ reminds us that this is the $L^2(S^1)$ norm.]

Sketch of Proof (Details may be found in Cheney[a]): Let us fix $k_0 > \frac{1}{2}$, and split the left side of (C1) into small-k and large-k pieces.

The small-k piece is easy: the results of Sec. 2 show that $\|S(k)-I\|$ is bounded for $|k|<k_0$, which implies that

$$\int_{-k_0}^{k_0} \|(S(k)-I)f\|_2^2\, dk \leqslant c\|f\|_2^2.$$

Now we consider $|k|>k_0$. The difficulty we face is to extract from the integrand enough negative powers of k to make the k integral converge. In order to obtain explicit formulas, we write out the first few terms of the Born series:

$$\begin{aligned} |V|^{1/2}\psi &= (I-K)^{-1}(|V|^{1/2}\exp(ik\theta'\cdot x)) \\ &= (I+K+(I-K)^{-1}K^2)(|V|^{1/2}\exp(ik\theta'\cdot x)). \end{aligned} \tag{C2}$$

This allows us to write the kernel of $S(k)-I$ as

$$\begin{aligned} &(S(k)-I)(\theta',\theta) \\ &\quad = -i(4\pi)^{-1} \int \exp(-ik\theta\cdot x) V(x) \psi(k,\theta',x)\, d^2x \\ &\quad = D_1 + D_2 + D_3, \end{aligned} \tag{C3}$$

where

$$\begin{aligned} D_1(k,\theta',\theta) &= -i(4\pi)^{-1} \int \exp(-ik\theta\cdot x) V(x) \\ &\quad \times \exp(ik\theta'\cdot x)\, d^2x, \end{aligned} \tag{C4}$$

$$\begin{aligned} D_2(k,\theta',\theta) &= -i(4\pi)^{-1} \int \exp(-ik\theta\cdot x) V_{1/2}(x) \\ &\quad \times \frac{i}{4} \int |V(x)|^{1/2} \\ &\quad \times H_0(k|x-y|) V_{1/2}(y) \\ &\quad \times |V(y)|^{1/2} \exp(ik\theta'\cdot y)\, d^2y\, d^2x, \end{aligned} \tag{C5}$$

$$\begin{aligned} D_3(k,\theta',\theta) &= -i(4\pi)^{-1} \int \exp(-ik\theta\cdot x) V_{1/2}(x) \\ &\quad \times (I-K)^{-1}K^2 \\ &\quad \times (|V|^{1/2}\exp(ik\theta'\cdot x))\, d^2x. \end{aligned} \tag{C6}$$

Because we know from Sec. 1 that $\|K\|$ behaves like $|k|^{-1/2}$ for large k, it is fairly easy to see that

$$\int_{|k|>k_0} \|D_3 f\|_2^2\, dk \leqslant c\|f\|_2^2.$$

The terms corresponding to D_1 and D_2, however, require more work.

First we consider the part of the (C1) integral corresponding to D_1

$$16\pi^2\int_{|k|>k_0}\left|\left|\int_{S^1} D_1(k,\theta,\theta')f(\theta')\,d\theta'\right|\right|_2^2 dk$$
$$=\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int \exp(-ik(\theta'-\theta)\cdot x)$$
$$\times V(x)\,d^2x f(\theta')\,d\theta'\int_{S^1}\int \exp(ik(\theta''-\theta)\cdot y)$$
$$\times V(y)d^2y\,\overline{f(\theta'')}\,d\theta''\,d\theta\,dk. \tag{C7}$$

The absolute convergence of the θ integral allows us to do the θ integration first:

$$\int_{S^1}\exp(ik\theta\cdot(y-x))\,d\theta = 2\pi J_0(|k|\,|x-y|). \tag{C8}$$

Next we let $z=x-y$ in (C8) and use the asymptotic expansion for J_0 to split up (C7) into pieces corresponding to $|z|<|k|^{-1}$ and $|z|>|k|^{-1}$, respectively.

The piece corresponding to $|z|<|k|^{-1}$ is fairly easy because J_0 is bounded near the origin. We obtain the necessary k decay by using the inequality $1<|kz|^{-1}$ and by noting that the domain of z integration shrinks as k grows.

The piece of (C7) corresponding to $|z|>|k|^{-1}$ is harder to estimate. We shall consider in detail only the leading term of the J_0 asymptotic expansion; the remainder term already contains a factor of $(|kz|)^{-3/2}$ and is therefore easier to estimate. We write the leading term as

$$F=2\pi\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int_{|z|>|k|^{-1}}\int V(z+y)$$
$$\times 2^{1/2}(\pi|kz|)^{-1/2}\cos(|kz|-\tfrac{1}{4}\pi)$$
$$\times\exp(ik\theta'\cdot z)\,d^2z\,V(y)\exp(ik(\theta''-\theta')\cdot y)$$
$$\times d^2y f(\theta')\,d\theta'\,\overline{f(\theta'')}\,d\theta''\,dk. \tag{C9}$$

We let $z=r\hat{\phi}$ in the innermost integral of (C9); the z integral is then

$$\int_{|k|^{-1}}^{\infty}\int_{S^1} V(r\hat{\phi}+y)2^{1/2}(\pi|k|r)^{-1/2}$$
$$\times\cos(|k|r-\tfrac{1}{4}\pi)\exp(ik\cos\phi)\,d\hat{\phi}\,r\,dr,$$

where the unit vectors are now adorned with hats and ϕ is the angle between the vectors $\hat{\phi}$ and $\hat{\theta}'$.

Use of the stationary phase approximation (Lemma D.1) on the ϕ integral gives

$$\int_{S^1} V(r\hat{\phi}+y)\exp(ik\cos\phi)\,d\hat{\phi} = R+U(k,r,\hat{\theta}',y),$$

where

$$U(k,r,\hat{\theta}',y)=M_1(|k|r)^{-1/2}(aV(r\hat{\theta}'+y)+bV(-r\hat{\theta}'+y)) \tag{C10}$$

and

$$|R|\leqslant M_2(|k|r)^{-1}$$
$$\times\max_{\hat{\phi}\in S^1}\{|V(r\hat{\phi}+y)|,|\nabla V(r\hat{\phi}+y)|,|\Delta V(r\hat{\phi}+y)|\}. \tag{C11}$$

This application of the stationary phase approximation to (C9) gives

$$F=2\pi\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int_{|k|^{-1}}^{\infty} V(y)$$
$$\times\exp(ik(\theta''-\theta')\cdot y)2^{1/2}(\pi rk)^{-1/2}$$
$$\times\cos(|k|r-\tfrac{1}{4}\pi)(U(k,r,\theta',y)+R)$$
$$\times r\,dr\,d^2y f(\theta')\,d\theta'\,\overline{f(\theta'')}\,d\theta''\,dk, \tag{C12}$$

where we have once again dropped the hats on unit vectors. Next we split up the y integral in (C12): $F=F_1+F_2$, where

$$F_1=(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int_{|y|<|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty} V(y)$$
$$\times\exp(ik(\theta''-\theta')\cdot y)\cos(|k|r-\tfrac{1}{4}\pi)$$
$$\times(|k|r)^{-1/2}(U(k,r,\theta',y)+R)r\,dr\,d^2y f(\theta')$$
$$d\theta'\,\overline{f(\theta'')}\,d\theta''\,dk, \tag{C13}$$

$$F_2=(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int_{|y|>|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty}$$
$$\text{(same integrand)}. \tag{C14}$$

To estimate F_1, we use the bounds (C10) and (C11) in (C13) to obtain decay of $|k|^{-1}$. We obtain additional decay by using the hypotheses on the potential and by noting that the domain of y integration shrinks as k grows.

Next we consider F_2 [Eq. (C14)]. We split F_2 into pieces corresponding to integration over different parts of S^1. We write $S^1=S_<\cup S_>$, where $S_<$ corresponds to $|\theta'-\theta''|<|k|^{-1+2\epsilon}$ and $S_>$ corresponds to $|\theta'-\theta''|>|k|^{-1+2\epsilon}$. Thus $F_2=C_1+C_2$, where

$$C_1=(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S_<}\int_{|y|>|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty} V(y)$$
$$\times\exp(ik(\theta''-\theta')\cdot y)$$
$$\times(|k|r)^{-1/2}\cos(|k|r-\tfrac{1}{4}\pi)(U(k,r,\theta',y)+R)$$
$$\times r\,dr\,d^2y f(\theta')\,d\theta'\,\overline{f(\theta'')}\,d\theta''\,dk, \tag{C15}$$

$$C_2=(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S_>}\int_{|y|>|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty}$$
$$\text{(same integrand)}. \tag{C16}$$

First we consider C_1: (C10) and (C11) applied to (C15) give us

$$|C_1|\leqslant(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S_<}\int_{|y|>|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty}(|k|r)^{-1/2}|V(y)|$$
$$\times|(2\pi)^{1/2}(|k|r)^{-1/2}F(|r\theta'+x_0+y|)$$
$$+(2\pi)^{1/2}(|k|r)^{-1/2}F(|-r\theta'+x_0+y|)+4M_2$$
$$\times(|k|r)^{-1}F(|r\phi_0+x_0+y|)|r\,dr\,d^2y|f(\theta')|$$
$$\times d\theta'\,|f(\theta'')|\,d\theta''\,dk. \tag{C17}$$

Over the range of integration $r>|k|^{-1}$, we can bound $(|k|r)^{-1}$ by $(|k|r)^{-1/2}$. We also use the fact that $|\pm r\theta'+x_0+y|$ and $|r\phi_0+x_0+y|$ can be bounded below by $||x_0+y|-r|$ to simplify (C17); we obtain

$$|C_1| < c \int_{|k|>k_0} |k|^{-1} \int_{S^1} \int_{S_<} \int |V(y)| \times \int_{|k|^{-1}}^{\infty} F(||x_0+y|-r|)\, dr\, d^2y \times |f(\theta')|\, d\theta' |f(\theta'')|\, d\theta''\, dk.$$

Carrying out the r and y integrations gives us

$$|C_1| \leqslant \int_{|k|>k_0} |k|^{-1} \int_{S^1} \int_{S_<} \|V\|_1 \times |f(\theta')| d\theta' |f(\theta'')|\, d\theta''\, dk.$$

We next apply the Schwarz inequality to the θ' integral, obtaining

$$|C_1| \leqslant c \int_{|k|>k_0} |k|^{-1} \int_{S^1} \|f\|_2 \left(\int_{S_<} d\theta' \right)^{1/2} \times |f(\theta'')|\, d\theta''\, dk. \tag{C18}$$

The θ' integral of (C18) is the measure of the angle subtending the chord of length $|k|^{-1+2\epsilon}$ between the unit vectors θ' and θ''. It is not hard to show that the measure of the angle also behaves like $|k|^{-1+2\epsilon}$ for large k. This gives us the additional k-decay we need in order to show $|C_1| \leqslant c\|f\|_2^2$.

We now turn our attention to C_2 [Eq. (C16)]. The right side of (C16) contains two pieces, one corresponding to U and the other to R. By (C11), the term corresponding to R already contains a factor of $(|k|r)^{-3/2}$; in order to make both the r integral and the k integral converge, we replace $(|k|r)^{-3/2}$ by $(|k|r)^{-1-\epsilon/2}$. This trick disposes of the remainder term, and we are left with the term corresponding to U. This term we write as

$$C_3 = (8\pi)^{1/2} \int_{|k|>k_0} \int_{S^1} f(\theta') \times \int_{S_>} f(\theta'') \int_{|y|>|k|^{-\epsilon}} \int_{|k|^{-1}}^{\infty} V(y) \times \exp(ik(\theta''-\theta')\cdot y) \times (|k|r)^{1/2} \cos(|k|r - \pi/4) M_1(|k|r)^{-1/2} \times [aV(-r\theta + y) + bV(r\theta' + y)] \times r\, dr\, d^2y\, d\theta'\, d\theta''\, dk. \tag{C19}$$

We note that the y integral of (C19) can be done first because the inner two integrals (r and y) converge absolutely. The y integral of (C19) can then be evaluated by letting $y = s\phi$ and applying the stationary phase approximation to the ϕ integral as follows. For notational convenience we define

$$W(s\phi, r\theta') = V(s\phi)[aV(-r\theta' + s\phi) + bV(r\theta' + s\phi)]$$

and

$$\tilde{\theta} = (\theta' - \theta'')/|\theta' - \theta''|.$$

Then the y integral is

$$\int_{|k|^{-\epsilon}}^{\infty} \int_{S^1} W(s\phi, r\theta') \exp(iks|\theta' - \theta''| \cos\phi)\, d\phi\, s\, ds. \tag{C20}$$

Application of the stationary phase approximation (Lemma D.1) to (C20) gives us

$$\int_{|k|^{-\epsilon}}^{\infty} \{M_1(|k|s|\theta' - \theta''|)^{1/2} \times [aW(s\tilde{\theta}, r\theta') + bW(-s\tilde{\theta}, r\theta')] + R'\} s\, ds,$$

where

$$|R'| \leqslant M_2(|k|s|\theta' - \theta''|)^{-1} \times \max_{\phi \in S^1} \{|W(s\phi, r\theta')|, |\nabla W(s\phi, r\theta')|, |\Delta W(s\phi, r\theta')|\}.$$

We use this in (C19) to obtain

$$|C_3| \leqslant c \int_{|k|>k_0} |k|^{-1} \int_{S^1} |f(\theta')| \times \int_{S_>} |f(\theta'')| \times \int_{|k|^{-\epsilon}}^{\infty} \int_{|k|^{-1}}^{\infty} \{M_1(|k|s|\theta' - \theta''|)^{-1/2} \times |aW(s\tilde{\theta}, r\theta') + bW(-s\tilde{\theta}, r\theta')| + |R'|\} \times dr\, s\, ds\, d\theta'\, d\theta''\, dk. \tag{C21}$$

In (C21) we use the assumptions on the potential

$$|W(s\tilde{\theta}, r\theta')| \leqslant F(|s - |x_0||) F(|\,|x_0| - s| - r|).$$

A similar bound holds for ∇W and ΔW. This allows us to estimate the right side of (C21) by

$$|C_3| \leqslant c \int_{|k|>k_0} |k|^{-1} \int_{S^1} |f(\theta'')| \times \int_{S_>} |f(\theta')| \int_{|k|^{-\epsilon}}^{\infty} \times \int_{|k|^{-1}}^{\infty} [(|k|s|\theta' - \theta''|)^{-1/2} + (|k|s|\theta' - \theta''|)^{-1}] \times F(|s - |x_0|\,|) F(|\,|x_0| - s| - r|) \times dr\, s\, ds\, d\theta'\, d\theta''\, dk. \tag{C22}$$

We now carry out the r integration and use the fact that over the range of integration, we have $|k|s|\theta' - \theta''| > |k|^{\epsilon}$. We can therefore bound the right side of (C22) by

$$|C_3| \leqslant c \int_{|k|>k_0} |k|^{-1} \int_{S^1} |f(\theta'')| \int_{S_>} |f(\theta')| \times \int_{|k|^{-\epsilon}}^{\infty} F(s - |x_0|) |k|^{-\epsilon/2} \times s\, ds\, d\theta'\, d\theta''\, dk \leqslant c\|f\|_2^2.$$

We have now shown that the right side of (C7) is bounded by $c\|f\|_2^2$; in other words, we have disposed of the D_1 term. Next we must consider the D_2 term.

We write out the piece of the (C1) integral corresponding to D_2 [(C5)]

$$16\pi^2 \int_{|k|>k_0} \left| \left| \int_{S^1} D_2(k, \theta, \theta') f(\theta')\, d\theta' \right| \right|_2^2 dk = \int_{|k|>k_0} \int_{S^1} \int_{S^1} \int \int \frac{i}{4} V(x) H_0(|k|\,|x-y|) V(y) \times \exp(-ik[\theta'\cdot y - \theta\cdot x])\, y\, d^2x\, f(\theta')\, d\theta' \times \int_{S^1} \int \int (-i/4) V(z) \overline{H_0(|k|\,|z-w|)} V(w) \times \exp(-ik[\theta\cdot z - \theta''\cdot w])\, d^2w\, d^2z\, f(\theta'')\, d\theta''\, d\theta\, dk. \tag{C23}$$

In the right side of (C23), we make the substitutions $y' = x - y$ and $w' = z - w$, and note that the θ integral is absolutely convergent. The θ integral can therefore be done first:

$$\int_{S^1} \exp(ik\theta\cdot(z-x))\, d\theta = 2\pi J_0(|k|\,|z-x|),$$

and so (C23) is

$$\begin{aligned}
&16\pi^2 \int_{|k|>k_0} \|D_2 f\|^2\, dk \\
&\quad = \frac{\pi}{8}\int_{|k|>k_0}\int_{S^1} f(\theta')\int\int\int_{S^1} f(\theta'') \\
&\quad \times \int\int J_0(|k(z-x)|)V(x)H_0(|ky'|)V(x-y') \\
&\quad \times \exp[-ik\theta'\cdot(x-y')]\, d^2y'\, d^2x\, d\theta'\, V(z) \\
&\quad \times \overline{H_0(|kw'|)}V(z-w')\exp[-ik\theta''\cdot(z-w')] \\
&\quad \times d^2w'\, d^2z\, d\theta''\, dk. \qquad \text{(C24)}
\end{aligned}$$

We shall obtain the sought-after k-decay from the spatial integrals. We therefore estimate the y' (or w') integral of (C24) first; we write

$$\int H_0(|ky'|)V(x-y')\, d^2y' = I_1 + I_2,$$

where I_1 and I_2 correspond to integration over the sets $|ky'| < 1$ and $|ky'| > 1$, respectively.

Use of the small-argument behavior of H_0 to estimate I_1 gives

$$|I_1| \leqslant c\int_{|ky'|<1} |\log|ky'|\, V(x-y')|\, d^2y'. \qquad \text{(C25)}$$

We apply Hölder's inequality to (C25), obtaining

$$\begin{aligned}
|I_1| &\leqslant c\left(\int_{|y|<|k|^{-1}} |\log|ky'|\,|^{(1+\epsilon)/\epsilon}\, d^2y\right)^{\epsilon/(1+\epsilon)} \\
&\quad \times \left(\int_{|y'|<|k|^{-1}} |V(x-y')|^{1+\epsilon}\, d^2y'\right)^{(1+\epsilon)^{-1}} \\
&\leqslant c\left(\int_0^{|k|^{-1}} F(|\,|x+x_0|-|y'|\,|)^{1+\epsilon}|y'|d\,|y'|\right)^{(1+\epsilon)^{-1}}.
\end{aligned} \qquad \text{(C26)}$$

To (C26) we apply Lemma D.2:

$$\begin{aligned}
|I_1| &\leqslant c\left(2|k|^{-1}\int_0^{|k|^{-1}} F(r)^{1+\epsilon}\, dr\right)^{(1+\epsilon)^{-1}} \\
&\leqslant c(|k|^{-1-\epsilon^2})^{(1+\epsilon)^{-1}}.
\end{aligned}$$

Use of the large-argument asymptotic behavior of H_0 to estimate I_2 shows

$$\begin{aligned}
|I_2| &\leqslant c\int_{|y'|>|k|^{-1}} |ky'|^{-1/2}|V(x-y')|\, d^2y' \\
&\leqslant c|k|^{-1/2}\int_{|k|^{-1}}^{\infty} F(|x-y+x_0|)|y|^{1/2}\, d\,|y| \\
&\leqslant c|k|^{-1/2}.
\end{aligned}$$

Thus the y' integral of (C24) can be estimated for large k by

$$\left|\int H_0(|ky'|)V(x-y')\, d^2y'\right| \leqslant c|k|^{-1/2}.$$

This shows that the right side of (C24) is bounded by

$$\begin{aligned}
&c\int_{|k|>k_0} |k|^{-1}\|f\|_1^2 \\
&\quad \times \int\int |J_0(|k(z-x)|)|\,|V(x)|\,|V(z)|\, d^2x\, d^2z\, dk. \qquad \text{(C27)}
\end{aligned}$$

It remains to do the x and z integrals of (C27); to do this end, we let $z' = z - x$, and split the z' integral into pieces corresponding to integration over $|kz'| < 1$ and $|kz'| > 1$, respectively. We obtain extra k-decay in the small-argument piece because the domain of integration shrinks as $k \to \infty$. Decay is obtained in the large-argument integral from the $|kz'|^{-1/2}$ behavior of J_0 at infinity. QED

APPENDIX D: TECHNICAL LEMMAS

Lemma D.1 (Stationary phase approximation): Let $Q \in C^2(R^2)$. Then

$$\begin{aligned}
&\int_{S^1} Q(r\hat{\phi})\exp(ikr\cos\phi)\, d\phi \\
&\quad = M_1(|k|\,|r|)^{-1/2}(aQ(r\hat{\phi})\,|_{\phi=0} + bQ(r\hat{\phi})\,|_{\phi=\pi}) + R, \qquad \text{(D1)}
\end{aligned}$$

where

$$|R| \leqslant M_2(|k|\,|r|)^{-1}\max_{\hat{\phi}\in S^1}\{|Q(r\hat{\phi})|, |\nabla Q(r\hat{\phi})|, |\Delta Q(r\hat{\phi})|\}; \qquad \text{(D2)}$$

here $\hat{\phi} = (\cos\phi, \sin\phi)$, a and b are constants of modulus 1 and the M_i are positive constants independent of Q.

Proof: The proof follows Erdelyi.[19] Our first task is to split up the integral

$$I = \int_{S^1} Q(r\hat{\phi})\exp(ikr\cos\phi)\, d\phi \qquad \text{(D3)}$$

so that we consider only one stationary point at a time. To this end, we write $I = I_1 + I_2$, where I_1 is the integral over $[0,\pi]$, I_2 the integral over $[\pi,2\pi]$. First we consider I_1, which we split into $I_1 = A + B$, where

$$A = \int_0^{\pi} Q(r\hat{\phi})\exp(ikr\cos\phi)\eta(\phi)\, d\phi, \qquad \text{(D4)}$$

$$B = \int_0^{\pi} Q(r\hat{\phi})\exp(ikr\cos\phi)[1-\eta(\phi)]\, d\phi, \qquad \text{(D5)}$$

and where η is an infinitely differentiable cutoff function with

$$\begin{aligned}
\eta(\phi) &= 1 \quad \text{for } 0 \leqslant \phi \leqslant \pi/4, \\
&= 0 \quad \text{for } 3\pi/4 \leqslant \phi \leqslant \pi.
\end{aligned}$$

We consider A first. In (D4) we make the change of variable $t^2 = 1 - \cos\phi$;

$$\begin{aligned}
A &= \exp(ikr)\int_0^{2^{1/2}} Q(r\hat{\phi}) \\
&\quad \times \exp(-ikrt^2)\tilde{\eta}(t)2t\,[1-(1-t^2)^2]^{-1/2}\, dt, \qquad \text{(D6)}
\end{aligned}$$

where $\tilde{\eta}(t) = -\eta(\arccos(1-t^2))$. Integration by parts of (D6) [differentiation of $2Q(r\hat{\phi})\tilde{\eta}(t)(2-t^2)^{-1/2}$ and integration of $\exp(-ikrt^2)$] gives

$$\begin{aligned}
A &= (ikr)\Big[2Q(r\hat{\phi})\tilde{\eta}(t)h_1(t)(2-t^2)^{-1/2}|_0^{2^{1/2}} \\
&\quad - \int_0^{2^{1/2}} \frac{\partial}{\partial t}(2Q(r\hat{\phi})\tilde{\eta}(t)(2-t^2)^{-1/2})h_1(t)\, dt\Big], \qquad \text{(D7)}
\end{aligned}$$

where

$$h_1(t) = -\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right) \times \int_0^\infty \exp\left[-ikr\left(t+\sigma\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right)\right)^2\right] d\sigma.$$

To compute the first term of (D7), we need to evaluate h_1 at zero:

$$h_1(0) = -\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right)\int_0^\infty \exp[-|k|r\sigma^2]\,d\sigma = -\pi^{1/2}(|k|r)^{-1/2}\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right).$$

We substitute this expression into (D7) and recall that $\tilde\eta(2^{1/2}) = 0$. Equation (D7) is then

$$A = (2\pi)^{1/2}(|k|r)^{-1/2}\exp(ikr) \times \exp(-i\,\mathrm{sgn}\,k\,\pi/4)Q(r\hat\phi)|_{\phi=0} + R_1,$$

where

$$R_1 = -\exp(ikr) \times \int_0^{2^{1/2}} \frac{\partial}{\partial t}[2Q(r\hat\phi)\tilde\eta(t)(2-t^2)^{-1/2}]h_1(t)\,dt.$$

We have now found the leading term of (D1); our next task is to obtain the correct decay for the remainder. To this end, we integrate R_1 by parts; this gives us

$$R_1 = -\exp(ikr) \times \frac{\partial}{\partial t}[2Q(r\hat\phi)\tilde\eta(t)(2-t^2)^{-1/2}]h_2(t)|_0^{2^{1/2}} + R_2, \qquad \text{(D8)}$$

where

$$R_2 = \exp(ikr) \times \int_0^{2^{1/2}} \frac{\partial^2}{\partial t^2}[2Q(r\hat\phi)\tilde\eta(t)(2-t^2)^{-1/2}]h_2(t)\,dt \qquad \text{(D9)}$$

and where h_2, given by

$$h_2(t) = -\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right) \times \int_0^\infty \sigma\exp\left[-ikr\left(t+\sigma\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right)\right)^2\right] d\sigma, \qquad \text{(D10)}$$

is the primitive of h_1 satisfying

$$h_2(0) = -\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right)\int_0^\infty \sigma\exp(-|k|r\sigma^2)\,d\sigma = -(|k|r)^{-1}\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right). \qquad \text{(D11)}$$

To estimate $h_2(t)$ for $t>0$, we note that along the path of integration, the quantity

$$\begin{aligned} &-ikr(t+\sigma\exp(-i\,\mathrm{sgn}\,k\,\pi/4))^2 + |k|r\sigma^2 \\ &\quad = -ikr[t^2 + 2t\sigma\exp(-i\,\mathrm{sgn}\,k\,\pi/4) \\ &\qquad + (\mathrm{sgn}\,k)i\sigma^2 - i\,\mathrm{sgn}\,k\,\sigma^2] \\ &\quad = -ikrt\,[t+2\sigma\exp(-i\,\mathrm{sgn}\,k\,\pi/4)] \end{aligned}$$

has negative real part; thus

$$\exp[-ikr(t+\exp(-i\,\mathrm{sgn}\,k\,\pi/4)\sigma)^2] \leqslant \exp(-|k|r\sigma^2).$$

With this estimate, we have

$$|h_2(t)| < \int_0^\infty \sigma\exp(-|k|r\sigma^2)\,d\sigma = (|k|r)^{-1}.$$

With this information, a bound on R_1 can be obtained as follows. We write $\mu(t) = \tilde\eta(t)(2-t^2)^{-1/2}$. Then we compute the derivatives appearing in (D8) and (D9) [$' = (d/dt)$]:

$$\frac{\partial}{\partial t}[Q(r\hat\phi)\mu(t)] = Q(r\hat\phi)\mu'(t) + \nabla Q(r\hat\phi)\cdot\hat\phi\,'\mu(t)$$

and

$$\begin{aligned} &\frac{\partial^2}{\partial t^2}[Q(r\hat\phi)\mu(t)] \\ &\quad = Q(r\hat\phi)\mu''(t) + 2\nabla Q(r\hat\phi)\cdot\hat\phi\,'\mu'(t) \\ &\qquad + \nabla Q(r\hat\phi)\cdot\hat\phi\,''\mu(t) + \nabla Q(r\hat\phi)\|\hat\phi\,'\|^2\mu(t). \end{aligned}$$

Let

$$M_2 = 6\max_{0<t<(1+2^{-1/2})^{1/2}}\{|\mu'(t)|,|\mu''(t)|,|\hat\phi\,'\mu(t)|,|2\hat\phi\,'\mu'(t)|, |\hat\phi\,''\mu(t)|\ \|\hat\phi\,'\|^2|\mu(t)|\}.$$

Then

$$|R_1| \leqslant M_2(|k|r)^{-1} \times \max_{0<t<(1+2^{-1/2})^{1/2}}\{|Q(r\hat\phi)|,|\Delta Q(r\hat\phi)|,|\Delta Q(r\hat\phi)|\}.$$

This concludes the estimate for A; now for B, the change of variables $t^2 = \cos\phi + 1$ gives a similar estimate; and in I_2 the change of variables $\beta = \phi - \pi$ converts I_2 to an integral of the form I_1.

Lemma D.2: Let $F(r)$ be a positive nonincreasing function on $[0,b]$, $b>0$. Then for $a>0$ and $\alpha>0$,

$$\int_0^b F(|a-r|)r^\alpha\,dr \leqslant 2b^\alpha\int_0^b F(r)\,dr. \qquad \text{(D12)}$$

Proof: In the left side of (D12) we note that $r \leqslant b$, and then we use the definition of absolute value

$$\begin{aligned} I &= \int_0^b F(|a-r|)r^\alpha\,dr \leqslant b^\alpha\int_0^b F(|a-r|)\,dr \\ &= b^\alpha\int_0^{\min(a,b)} F(a-r)\,dr + b^\alpha\int_{\min(a,b)}^b F(r-a)\,dr. \end{aligned}$$

In the first integral let $s = a - r$; in the second let $s = r - a$. Then

$$I \leqslant b^\alpha\int_{a-\min(a,b)}^a F(s)\,ds + b^\alpha\int_{\min(a,b)-a}^{b-a} F(s)\,ds.$$

Case $a<b$:

$$I \leqslant b^\alpha\int_0^a F(s)\,ds + b^\alpha\int_0^{b-a} F(s)\,ds \leqslant 2b^\alpha\int_0^b F(s)\,ds.$$

Case $b<a$:

$$I \leqslant b^\alpha\int_{a-b}^a F(s)\,ds \leqslant b^\alpha\int_0^b F(s)\,ds.$$

[1] I. Kay and H. E. Moses, Nuovo Cimento **22**, 689 (1961).
[2] I. Kay and H. E. Moses, Comm. Pure Appl. Math. **14**, 435 (1961).
[3] L. D. Faddeev, Itogi Nauk Tekh. Sov. Probl. Mat. **3**, 93 (1974) [J. Sov. Math. **5**, 334 (1976)].

[4]R. G. Newton, *Scattering Theory in Mathematical Physics*, edited by J. A. Lavita and J.-P. Marchand (Reidel, Dordrecht, 1974).
[5]L. D. Faddeev, Dokl. Akad. Nauk SSSR **165**, 514 (1965) [Sov. Phys. Dokl. **10**, 1033 (1966)].
[6]L. D. Faddeev, Dokl. Akad. Nauk SSSR **167**, 69 (1966) [Sov. Phys. Dokl. **11**, 209 (1966)].
[7]R. T. Prosser, J. Math. Phys. **10**, 1819 (1969).
[8]R. T. Prosser, J. Math. Phys. **17**, 1775 (1976).
[9]R. T. Prosser, J. Math. Phys. **21**, 2635 (1980).
[10]C. Morawetz, Comp. Math. Appls. **7**, 319 (1981).
[11]P. Deift and E. Trubowitz, Comm. Pure. Appl. Math. **32**, 121 (1979).
[12]R. G. Newton, J. Math. Phys. **21**, 1698 (1980).
[13]R. G. Newton, J. Math. Phys. **22**, 2191 (1981).
[14]R. G. Newton, J. Math. Phys. **23**, 594 (1982).
[15]S. Agmon, Ann. Scuola Norm. Sup. Pisa, Ser. IV, **2**, 151 (1975).
[16]M. Cheney, "Two-dimensional scattering: the number of bound states from scattering data," J. Math. Phys. (in press).
[17]R. G. Newton, *Scattering Theory of Waves and Particles*, (Springer, New York, 1982), 2nd ed., p. 286.
[18]M. Reed and B. Simon, *Methods of Modern Mathematical Physics. I. Functional Analysis* (Academic, New York, 1972), p. 237.
[19]A. Erdelyi, *Asymptotic Expansions* (Dover, New York, 1965).

Eigenvalues and eigenfunctions associated with the Gel'fand–Levitan equation

Harry E. Moses[a)]
Center for Atmospheric Research, University of Lowell, Lowell, Massachusetts 01854

Reese T. Prosser
Department of Mathematics, Dartmouth College, Hanover, New Hampshire 03755

(Received 8 June 1983; accepted for publication 10 August 1983)

It is shown here that the solutions of the Gel'fand–Levitan equation for inverse potential scattering on the line may be expressed in terms of the eigenvalues and eigenfunctions of certain associated operators of trace class. The details are sketched for the case of rational reflection coefficients, and carried out for the simplest class of examples.

PACS numbers: 03.80. + r, 03.65.Nk

1. INTRODUCTION

The Gel'fand–Levitan equation plays a central role in solving inverse scattering problems in one dimension.[1] In the case where the problem involves a *scattering potential* $V(x)$ defined for $-\infty < x < +\infty$, for example, we know that $V(x)$ may be recovered from the *reflection coefficient* $r(k)$, defined for $-\infty < k < +\infty$, as follows: set

$$R(x,y) = \hat{r}(x+y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ikx} r(k) e^{-iky}\, dk, \quad (1)$$

and then solve for $K(x,y)$ the Gel'fand–Levitan equation

$$K(x,y) + R(x,y) + \int_{-\infty}^{x} K(x,z)R(z,y)dz = 0. \quad (2)$$

Then the potential $V(x)$ appears as

$$V(x) = 2\frac{d}{dx}K(x,x). \quad (3)$$

(See Ref. 2 for a general discussion of this procedure.)

In order to study the behavior of the solutions of (2), it is useful to consider the associated equation, to be solved for $K(x,y,w)$:

$$K(x,y,w) + R(x,y) + \int_{-\infty}^{w} K(x,z,w)R(z,y)dz = 0. \quad (4)$$

Evidently $K(x,y,x) = K(x,y)$. Now (4) may be expressed in operator form with w as a parameter:

$$K(w) + R + K(w)P(w)R = 0. \quad (5)$$

Here R, $K(w)$, and $P(w)$ are integral operators with kernels $R(x,y)$, $K(x,y,w)$, and $P(x,y,w)$, with

$$P(x,y,w) = \theta(w-x)\delta(x-y). \quad (6)$$

Here $\theta(z)$ is the Heaviside function, and $\delta(z)$ its derivative.

Now (4) yields

$$K(w)(I + P(w)R) = -R, \quad (7)$$

and hence, whenever $(I + P(w)R)$ is invertible,

$$K(w) = -R(I + P(w)R)^{-1}. \quad (8)$$

Now suppose that the reflection coefficient $r(k)$ is such that its Fourier transform $\hat{r}(z)$ is smooth and integrable. Then it follows that the operator $P(w)R$ is of *trace class* for each w, and

$$\operatorname{tr} P(w)R = \int_{-\infty}^{w} \hat{r}(2z)dz. \quad (9)$$

One can then define the *Fredholm determinant* $\Delta(w)$ of the operator $(I + P(w)R)$ by (cf. Ref. 3, p. 255ff)

$$\Delta(w) = \det(I + P(w)R) = \exp \operatorname{tr} \log(I + P(w)R). \quad (10)$$

Evidently

$$\log \Delta(w) = \operatorname{tr} \log(I + P(w)R) \quad (11)$$

and so

$$\frac{d}{dw}\log \Delta(w) = \frac{\Delta'(w)}{\Delta(w)} = \operatorname{tr} P'(w)R(I + P(w)R)^{-1} = -\operatorname{tr} P'(w)K(w). \quad (12)$$

Here we have used (8). But $P'(w)K(w)$ has kernel $\delta(w-x)K(x,y,w)$, so

$$-\operatorname{tr} P'(w)K(w) = -\int_{-\infty}^{w} \delta(w-x)K(x,x,w)dx = -K(w,w,w) = -K(w,w). \quad (13)$$

Hence by (3)

$$V(w) = 2\frac{d}{dw}K(w,w) = -2\frac{d^2}{dw^2}\log \Delta(w). \quad (14)$$

This formula, which gives V directly in terms of R, first appears in Ref. 4, and has since been rediscovered by several authors, including us.[5] In one sense, this formula by-passes the Gel'fand–Levitan equation, since it gives V directly in terms of R, and once V is known everything about the scattering problem is known, at least in principle.

In another sense (14) is no better than (4), since the calculation of the determinant $\Delta(w)$ of $(I + P(w)R)$ is not usually an easy matter in practice. One possible approach is to calcu-

a)Research Sponsored in part by AFOSR Grant No. 81-0253A.

 0022-2488/84/010108-05$02.50

late the *eigenvalues* $\lambda_n(w)$ of the operator $P(w)R$ and use them to calculate $\Delta(w)$:

$$\Delta(w) = \prod_{n=1}^{\infty} (1 + \lambda_n(w)). \tag{15}$$

We indicate here how this might be done in the case where the reflection coefficient $r(k)$ is a rational function of k. (This case has already been treated by other methods in Refs. 6 and 7.)

Accordingly, we assume now that $r(k)$ has the form

$$r(k) = p(-ik)/q(-ik), \tag{16}$$

where p and q are polynomials with real coefficients, chosen so that degree $p <$ degree q, and so that $r(k)$ is regular in the upper half k-plane. If $r(k)$ is to be a reflection coefficient, then we should require that $|r(k)| \leqslant 1$ and $r(0) = -1$, but these requirements will play no role in solving (4).

It follows from our assumptions that $R(x, y) = \hat{r}(x + y)$ vanishes if $x + y < 0$, and satisfies an ordinary differential equation if $x + y > 0$, of the form

$$q(D)R(x, y) = p(D)\delta(x + y), \tag{17}$$

where $D = d/dx$.

The eigenvalues $\lambda_n(w)$ of the trace-class operator $P(w)R$ are discrete and the corresponding eigenfunctions $\phi_n(w)$ satisfy

$$P(w)R\phi_n(w) = \lambda_n(w)\phi_n(w). \tag{18}$$

It follows that $\phi_n(w) = P(w)\phi_n(w)$ and, hence, that

$$R(w)\phi_n(w) \equiv P(w)RP(w)\phi_n(w) = \lambda_n(w)\phi_n(w). \tag{19}$$

Moreover, it is easy to verify from (1) that if $|r(k)| \leqslant M$, then the operator R^2 is positive and satisfies $0 \leqslant R^2 \leqslant M^2 I$. It follows that the same is true of $R(w)^2$. Hence we have

$$0 \leqslant \lambda_n^2(w) \leqslant M^2. \tag{20}$$

Since $R(w)$ is of trace class, we also have, after a suitable rearrangement,

$$M^2 \geqslant \lambda_1^2 \geqslant \lambda_2^2 \geqslant \cdots \geqslant \lambda_n^2 \downarrow 0, \tag{21}$$

$$\sum_{n=1}^{\infty} \lambda_n(w) = \operatorname{tr}(P(w)RP(w)) = \operatorname{tr}(R(w)), \tag{22}$$

$$\prod_{n=1}^{\infty} (1 + \lambda_n(w)) = \Delta(w). \tag{23}$$

Now Eq. (19) may be written, for $-\infty < x \leqslant w$,

$$\int_{-\infty}^{w} R(x + y)\phi_n(y, w)dy = \lambda_n(w)\phi_n(x, w). \tag{24}$$

Applying (17) to (24), we get, for $-w \leqslant x \leqslant w$,

$$\lambda_n(w)q(D)\phi_n(x, w) = \int_{-\infty}^{w} q(D)R(x + y)\phi_n(y, w)dy$$
$$= p(D)\phi_n(-x, w). \tag{25}$$

It follows that, for $-w \leqslant x \leqslant w$,

$$\lambda_n^2(w)q(-D)q(D)\phi_n(x, w) = \lambda_n(w)p(D)q(-D)\phi_n(-x, w)$$
$$= p(D)p(-D)\phi_n(x, w). \tag{26}$$

Thus we see that the eigenfunctions $\phi_n(x, w)$ of the operator $R(w) = P(w)RP(w)$ satisfy an ordinary differential equation of even order with constant coefficients. By inserting the known form of the solutions of this equation back into (24), we may determine the integration constants and the admissible values of $\lambda_n(w)$. Specifically, the solutions of (26) all have the form

$$\phi_n(x, w) = \sum_{j=1}^{m} (A_j e^{ik_j x} + B_j e^{-ik_j x}), \tag{27}$$

where the $\pm k_j$ are the $2m$ solutions of the equation

$$r(-k)r(k) = \lambda_n^2(w). \tag{28}$$

Here m is the degree of the polynomial q. Note that if k_j is a solution of this equation, then so is $-k_j$, and so is $\bar{k}_j$. We assume here that these solutions are all distinct, and that $\operatorname{Im}(+k_j) \geqslant 0$.

Now if we insert (27) back into (24), do the integration, and equate coefficients of the various resulting exponentials, we get $2m - 1$ equations relating the A_j and B_j, and one equation determining the admissible values of $\lambda_n(w)$ for given w. Details are presented in the next section.

Once the eigenvalues $\lambda_n(w)$ and eigenfunctions $\phi_n(x, w)$ of the operator $R(w) = P(w)RP(w)$ are known, then we can calculate the determinant $\Delta(w)$ by (15). Moreover, we can also calculate the kernel of $K(w)$, since if the eigenfunctions $\phi_n(w)$ are normalized by

$$\|\phi_n(w)\|_2 = 1, \tag{29}$$

then we have, for $-\infty < x, y \leqslant w$,

$$R(x, y, w) = \sum_{n=1}^{\infty} \lambda_n(w)\phi_n(x, w)\phi_n(y, w), \tag{30}$$

and so, by (8), for $-\infty < x, y \leqslant w$,

$$K(x, y, w) = \sum_{n=1}^{\infty} \frac{-\lambda_n(w)}{1 + \lambda_n(w)} \phi_n(x, w)\phi_n(y, w) \tag{31}$$

and

$$K(w, w, w) = \sum_{n=1}^{\infty} \frac{-\lambda_n(w)}{1 + \lambda_n(w)} \phi_n(w, w)^2. \tag{32}$$

But from (12) and (13) we have

$$K(w, w, w) = -\frac{d}{dw} \log \Delta(w)$$
$$= -\sum_{n=1}^{\infty} \frac{\lambda_n'(w)}{1 + \lambda_n(w)}. \tag{33}$$

Comparing (32) and (33), we see that when $x = w$, we have

$$\phi_n(w, w)^2 = \lambda_n'(w)/\lambda_n(w). \tag{34}$$

On the other hand, since $R(x + y) = 0$ if $x + y < 0$, we see from (24) that, when $x = -w$, we have

$$\phi_n(-w, w) = 0. \tag{35}$$

Thus we see that the eigenfunction $\phi_n(x, w)$ of $R(w)$ vanishes unless $|x| \leqslant w$, and then is a real exponential polynomial which vanishes at $x = -w$ and takes the value $(\lambda_n'(w)/\lambda_n(w))^{1/2}$ at $x = +w$.

It is not clear to us yet what role these eigenvalues and eigenfunctions may play in a further study of the Gel'fand–Levitan equation, nor what physical significance, if any, may be attached to them. We note here only that Eq. (4) admits an iterative solution

$$K(w) = -R(w) + R(w)^2 - R(w)^3 + \cdots \tag{36}$$

which converges in operator norm, according to the Fredholm theory, if and only if the eigenvalues $\lambda_n(w)$ of $R(w)$ all satisfy

$$|\lambda_n(w)| < 1. \tag{37}$$

This condition provides a natural obstacle to the convergence of any iterative procedure. In the physically interesting case $|r(k)| < 1$, and so (20) implies (37). We conclude that in this case the iterative solution (36) actually converges geometrically in operator norm to the operator $K(w)$.

It may also be possible to develop effective approximate solutions to the Gel'fand–Levitan equation by using a finite number of the eigenvalues and eigenfunctions as normal modes, to be computed numerically, e.g., by a suitable variational principle.

2. CALCULATIONS

Now we assume that $r(k)$ is rational, of the form (16), and rewrite it as

$$r(k) = \sum_{i=1}^{m} \frac{a_i}{k - b_i}. \tag{38}$$

Here a_i and b_i are complex constants, with $\operatorname{Im} b_i < 0$. We assume that the b_i are all distinct. It follows from (1) that

$$R(x+y) = \theta(x+y) \sum_{i=1}^{m} (-ia_i) e^{-ib_i(x+y)}. \tag{39}$$

We now insert the forms (27) and (39) into Eq. (21) and equate the coefficients of the exponential terms $e^{\pm ik_j x}$. In this way we find

$$\left(\sum_{i=1}^{m} \frac{a_i}{k_j - b_i}\right) A_j = r(k_j) A_j = -\lambda B_j, \tag{40}$$

$$\left(\sum_{i=1}^{m} \frac{a_i}{-k_j - b_i}\right) B_j = r(-k_j) B_j = -\lambda A_j, \tag{41}$$

$$\sum_{j=1}^{m} \left(\frac{A_j}{i(k_j - b_i)} e^{i(k_j - b_i)w} - \frac{B_j}{i(k_j + b_i)} e^{-i(k_j + b_i)w} \right) = 0. \tag{42}$$

To satisfy (40) and (41), we put

$$s(k_j) = (r(-k_j))^{1/2}, \tag{43}$$

$$t(k_j) = (r(+k_j))^{1/2}, \tag{44}$$

where the square roots are chosen so that $s(k_j)t(k_j) = -\lambda$. Then we put

$$A_j = s(k_j) C_j, \tag{45}$$

$$B_j = t(k_j) C_j, \tag{46}$$

with C_j to be determined, and note that (40) and (41) are satisfied for any choice of C_j.

Now (42) takes the form

$$\sum_{j=1}^{m} A_{ij} C_j = 0, \tag{47}$$

where the matrix A_{ij} is given by

$$A_{ij} = A_{ij}(w, \lambda) = \frac{s(k_j)}{i(k_j - b_i)} e^{i(k_j - b_i)w} - \frac{t(k_j)}{i(k_j + b_i)} e^{-i(k_j + b_i)w} \tag{48}$$

Note that the k_j, and hence the A_{ij}, depend on λ. Equation (47), and hence (42), admits a nontrivial solution if and only if

$$\det(A_{ij}(w, \lambda)) = 0. \tag{49}$$

This is the case only for certain values λ_n of λ; these values λ_n are then the eigenvalues, and the corresponding functions ϕ_n are the eigenfunctions of (24). In this way the eigenvalue problem for (24) reduces to the problem of solving (49).

It is instructive to apply this same procedure to obtain a solution $K(x, y, w)$ for the integral equation (4). An argument similar to that leading to (26) shows that if $y < x$, then $K(x, y, w)$ satisfies a differential equation in y of the form

$$q(-D) q(D) K(x, y, w) = p(D) p(-D) K(x, y, w). \tag{50}$$

Here $D = \partial/\partial y$. Hence $K(x, y, w)$ must have the form

$$K(x, y, w) = \sum_{j=1}^{m} A_j(x, w) e^{ik_j y} + B_j(x, w) e^{-ik_j y}, \tag{51}$$

where the k_j are now solutions of

$$r(-k) r(k) = 1. \tag{52}$$

This is just (28) with $\lambda^2 = 1$. If we insert (50) and (39) into (4) and equate coefficients of $e^{\pm ik_j y}$, we find

$$\left(\sum_{i=1}^{m} \frac{a_i}{k_j - b_i}\right) A_j = r(k_j) A_j = B_j, \tag{53}$$

$$\left(\sum_{i=1}^{m} \frac{a_i}{-k_j - b_i}\right) B_j = r(-k_j) B_j = A_j, \tag{54}$$

$$\sum_{j=1}^{m} \left(\frac{A_j}{i(k_j - b_i)} e^{i(k_j - b_i)w} - \frac{B_j}{i(k_j + b_i)} e^{-i(k_j + b_i)w} \right) = -e^{-ib_i x}. \tag{55}$$

Note that (53) and (54) are just (40) and (41) with $\lambda = -1$, and (42) is the homogeneous form of (55). Hence, with the choices (45) and (46) (with $\lambda = -1$) for A_j and B_j, we know that (53) and (54) are satisfied, and (55) becomes

$$\sum_{j=1}^{m} A_{ij} C_j = -e^{-ib_i x} \tag{56}$$

with the matrix $A_{ij} = A_{ij}(w, -1)$ given again by (48), with $\lambda = -1$. When $|r(k)| < 1$, we know [cf. (20)] that $\lambda = -1$ cannot be an eigenvalue of $R(w)$, and hence that $\det A_{ij}(w, -1)$ cannot vanish. Hence $A_{ij}(w, -1)$ must be invertible. Setting

$$B_{jk}(w) = (A^{-1}(w, 1))_{jk}, \tag{57}$$

we have

$$C_j(x, w) = \sum_{k=1}^{m} B_{jk}(w) e^{-ib_k x}, \tag{58}$$

and so

$$K(x, y, w) = -\sum_{j,k=1}^{m} B_{jk}(w) e^{-ib_k x} (s(k_j) e^{ik_j y} + t(k_j) e^{-ik_j y}). \tag{59}$$

Since

$$s(k_j) e^{ik_j y} + t(k_j) e^{-ik_j y} = A'_{kj}(y) e^{ib_k y}, \tag{60}$$

where $A'_{kj}(y) = dA_{kj}(y, -1)/dy$, we may rewrite (59) as

$$K(x, y, w) = -\sum_{j,k=1}^{m} B_{jk}(w) A'_{kj}(y) e^{-ib_k(x-y)}. \tag{61}$$

In particular, when $x = y = w$, (61) becomes

$$K(w,w,w) = -\frac{d}{dw}\mathrm{tr}\log A(w,-1). \tag{62}$$

Comparing (62) with (33), we see that

$$\Delta(w) = \mathrm{const}\times\det A(w,-1). \tag{63}$$

The constant in (63) need not be 1, as our example in the next section shows, but it plays no role in determining $K(w,w)$ or $V(w)$.

3. EXAMPLES

Here we work through the simplest class of examples. We assume that $m = 1$ in (38) and set $a_1 = i\alpha$, $b_1 = -i\beta$, so that

$$r(k) = \frac{i\alpha}{k+i\beta} = \frac{\alpha}{\beta - ik}, \tag{64}$$

with α,β real constants, $\alpha,\beta > 0$. (The potentials for these reflection coefficients have been obtained using Gel'fand–Levitan methods for $-\beta \leqslant \alpha < \beta$ in Refs. 8 and 9 and for $\alpha = \beta$ in Ref. 10. Note that the case $\alpha = \beta$ is a pathological case in which two distinct potentials can be found which have the same reflection coefficient.[10]) Then we have from (1)

$$R(x+y) = \theta(x+y)\alpha e^{-\beta(x+y)}, \tag{65}$$

and the eigenvalue equation (24) becomes

$$\alpha\int_{-x}^{w} e^{-\beta(x+y)}\phi(y,w)dy = \lambda\phi(x,w). \tag{66}$$

One may verify that (26) holds:

$$\begin{aligned}\lambda^2 q(-D)q(+D)\phi(x,w) &= \lambda^2(\beta^2 - D^2)\phi(x,w)\\ &= p(D)p(-D)\phi(x,w) = \alpha^2\phi(x,w),\end{aligned} \tag{67}$$

from which it follows that $\phi(x,w)$ must have the form (setting $k_1 = \kappa$)

$$\phi(x,w) = Ae^{i\kappa x} + Be^{-i\kappa x}, \tag{68}$$

with $\pm\kappa$ chosen so that

$$r(\kappa)r(-\kappa) = \alpha^2/\kappa^2 + \beta^2 = \lambda^2. \tag{69}$$

We assume first that $\lambda^2 < \alpha^2/\beta^2$, so that $\pm\kappa$ are real. Inserting (68) into (66), integrating, and equating the coefficients of $e^{\pm i\kappa x}$, we find

$$(i\alpha/(\kappa + i\beta))A = -\lambda B, \tag{70}$$

$$(i\alpha/(-\kappa + i\beta))B = -\lambda A. \tag{71}$$

Setting

$$s(\kappa) = (i\alpha/(-\kappa + i\beta))^{1/2}, \tag{72}$$

$$t(\kappa) = (i\alpha/(\kappa + i\beta))^{1/2}, \tag{73}$$

$$A = s(\kappa)C, \tag{74}$$

$$B = t(\kappa)C, \tag{75}$$

we get

$$\phi(x,w) = C(s(\kappa)e^{i\kappa x} + t(\kappa)e^{-i\kappa x}). \tag{76}$$

The matrix $A_{ij}(w,\lambda)$ in this case reduces to a single entry

$$A_{11}(w,\lambda) = \frac{s(\kappa)e^{(i\kappa-\beta)w}}{(i\kappa-\beta)} - \frac{t(\kappa)e^{(-i\kappa-\beta)w}}{(i\kappa+\beta)}. \tag{77}$$

It follows that

$$\begin{aligned}\alpha A_{11}(w,\lambda) &= -r(\kappa)s(\kappa)e^{(i\kappa-\beta)w} - r(-\kappa)t(\kappa)e^{(-i\kappa-\beta)w}\\ &= \lambda t(\kappa)e^{(i\kappa-\beta)w} + \lambda s(\kappa)e^{(-i\kappa-\beta)w}\\ &= \lambda e^{-\beta w}\phi(-w,w)/C,\end{aligned} \tag{78}$$

and

$$\alpha A'_{11}(x,\lambda) = \alpha e^{-\beta x}\phi(x,w)/C. \tag{79}$$

Thus the eigenvalue condition (49) in this case reduces to the condition

$$\phi(-w,w) = 0. \tag{80}$$

To satisfy (80), we set

$$s(\kappa) = \rho e^{-i\gamma}, \tag{81}$$

with $\rho = |r(\kappa)|^{1/2}$ and $\gamma = \frac{1}{2}\arg r(\kappa)$:

$$\rho = |\lambda|^{1/2}, \quad \gamma = \tfrac{1}{2}\arctan(\kappa/\beta). \tag{82}$$

Then we have

$$t(\kappa) = \begin{cases} \rho e^{i\gamma} & \text{if } \lambda < 0,\\ -\rho e^{i\gamma} & \text{if } \lambda > 0,\end{cases} \tag{83}$$

and (76) becomes

$$\phi(x,w) = \begin{cases} 2|\lambda|^{1/2}C\cos(\kappa x - \gamma) & \text{if } \lambda < 0,\\ 2i|\lambda|^{1/2}C\sin(\kappa x - \gamma) & \text{if } \lambda > 0.\end{cases} \tag{84}$$

Then (80) requires

$$\begin{aligned}&\cos(\kappa w + \gamma) = 0 \quad \text{if } \lambda < 0,\\ &\sin(\kappa w + \gamma) = 0 \quad \text{if } \lambda > 0,\end{aligned} \tag{85}$$

or

$$\kappa w + \gamma = \begin{cases} (n+\frac{1}{2})\pi & \text{if } \lambda < 0,\\ (n+1)\pi & \text{if } \lambda > 0,\end{cases} \tag{86}$$

where $n = 0, \pm 1, \pm 2, \ldots$, and in either case we are led to the transcendental equation

$$\kappa/\beta + \tan 2\kappa w = 0 \tag{87}$$

for the admissible solutions of κ, and hence of λ, in terms of w. The associated eigenvalues and eigenfunctions are then just the admissible values λ_n of λ, and

$$\phi_n(x,w) = \begin{cases} C_n\cos(\kappa_n x - \gamma_n) & \text{if } \lambda_n < 0,\\ C_n\sin(\kappa_n x - \gamma_n) & \text{if } \lambda_n > 0,\end{cases} \tag{88}$$

where C_n is merely a normalizing constant.

The reader can now verify that if $\lambda^2 > \alpha^2/\beta^2$, then $\pm\kappa$ are replaced throughout by $\pm i\mu$ with μ real, so that (85) is replaced by

$$\begin{cases}\cosh(\mu w + \gamma) = 0 & \text{if } \lambda < 0,\\ \sinh(\mu w + \gamma) = 0 & \text{if } \lambda > 0,\end{cases} \tag{89}$$

admitting no new admissible values for λ. This verifies what already seems reasonable, that

$$0 \leqslant \lambda_n^2 \leqslant \alpha^2/\beta^2, \tag{90}$$

i.e., that the λ_n^2 must lie in the range of $|r(k)|^2$.

The kernel $K(x,y,w)$ is given by (61) with $\lambda = -1$, which, in view of (78), reduces to

$$K(x,y,w) = -(\alpha\cos(\kappa y - \gamma)/\cos(\kappa w + \gamma))e^{\beta(w-x)}. \tag{91}$$

Here we suppose that $\alpha^2 > \beta^2$, in which case $\lambda^2 = 1 < \alpha^2/\beta^2$, $\kappa = +(\alpha^2 - \beta^2)^{1/2}$ is real, and $\gamma = \frac{1}{2}\arctan((\alpha^2/\beta^2) - 1)^{1/2}$. If $1 > \alpha^2/\beta^2$, then $\kappa = i\mu$ is imaginary, with $\mu = +(\beta^2 - \alpha^2)^{1/2}$. Then $s(\kappa) = (\alpha/(\beta - \mu))^{1/2} = e^{\delta}$, and $t(\kappa) = (\alpha/(\beta - \mu))^{-1/2} = e^{-\delta}$, where now $\delta = \frac{1}{2}\log(\alpha/(\beta - \mu)) = \operatorname{arctanh}(\mu/\beta)$. Then $\phi(x,w) = 2C\cosh(\mu x - \delta)$ and $\alpha A_{11}(\omega, -1) = -e^{-\beta w} 2\cosh(\mu w + \delta)/C$. Hence if $\alpha^2 < \beta^2$, then (91) is replaced by

$$K(x,y,w) = -(\alpha\cosh(\mu y - \delta)/\cosh(\mu w + \delta))e^{\beta(w-x)}. \tag{92}$$

The intractability of (87) prohibits an explicit determination of $\lambda_n(w)$, or of $\Delta(w)$, in general. In the limiting case $\alpha = 1$, $\beta = 0$, however, we have $\gamma = \pi/4$, and (87) becomes

$$\kappa w + \frac{\pi}{4} = \begin{cases} (n + \frac{1}{2})\pi & \text{if } \lambda < 0, \\ (n+1)\pi & \text{if } \lambda > 0. \end{cases} \tag{93}$$

The positive admissible values of κ are

$$\kappa_n = \begin{cases} (4n+1)\pi/4w & \text{if } \lambda < 0, \\ (4n+3)\pi/4w & \text{if } \lambda > 0, \end{cases} \tag{94}$$

for $n = 0,1,2,\ldots$, and the admissible values of λ are

$$\lambda_n = \begin{cases} -1/\kappa_n = -4w/(4n+1)\pi & \text{if } \lambda_n < 0, \\ +1/\kappa_n = +4w/(4n+3)\pi & \text{if } \lambda_n > 0, \end{cases} \tag{95}$$

or

$$\lambda_n = (-1)^{n+1} 4w/(2n+1)\pi, \quad n = 0,1,2,\ldots. \tag{96}$$

Hence in this case

$$\begin{aligned} \Delta(w) &= \prod_{n=0}^{\infty}(1 + \lambda_n) \\ &= \prod_{n=0}^{\infty}\left(1 + \frac{(-1)^{n+1}4w}{(2n+1)\pi}\right) \\ &= 2^{1/2}\cos(w + \pi/4). \end{aligned} \tag{97}$$

On the other hand, from (78) we have in this case

$$A_{11}(w,\lambda) = (\lambda/C)\phi(-w,w). \tag{98}$$

In particular, for $\lambda = -1$, $\kappa = +1$,

$$A_{11}(w,-1) = -2(\cos\kappa w + \pi/4) \tag{99}$$

so that $\Delta(w)$ and $\det(A_{11}(w,-1))$ differ by the constant factor $-2^{1/2}$. The eigenfunctions in this case are given by

$$\phi_n(x,w) = \begin{cases} C_n\cos((4n+1)\pi x/4w - \pi/4) & \text{if } \lambda_n < 0, \\ C_n\sin((4n+3)\pi x/4w - \pi/4) & \text{if } \lambda_n < 0, \end{cases}$$

and the kernel $K(x,y,w)$ is given by [cf. (91)]

$$K(x,y,w) = \frac{-\cos(y - \pi/4)}{\cos(w + \pi/4)}. \tag{100}$$

We have assumed throughout this section that $\alpha > 0$ in (64). The reader may verify that if $\alpha < 0$, then everything is exactly the same except that the phase $\gamma = \frac{1}{2}\arg r(\kappa)$ is then augmented by π. We have avoided the case $\alpha^2/\beta^2 = 1$, since then, when $\lambda = -1$, $\kappa = 0$, and so $\pm\kappa$ are not distinct (cf. Ref. 10).

[1] I. M. Gel'fand and B. M. Levitan, Izv. Akad. Nauk SSSR, Math Series **15**, 309 (1951).

[2] I. Kay and H. E. Moses, *Inverse Scattering Papers: 1955–1963* (Math. Sci. Press, Brookline, MA, 1982).

[3] B. Simon, "Notes on Infinite Determinants of Hilbert Space Operators," Advances in Math. **24**, 244 (1977).

[4] H. Cornille, "Connection between the Marchenko formalism and N/D equations I.," J. Math. Phys. **8**, 2268 (1967).

[5] F. J. Dyson, "Fredholm Determinants and Inverse Scattering Problems," Comm. Math. Phys. **47**, 171 (1976); R. G. Newton, *Scattering Theory of Waves and Particles*, 2nd. ed. (McGraw-Hill, New York) (to appear); R. T. Prosser, "A General Solution for the One-Dimensional Inverse Scattering Problem," Dartmouth College, 1981 (unpublished).

[6] I. Kay, "The Inverse Scattering Problem When the Reflection Coefficient is a Rational Function," Comm. Pure Appl. Math. **13**, 371 (1960).

[7] K. R. Pechenick and J. Cohen "Inverse scattering—exact solution of the Gel'fand–Levitan equation," J. Math. Phys. **22**, 1513 (1981).

[8] H. E. Moses, "An Example of the Effect of the Rescaling of the Reflection Coefficient on the Scattering Potential for the One-Dimensional Schrödinger Equation," Stud. Appl. Math. **60**, 177 (1979).

[9] P. B. Abraham and H. E. Moses, "Exact Solutions of the One-Dimensional Acoustic Wave Equation for Several New Velocity Profiles. Transmission and Reflection Coefficients," J. Acoust. Soc. Am. **71**, 1391 (1982).

[10] P. B. Abraham, B. De Facio, and H. E. Moses, "Two Distinct Local Potentials with No Bound States Can Have the Same Scattering Operator: A Non-Uniqueness in Inverse Spectral Transformations," Phys. Rev. Lett. **46**, 1657 (1981).

The causal automorphism of de Sitter and Einstein cylinder spacetimes

J. A. Lester
Department of Pure Mathematics, University of Waterloo, Waterloo, Ontario, Canada

(Received 20 July 1982; accepted for publication 10 December 1982)

A well-known result, due originally to Alexandrov in 1953 and subsequently rediscovered by Zeeman in 1964, states that transformations of Minkowski spacetime which preserve causality are essentially orthochronous Lorentz transformations. In this article, we first exhibit a proof of this result by using a lemma of Zeeman to reduce the proof to another well-known theorem of Alexandrov involving transformations preserving light speed. Then, by generalizing Zeeman's lemma and using recent extensions of Alexandrov's light-speed theorem, we determine the causal automorphisms of de Sitter and Einstein cylinder spacetimes.

PACS numbers: 04.20. − q

1. INTRODUCTION: THE CAUSAL AUTOMORPHISM OF MINKOWSKI SPACETIME

Minkowski spacetime may be thought of as $\mathbb{R}^4$ equipped with the metric $(\ ,\ )$ given by

$$(x, y) := -x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4$$

for all $x := (x_1,x_2,x_3,x_4)$, $y := (y_1, y_2, y_3, y_4) \in \mathbb{R}^4$. The separation between events $x, y \in M_4$ is the quantity $(x-y, x-y)$, and is preserved by all translations and Lorentz transformations (linear, metric-preserving bijections) of M_4.

The separation between events in M_4 is zero iff they are joined by an unreflected light signal. Alexandrov's "light-speed" theorem[1,2] states that bijections of M_4 preserving separation zero in both directions must be Lorentz transformations, up to translations and dilatations (scale changes). The significance of this result is that, unlike Einstein's original derivation of Lorentz transformations,[3] it assumes no regularity conditions (e.g., linearity, or even continuity) for the transformations.

A vector $x \in M_4$ is said to be timelike, null, or spacelike whenever (x,x) is negative, zero, or positive, respectively. The nonzero null and timelike vectors lie, respectively, on and inside one of the two halves of a circular cone in M_4. They are thus segregated into two disconnected components, which we may (arbitrarily) label future-pointing vectors and past-pointing vectors. It is easily checked that, unless they are parallel null vectors, two nonspace-like vectors $x \neq 0$, $y \neq 0$ lie in the same component iff $(x, y) < 0$. Lorentz transformations which preserve future-pointing vectors are said to be orthochronous, and form a subgroup of the full Lorentz group.

Causality on M_4 may be defined in terms of future-pointing vectors as follows. A line in M_4 with timelike direction represents the spacetime history of a material particle experiencing no external force, while a line with null direction describes the history of an unreflected photon. Since an event $x \in M_4$ can cause an event $y \in M_4$ iff a material particle or photon can experience both events in that order, two corresponding causal relations, symbolized by $\prec$ and $\prec\cdot$, may be formulated.

Definition 1.1: For $x,y \in M_4$,

(i) $x \prec y$ iff $y - x$ is timelike and future pointing,

(ii) $x \prec\cdot y$ iff $y - x$ is null and future pointing.

The result which interests us here appeared first as one of several related results in Ref. 4; its rediscovery by Zeeman[5] appears to be better known (at least among physicists), possibly because the former article is in Russian (see Ref. 6 for historical background). The theorem states that bijections of M_4, which preserve the relation $\prec$ in both directions (Zeeman's "causal automorphisms"), must be orthochronous Lorentz transformations, up to translations and dilatations.

The significance of this result is again, as with Alexandrov's light-speed theorem, the absence of regularity assumptions on the transformations involved: preservation of a simple, physical condition is sufficient. For this reason, interest in these and similar characterizations has been growing steadily in recent years, particularly among geometers. Many generalizations now exist; these involve other spacetimes, other separations, more abstract light-cone structures, spaces over more general fields, etc. The bibliographies of Refs. 6 and 7, for example, provide a cross section of such works.

Zeeman's proof of the causality-preservation theorem on M_4 begins by showing that causal automorphisms must also preserve the relation $\prec\cdot$ in both directions. The crux of the matter is the following condition, for which we supply the proof omitted in Ref. 5.

Lemma 1.1: For distinct $x, y \in M_4$,

$$x \prec\cdot y \text{ iff } \begin{cases} x \not\prec y, \\ \text{for all } z \in M_4,\ z \prec x \text{ implies } z \prec y. \end{cases}$$

Proof: (a) Assume that $x \prec\cdot y$; then clearly $x \not\prec y$. For any $z \in M_4$ with $z \prec x$, write $y - z = (y - x) + (x - z)$; then

$$(y - z, y - z) = (y - x, y - x) + 2(y - x, x - z) + (x - z, x - z) < 0$$

since $y - x$ is null, $x - z$ is timelike, and both are future pointing. Thus $y - z$ is timelike, and, from $(y - z, y - x) = (x - z, y - x) < 0$, $y - z$ is future pointing (since $y - x$ is). Hence $z \prec y$.

(b) Assume that $x \not\prec\cdot y$ and $x \not\prec y$. For any timelike future-pointing vector t not parallel to $y - x$, the two-space spanned by t and $y - x$ contains a spacelike vector of the form $(y - x) + \alpha t$. If $y - x$ is spacelike, we may choose $\alpha > 0$ for small enough α; otherwise $y - x$ must be past pointing (else

$x \lessdot y$ or $x \lessdot \cdot y$), which implies $\alpha > 0$. In either case, the vector $z := x - \alpha t$ satisfies $z \lessdot x$, but $z \not\lessdot y$. ■

After restricting attention to the relation $\lessdot \cdot$, Zeeman's proof proceeds through properties of quadric surfaces, compositions of parallel displacements, Cauchy's functional equation, etc., eventually reaching the required result. However, since two events $x, y \in M_4$ have zero separation iff $x \lessdot \cdot y$ or $y \lessdot \cdot x$, the theorem follows immediately via Alexandrov's light-speed theorem. (The above lemma and the consequent shortcut actually work for Minkowski space M_n of any dimension $n \geqslant 3$, where both theorems are valid. For $n = 2$, both theorems fail.) In the next sections, generalizations of Lemma 1.1 and Alexandrov's result will yield the causal automorphisms of de Sitter and Einstein cylinder spacetimes.

2. CAUSAL AUTOMORPHISMS OF DE SITTER SPACETIME

de Sitter spacetime $\mathscr{S}_4$ can be embedded as a hyperboloid in five-dimensional Minkowski space M_5 (see Ref. 8, Sec. 5.2), i.e., if $(\ ,\ )$ denotes the metric of M_5, then $\mathscr{S}_4 := \{x | x \in M_5, (x,x) = 1\}$, and the (differential) metric of $\mathscr{S}_4$ is given by $ds^2 := (dx,dx)$. Events $x, y \in \mathscr{S}_4$ with $(x, y) > -1$ are joined by a geodesic (given by a section of $\mathscr{S}_4$ with a two-space in M_5; see Ref. 7) and their separation is s^2, where s (found by integrating ds along this geodesic) is the real or pure imaginary number given by $4 \sin^2(s/2) = (x - y, x - y)$. A direct generalization of Alexandrov's light-speed theorem[7] states that bijections of $\mathscr{S}_4$ preserving separation $s = 0$ [i.e., preserving the relation $(x, y) = 1$] in both directions must be induced on $\mathscr{S}_4$ by the Lorentz transformations of M_5.

The causal structure of $\mathscr{S}_4$ is induced by that of M_5; upon distinguishing the past-pointing and future-pointing vectors of M_5 as in Sec. 1, we define the causal relations $\lessdot$ and $\lessdot \cdot$ on $\mathscr{S}_4$ essentially as before.

Definition 2.1: For $x, y \in \mathscr{S}_4$,

(i) $x \lessdot y$ iff $y - x$ is timelike in M_5 and future pointing,
(ii) $x \lessdot \cdot y$ iff $y - x$ is null in M_5 and future pointing.

Clearly, the orthochronous Lorentz transformations of M_5 induce causality-preserving transformations of $\mathscr{S}_4$. Zeeman's condition, which generalizes exactly to $\mathscr{S}_4$ (see below), will enable us to establish these induced transformations as the only causal automorphisms of $\mathscr{S}_4$.

Lemma 2.1: For distinct $x, y \in \mathscr{S}_4$,

$$x \lessdot \cdot y \text{ iff } \begin{cases} x \not\lessdot y, \\ \text{for all } z \in \mathscr{S}_4,\ z \lessdot x \text{ implies } z \lessdot y. \end{cases}$$

Proof: If $x \lessdot \cdot y$, repeat part (a) of the proof of lemma 1.1 with $z \in \mathscr{S}_4$ to get the required results.

Assume that $x \not\lessdot \cdot y$ and $x \not\lessdot y$. Choose a future-pointing $t \in M_5$ with $(t,t) = -1$, $(t,x) = 0$, and for $\epsilon > 0$, define $z := (1 + \epsilon^2)^{1/2} x - \epsilon t$. Then $z \in \mathscr{S}_4$, $z \lessdot x$, and, for $\lambda := (x, y)$ and $\mu := (t, y)$, $(y - z, t) = \mu - \epsilon$ and $(y - z, y - z) = 2\{1 + \epsilon\mu - (1 + \epsilon^2)^{1/2}\lambda\}$. We find choices of ϵ for which $z \not\lessdot y$.

If $\mu > 0$, choose $\epsilon < \mu$; then $(y - z, t) > 0$, so $z \not\lessdot y$.

If $\mu = 0$, the space spanned by x and y is orthogonal to t, and is thus positive definite. The Cauchy–Schwarz inequality gives $\lambda^2 \leqslant 1$, so since $\lambda \neq 1$ $(x \neq y)$ we have $\lambda < 1$. Then for some $\epsilon > 0$, $(1 + \epsilon^2)^{1/2}\lambda < 1$, so for this ϵ, $(y - z, y - z) > 0$ and hence $z \not\lessdot y$.

If $\mu < 0$, then $\lambda < 1$ (else $x \lessdot \cdot y$ or $x \lessdot y$). If $\lambda > 0$, choose ϵ with $\epsilon(\lambda - \mu) < 1$; then $(1 + \epsilon^2)^{1/2}\lambda < (1 + \epsilon)\lambda < 1 + \mu\epsilon$, so $(y - z, y - z) > 0$. If $\lambda \leqslant 0$, choose $\epsilon < -\mu^{-1}$; then $1 + \mu\epsilon > 0$, so again $(y - z, y - z) > 0$. In either case, then, $z \not\lessdot y$. ■

Using the generalized light-speed theorem exactly as in Sec. 1, we have that bijections of $\mathscr{S}_4$ which preserve the relation $\lessdot$ in both directions must be induced by the orthochronous Lorentz transformations of M_5. We note that since the generalized Alexandrov result is in fact true for de Sitter spaces $\mathscr{S}_n$ of any dimension $n \geqslant 3$, so is our present result.

3. CAUSAL AUTOMORPHISMS OF EINSTEIN'S CYLINDER UNIVERSE

Einstein's cylinder universe $\mathscr{C}_4$ can be visualized as a circular cylinder in $\mathbb{R}^5$ (see Ref. 8, p. 121), i.e., if "$\cdot$" denotes the usual dot product of $\mathbb{R}^4$, then

$$\mathscr{C}_4 := \{(\rho, r) | \rho \in \mathbb{R}, r \in \mathbb{R}^4, r \cdot r = 1\} \text{ and}$$
$$ds^2 := -d\rho^2 + dr \cdot dr.$$

Its geodesics are either circular sections of $\mathscr{C}_4$ (which are spacelike) or of the form $r = \cos(\alpha\rho)a + \sin(\alpha\rho)b$ for some constant $\alpha \geqslant 0$ and orthonormal $a, b \in \mathbb{R}^4$ (and are timelike, null, or spacelike whenever $\alpha < 1$, $\alpha = 1$, or $\alpha > 1$, respectively). In general, two points of $\mathscr{C}_4$ are joined by many geodesics (e.g., for orthonormal $a, b \in \mathbb{R}^4$, the points (0,a) and $(\pi/2, b)$ are joined by all geodesics of the form $r = \cos[(1 + 4k)\rho]a + \sin[(1 + 4k)\rho]b$ for integral k, thus the separation s^2 between them (obtained by integrating ds along a joining geodesic) will be multivalued. For events (ρ_1, r_1), $(\rho_2, r_2) \in \mathscr{C}_4$ we obtain $s^2 = -(\rho_1 - \rho_2)^2 + \{\cos^{-1}(r_1 \cdot r_2)\}^2$, which is negative, zero, or positive whenever the geodesic is timelike, null, or spacelike, respectively.

The transformation group of $\mathscr{C}_4$ (i.e., the group of transformations of $\mathscr{C}_4$ which preserve ds^2 at each point) consists of mappings of the form $(\rho, r) \to (\pm\rho + \text{const.}, Ar)$, where A is a 4×4 orthogonal matrix. The light-speed theorem does *not* generalize to $\mathscr{C}_4$, since there exist rather pathological transformations of $\mathscr{C}_4$ which preserve separation zero (see Ref. 9 for details; the relevant points follow). For example, for fixed $(\rho, r) \in \mathscr{C}_4$, arbitrary permutations within the subset $\{(\rho + k\pi, (-1)^k r) | k \text{ an integer}\} \subset \mathscr{C}_4$ preserve separation zero. Up to such permutations, bijections of $\mathscr{C}_4$ which preserve separation zero [or equivalently, the relation $\cos(\rho_1 - \rho_2) = r_1 \cdot r_2$] in both directions have the form $(r, \cos\rho, \sin\rho)^t \to \lambda T(r, \cos\rho, \sin\rho)^t$ for a scalar function $\lambda = \lambda(\rho, r)$ (determined up to sign by the requirement that the condition $r \cdot r = 1$ be preserved) and a (constant) 6×6 matrix T satisfying $T^t G T = G$, where $G := \text{diag}\{1,1,1,1,-1,-1\}$ (superscript t denotes transpose).

Causality is defined on $\mathscr{C}_4$ by using the ρ-coordinate as a criterion of temporal order, i.e., an event $(\rho_1, r_1) \in \mathscr{C}_4$ can cause an event $(\rho_2, r_2) \in \mathscr{C}_4$ iff they can be joined by a timelike

or null geodesic and $\rho_1 < \rho_2$. Specifically, we define the causal relations $\prec$ and $\prec\cdot$ as follows.

Definition 3.1: For $(\rho_1,r_1), (\rho_2,r_2) \in \mathscr{C}_4$,

(i) $(\rho_1,r_1) \prec (\rho_2,r_2)$ iff $s^2 < 0$ for some geodesic joining them and $\rho_1 < \rho_2$,

(ii) $(\rho_1,r_1) \prec\cdot (\rho_2,r_2)$ iff $s^2 = 0$ for some geodesic joining them and $\rho_1 < \rho_2$.

A more useful characterization of these relations follows:

Lemma 3.1: For $(\rho_1,r_1), (\rho_2,r_2) \in \mathscr{C}_4$,

(i) $(\rho_1,r_1) \prec (\rho_2,r_2)$ iff $\rho_1 < \rho_2$ and either $\rho_2 - \rho_1 > \pi$ or $\cos(\rho_2 - \rho_1) < r_1 \cdot r_2$,

(ii) $(\rho_1,r_1) \prec\cdot (\rho_2,r_2)$ iff $\rho_1 < \rho_2$ and $\cos(\rho_1 - \rho_2) = r_1 \cdot r_2$.

Proof: for some $0 \leqslant \theta \leqslant \pi$, $r_1 \cdot r_2 = \cos\theta$, thus

$$\begin{aligned} s^2 &= -(\rho_1 - \rho_2)^2 + \{\cos^{-1}(\cos\theta)\}^2 \\ &= -(\rho_1 - \rho_2)^2 + (\pm\theta + 2k\pi)^2 \end{aligned}$$

for integral k. A tedious but elementary analysis of which k's are possible for $s^2 < 0$ and $s^2 = 0$ yields the required results. ■

Zeeman's condition must be slightly modified to hold on $\mathscr{C}_4$.

Lemma 3.2: For distinct $(\alpha,a), (\beta,b) \in \mathscr{C}_4$,

$$\left.\begin{array}{l}(\alpha,a) \prec\cdot (\beta,b) \\ \text{and } \beta - \alpha \leqslant \pi\end{array}\right\}$$

$$\text{iff} \begin{cases} (\alpha,a) \not\prec (\beta,b), \text{ and for all } (\gamma,c) \in \mathscr{C}_4, \\ (\gamma,c) \prec (\alpha,a) \text{ implies } (\gamma,c) \prec (\beta,b). \end{cases}$$

Proof: Without loss of generality we may assume that $\alpha = 0$. For some $0 \leqslant \omega \leqslant \pi$, $a \cdot b = \cos\omega$. Recall that the cosine function is decreasing on $[0,\pi]$.

(a) Assume that $\beta \leqslant \pi$ and that $(0,a) \prec\cdot (\beta,b)$. Then $0 < \beta \leqslant \pi$ and $\cos\beta = b \cdot a = \cos\omega$, so $\omega = \beta > 0$. Suppose there exists a $(\gamma,c) \in \mathscr{C}_4$ with $(\gamma,c) \prec (0,a)$ but $(\gamma,c) \not\prec (\beta,b)$. We have $\gamma < 0 < \beta$ and $0 - \gamma \leqslant \pi$ [else $\beta - \gamma > \pi$, which implies $(\gamma,c) \prec (\beta,b)$]; thus $\cos(-\gamma) < a \cdot c$. For $0 \leqslant \theta, \phi \leqslant \pi$ defined by $\cos\theta := a \cdot c$, $\cos\phi = b \cdot c$, we have $\cos(-\gamma) < \cos\theta$, $\cos(\beta - \gamma) \geqslant \cos\phi$; thus $-\gamma > \theta$ and $\beta - \gamma \leqslant \phi$, from which $\phi > \theta + \omega$ and $\cos\phi < \cos(\theta + \omega)$.

Since the subspace of $\mathbb{R}^4$ spanned by a, b, and c is positive definite,

$$0 \leqslant \begin{vmatrix} a\cdot a & b\cdot a & c\cdot a \\ a\cdot b & b\cdot b & c\cdot b \\ a\cdot c & b\cdot c & c\cdot c \end{vmatrix} = \begin{vmatrix} 1 & \cos\omega & \cos\theta \\ \cos\omega & 1 & \cos\phi \\ \cos\theta & \cos\phi & 1 \end{vmatrix},$$

which may be written

$$\{\cos(\theta + \omega) - \cos\phi\}\{\cos(\theta - \omega) - \cos\phi\} \leqslant 0.$$

The first factor has been proven positive, so $\cos\phi \geqslant \cos(\theta - \omega)$, whence $\theta + \omega \leqslant |\theta - \omega|$. This last implies the contradiction that either θ or ω is negative, thus no $(\gamma,c) \in \mathscr{C}_4$ with $(\gamma,c) \prec (0,a)$ and $(\gamma,c) \not\prec (\beta,b)$ exists.

(b) Assume that $(0,a) \not\prec (\beta,b)$ and that for all $(\gamma,c) \in \mathscr{C}_4$, $(\gamma,c) \prec (0,a)$ implies that $(\gamma,c) \prec (\beta,b)$. If $\beta > \omega$, then $\pi \geqslant \beta > \omega \geqslant 0$, and consequently $\cos\beta < \cos\omega = a \cdot b$. Hence $(0,a) \prec (\beta,b)$, a contradiction. If $\beta < \omega$, define $(\gamma,c) := (-\epsilon,a)$ for $0 < \epsilon < \omega - \beta$; then $\gamma < 0$ and $\cos(-\gamma) = \cos\epsilon < 1 = a \cdot c$, so $(\gamma,c) \prec (0,a)$. But $\cos(\beta - \gamma) = \cos(\beta + \epsilon) > \cos\omega = b \cdot c$ and $\beta - \gamma = \beta + \epsilon < \omega \leqslant \pi$, so $(\gamma,c) \not\prec (\beta,b)$, a contradiction.

If $\beta = \omega = 0$, then $a \cdot b = 1$, so $a = b$ and $(0,a) = (\beta,b)$, a contradiction. There remains the case $\beta = \omega > 0$, which yields $0 < \beta \leqslant \pi$ and $\cos\beta = a \cdot b$, from which $(0,a) \prec\cdot (\beta,b)$ as required. ■

We see that bijections of $\mathscr{C}_4$ which preserve the relation $\prec$ in both directions preserve zero separation for "close enough" points. The following lemma, which rules out the existence of "null triangles" in $\mathscr{C}_4$, will enable us to extend this result to more distant points.

Lemma 3.3: Three distinct points $(\alpha,a), (\beta,b), (\gamma,c) \in \mathscr{C}_4$ with pairwise zero separation lie on a common null geodesic.

Proof: Without loss of generality $\alpha = 0$, so $\cos\beta = a \cdot b$, $\cos\gamma = a \cdot c$, and $\cos(\beta - \gamma) = b \cdot c$. If b and c are parallel to a, then for some integers k,n, $(\beta,b) = (k\pi,(-1)^k a)$ and $(\gamma,c) = (n\pi,(-1)^n a)$. Then for any unit $d \in \mathbb{R}^4$ orthogonal to a, all three points lie on the null geodesic with equation $r = (\cos\rho)a + (\sin\rho)d$.

We may now assume that b is not parallel to a; thus $\sin\beta \neq 0$. Define $d := -\cot\beta a + \csc\beta b$; then d is unit and orthogonal to a, and $(0,a)$ and (β,b) lie on the null geodesic with equation $r = (\cos\rho)a + (\sin\rho)d$. For some scalars Ψ,ϕ and some $e \in \mathbb{R}^4$ orthogonal to a and d, $c = \Psi a + \phi d + e$, so $\Psi = a \cdot c = \cos\gamma$ and $\phi = c \cdot d = \sin\gamma$, from which $1 = c \cdot c = \cos^2\gamma + \sin^2\gamma + e \cdot e$. It follows that $e = 0$, so (γ,c) is also on the null geodesic. ■

Now consider two "distant" points $(\rho_1,r_1), (\rho_2,r_2) \in \mathscr{C}_4$ with separation zero. Cover the null geodesic segment joining them by a collection of open, overlapping subsegments whose points are "close enough," i.e., whose ρ-coordinates differ by at most π. By Lemma 3.3, the images of these subsegments under a causal automorphism are also overlapping segments of null geodesics. But the points of each image overlap lie on at most a single null geodesic, so in fact, all image segments lie on the *same* null geodesic. This geodesic joins the images of (ρ_1,r_1), (ρ_2,r_2), so these image points also have separation zero.

Since they preserve separation zero in both directions, our causal automorphisms have the form

$$(r,\cos\rho,\sin\rho)^t \to \lambda T (r,\cos\rho,\sin\rho)^t$$

for scalar λ and matrix T as described earlier, up to permutations within subsets of the form

$$\{(\rho + k\pi,(-1)^k r) | k \text{ an integer}\} \text{ for fixed } (\rho,r) \in \mathscr{C}_4.$$

But such permutations must now preserve causality, so since all points of the subset lie on a common null geodesic, their order must be preserved. It follows that if $(\bar\rho,\bar r)$ denotes the image of (ρ,r) under a causal automorphism, then the image of $(\rho + k\pi,(-1)^k r)$ is $(\bar\rho + k\pi,(-1)^k \bar r)$ for all integers k.

The scalar $\lambda = \lambda(\rho,r)$ is fixed up to sign by the requirement that $\bar r \cdot \bar r = 1 = \cos^2\bar\rho + \sin^2\bar\rho$. For (β,b), $(\gamma,c) \in \mathscr{C}_4$ we have

$$\begin{aligned} b \cdot c - \cos(\bar\beta - \bar\gamma) &= (\bar b,\cos\bar\beta,\sin\bar\beta) G (\bar c,\cos\bar\gamma,\sin\bar\gamma)^t \\ &= \lambda(\beta,b)\lambda(\gamma,c)(b,\cos\beta,\sin\beta)(T^t G T)(c,\cos\gamma,\sin\gamma)^t \\ &= \lambda(\beta,b)\lambda(\gamma,c)\{b \cdot c - \cos(\beta - \gamma)\} \end{aligned}$$

since $T^t G T = G := \operatorname{diag}\{1,1,1,1,-1,-1\}$. Since causality is preserved, all λ's have the same sign. We may in fact take

$\lambda > 0$: since $T^t GT = G$ iff $(-T)^t G(-T) = G$, minus signs may be absorbed into T.

For $(\alpha,a) \in \mathscr{C}_4$, consider the subset

$$\mathscr{M}(\alpha,a) := \{(\rho,r) \in \mathscr{C}_4 \mid (\alpha,a) \lessdot (\rho,r) \text{ and either } (\rho,r) \lessdot (\alpha + 2\pi,a) \text{ or } (\rho,r) \lessdot \cdot (\alpha + 2\pi,a) \text{ or } (\rho,r) = (\alpha + 2\pi,a)\}.$$

Clearly, if $(\alpha,a) \to (\overline{\alpha},\overline{a})$, then $\mathscr{M}(\alpha,a)$ maps into $\mathscr{M}(\overline{\alpha},\overline{a})$. Furthermore, careful examination of the definitions of $\lessdot$ and $\lessdot\cdot$ shows that

$$\mathscr{M}(\alpha,a) = \{(\rho,r) \in \mathscr{C}_4 \mid 0 < \rho - \alpha \leqslant 2\pi,\ \cos(\rho - \alpha) \leqslant r \cdot a, \text{ and if } \cos(\rho - a) = r \cdot a, \text{ then } \rho - \alpha > \pi\},$$

from which it can be checked that any point $(\rho,r) \in \mathscr{C}_4$ can be uniquely expressed as $(\sigma + k\pi, (-1)^k s)$ for some $(\sigma,s) \in \mathscr{M}(\alpha,a)$ and integer k. It follows that the image of any point of $\mathscr{C}_4$ is determined by the image of $\mathscr{M}(\alpha,a)$ for any given $(\alpha,a) \in \mathscr{C}_4$.

In summary, we may describe any bijection of $\mathscr{C}_4$ which preserves the relation $\lessdot$ in both directions as follows: choose $(\alpha,a) \in \mathscr{C}_4$ with image $(\overline{\alpha},\overline{a})$. Then for some 6×6 matrix T with $T^t GT = G := \operatorname{diag}\{1,1,1,1,-1,-1\}$ and for a uniquely determined scalar function $\lambda = \lambda(\sigma,s) > 0$, the causal automorphism maps $\mathscr{M}(\alpha,a)$ onto $\mathscr{M}(\overline{\alpha},\overline{a})$, and has the form $(\sigma,s) \to (\overline{\sigma},\overline{s})$, where

$$(\overline{s},\cos\overline{\sigma},\sin\overline{\sigma})^t = \lambda T(s,\cos\sigma,\sin\sigma)^t$$

on $\mathscr{M}(\alpha,a)$. Any point $(\rho,r) \in \mathscr{C}_4$ has the form $(\rho,r) = (\sigma + k\pi, (-1)^k s)$ for some unique $(\sigma,s) \in \mathscr{M}(\alpha,a)$ and integer k: its image is then $(\overline{\sigma} + k\pi, (-1)^k \overline{s})$.

It is easily checked that, given any point $(\alpha,a) \in \mathscr{C}_4$, any image point $(\overline{\alpha},\overline{a})$, and a 6×6 matrix T satisfying $T^t GT = G$, then the bijection of $\mathscr{C}_4$ defined as above by $(\alpha,a),(\overline{\alpha},\overline{a})$, and T is a causal automorphism of $\mathscr{C}_4$; we have thus characterized all causal automorphisms. As for Minkowski and de Sitter spacetimes, the characterization is in fact valid for n-dimensional Einstein cylinder spaces $\mathscr{C}_n$ for $n \geqslant 3$.

We note finally that the occurrence of the subsets $\mathscr{M}(\alpha,a)$ above is no accident: the interior of each is conformal to Minkowski spacetime M_4 (see Ref. 8, p. 122). The translations, dilatations, and orthochronous Lorentz transformations of M_4 induced transformations on $\mathscr{M}(\alpha,a)$ which, since they preserve the signs of separations between points, preserve causality on $\mathscr{M}(\alpha,a)$. The causal automorphisms obtained above are in fact compositions of these transformations with those of the transformation group of $\mathscr{C}_4$ described earlier.

[1] A. D. Alexandrov, "On Lorentz transformations," Usp. Mat. Nauk **5**, 187 (1950).

[2] A. D. Alexandrov, "A contribution to chronogeometry," Can. J. Math. **19**, 1110–1128 (1967).

[3] A. Einstein, "Zur Elektrodynamik bewegter Körper," Ann. Phys. **17**, 891–921 (1905).

[4] A. D. Alexandrov and V. V. Ovchinnikova, "Notes on the foundations of relativity theory," Vestn. Leningr. Univ. **11**, 95–100 (1953).

[5] E. C. Zeeman, "Causality implies the Lorentz group," J. Math. Phys. **5**, 490–493 (1964).

[6] A. D. Alexandrov, "Mappings of space with families of cones and space-time transformations," Ann. Mat. Pura Appl. (4) **103**, 229–257 (1975).

[7] J. A. Lester, "Separation-preserving transformations of de Sitter space-time," Abh. Math. Sem. Univ. Hamburg (to appear).

[8] S. W. Hawking and G. F. R. Ellis, in *The Large-Scale Structure of Space-time* (Cambridge University Press, Cambridge, 1973).

[9] J. A. Lester, "Alexandrov-type transformations of Einstein's cylinder universe," C. R. Math. Rep. Acad. Sci. Can. **4**, 175–178 (1982).

Spherically symmetric solution in the nonsymmetric Kaluza–Klein theory

M. W. Kalinowski[a] and G. Kunstatter
Physics Department, University of Toronto, Toronto, Ontario, M5S 1A7, Canada

(Received 6 May 1983; accepted for publication 26 August 1983)

In this paper we find an exact, static, spherically symmetric solution for the nonsymmetric Kaluza–Klein theory. This solution has the remarkable property of describing "mass without mass" and "charge without charge." We examine its properties and a physical interpretation.

PACS numbers: 04.50. + h, 11.10.Ef

INTRODUCTION

The aim of this paper is to find an exact spherically symmetric solution to the nonsymmetric Kaluza–Klein equations (see Refs. 1–7) in the electromagnetic case.[1,3]

The nonsymmetric Kaluza–Klein theory provides a true unification of the electromagnetic and gravitational fields in the following sense. It not only reduces two major principles of invariance (i.e., the local coordinate invariance principle and the local gauge invariance principle) to the local coordinate invariance principle, but it also gives rise to new effects, which are absent in the classical Kaluza–Klein theory. These effects do not appear in either Moffat's theory of gravitation (see Refs. 8–10) or in Maxwell's electromagnetism. They are therefore interference effects between the gravitational and electromagnetic fields. We outline these new features of the nonsymmetric Kaluza–Klein theory below (see Ref. 1):

1. A new term appears in the electromagnetic Lagrangian of the form

$$(1/4\pi)(g^{[\mu\nu]}F_{\mu\nu})^2.$$

2. There exists a vacuum electromagnetic polarization tensor $M_{\alpha\beta}$ which has a geometrical interpretation as torsion in the fifth dimension. Thus, there are two electromagnetic field strength tensors $F_{\alpha\beta}$ and $H_{\alpha\beta}$.

3. There is an additional term for the Lorentz force in the equation of motion for a test particle:

$$(q/m_0)g^{[\gamma\alpha]}H_{\gamma\beta}U^{\beta},$$

where q is the charge of the test particle and m_0 is its rest mass. This term plays the role of a reaction force for nonholonomic constraints.[1]

4. A new traceless energy-momentum tensor $T^{\rm em}_{\alpha\beta}$ appears for the electromagnetic field.

5. There exists a source for the electromagnetic field, i.e., the conserved current j^{α}.

All of the above effects vanish when the metric of spacetime is symmetric, in which case we get the classical Kaluza–Klein theory. Moreover, the new effects do not contradict any experimental or observational data.[1] The nonsymmetric Kaluza–Klein theory has a well-defined linear approximation.[11] In the electromagnetic case it has been shown[11] that there is no coupling between skewon and electromagnetic fields up to the first order in $h_{\mu\nu}\equiv g_{\mu\nu}-\eta_{\mu\nu}$ (where $\eta_{\mu\nu}$ is the Minkowski tensor). The nonsymmetric Kaluza–Klein theory also has a well-defined geometry on the five-dimensional manifold, which one calls Einstein geometry.[1] When the electromagnetic field vanishes, we get Moffat's nonsymmetric gravitation theory (NGT) which is able to fit the perihelion shift of Mercury in the presence of a nonzero quadrupole moment of mass for the sun.[12,13]

It is possible to extend the formalism of the nonsymmetric Kaluza–Klein theory to the nonabelian case[2,6] (including such features as spontaneous symmetry breaking and the Higgs mechanism) as well as to the Jordan–Thiry case[4,5,7], which possesses a scalar field connected to the gravitational constant. Material sources have also been incorporated[3] into this formalism.

It is of course important to find significant physical consequences of the "interference effects" present in the nonsymmetric Kaluza–Klein theory. The best way to achieve this is to find an exact solution of the full field equations, and this is the aim of this paper. We find an exact solution of the field equations in the static, spherically symmetric case in the form suggested in Sec. 6 of Ref. 1. Even in this, the simplest case, we get the following interesting results:

1. The electric field is nonsingular at $r=0$ and has Coulomb like behavior for large r. This is similar to the situation in Born–Infeld electrodynamics.[14] Thus, there is a maximal value of the electric field.

2. Asymptotically (for large r) the full solution behaves like the charged Reissner–Nördström type solution in NGT.[10]

3. The Newtonian mass is constructed from an electric charge Q and from a fermion charge l.

4. The energy distribution is not singular and is negative in a small region around $r=0$. This means that the solution describes a bounded system of electromagnetic and gravitational fields.

5. The total mass (i.e., total energy) of the solution is greater than the Newtonian mass (the mass which is seen at infinity).

6. There is no singularity at $r=0$ in the function $\alpha=g_{11}$; that is, $g_{11}(r=0)=1$.

7. The only singularities at $r=0$ are in $\omega\equiv g_{[14]}=l^2/r^2$ and in a factor $(1+l^4/r^4)$ in the function $\gamma=g_{44}$. There is also the usual singularity in the determinant of the full nonsymmetric tensor $\sqrt{-g}=r^2\sin\theta$ at $r=0$.

8. The charge distribution is nonsingular.

9. For sufficiently large charge Q there exists one or two

[a] On leave of absence from the Institute of Philosophy and Sociology of the Polish Academy of Sciences, 00-330 Warsaw, Nowy Swiat 72, Poland.

event horizons, just as in the Reissner–Nördström solution to the Einstein–Maxwell equations. Sufficiently large charge in the present case means sufficiently large Newtonian mass as well.

This solution is interesting as a classical model of a charged particle constructed from gravitational and electromagnetic fields. If we suppose that the Newtonian mass of our solution is the mass of an electron, we get a relationship between the classical radius of an electron and the parameter l from Moffat's theory of gravitation. The most facinating aspect of our solution is that it describes "mass without mass" and "charge without charge" in the following sense. At the origin $r = 0$ (or anywhere) there are no Coulomb-like or Newton-like first- and second-order poles with charge and mass as residues. This is true for the metric and for the electric field.

The paper is organized as follows. In the first section we describe some elements of the nonsymmetric Kaluza–Klein theory. The second section deals with the spherically symmetric fields in the nonsymmetric Kaluza–Klein theory, and presents the field equations in this case. The third section is devoted to the exact, static, spherically symmetric solution of the nonsymmetric Kaluza–Klein theory. We find this solution and examine its properties. In the fourth section we discuss our conclusions and prospects for further research. Appendices A and B contain some details of calculations; in Appendix A we derive the Ricci tensor in the general (nonstatic) spherically symmetric case, while in Appendix B we deal with some details concerning the static, spherically symmetric case. In Appendix C we write down the coefficients of the connection $\overline{\Gamma}$ and the Christoffel symbols for our solution as well as the equations of motion for uncharged and charged test particles.

1. ELEMENTS OF THE NONSYMMETRIC KALUZA–KLEIN THEORY

Let P be a principal fiber bundle with structural group $G = \mathrm{U}(1)$ over space-time E with projection π and let us define on this bundle a connection α. We call this bundle an electromagnetic bundle and α an electromagnetic connection. We define a curvature 2-form for the connection α:

$$\Omega = d\alpha = \tfrac{1}{2}\pi^*(F_{\mu\nu}\overline{\theta}^{\mu}\wedge\overline{\theta}^{\nu}), \tag{1.1}$$

where

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \quad e^*\alpha = A_\mu\overline{\theta}^{\mu}. \tag{1.2}$$

A_μ is a 4-potential of the electromagnetic field, e is a local section of $\underline{P}$, $F_{\mu\nu}$ is an electromagnetic field strength, and $\overline{\theta}^{\mu}$ is a frame on E. Bianchi's identity is

$$d\Omega = 0, \tag{1.3}$$

so that the 4-potential exists. This is of course simply the first Maxwell equation. On space-time E we define a nonsymmetric metric tensor $g_{\alpha\beta}$ such that

$$\begin{aligned} &g_{\alpha\beta} = g_{(\alpha\beta)} + g_{[\alpha\beta]}, \\ &g_{\alpha\beta}g^{\gamma\beta} = g_{\beta\alpha}g^{\beta\gamma} = \delta^{\gamma}_{\alpha}, \end{aligned} \tag{1.4}$$

where the order of indices is important. We define also on E two connections $\overline{\omega}^{\alpha}{}_{\beta}$ and $\overline{W}^{\alpha}{}_{\beta}$:

$$\overline{\omega}^{\alpha}{}_{\beta} = \overline{\Gamma}^{\alpha}_{\beta\gamma}\overline{\theta}^{\gamma} \tag{1.5}$$

and

$$\overline{W}^{\alpha}{}_{\beta} = \overline{W}^{\alpha}{}_{\beta\gamma}\overline{\theta}^{\alpha},$$

such that

$$\overline{W}^{\alpha}{}_{\beta} = \overline{\omega}^{\alpha}{}_{\beta} - \tfrac{2}{3}\delta^{\alpha}_{\beta}\overline{W}, \tag{1.6}$$

where

$$\overline{W} = \overline{W}_{\gamma}\overline{\theta}^{\gamma} = \tfrac{1}{2}(\overline{W}^{\sigma}_{\cdot\gamma\sigma} - \overline{W}^{\sigma}_{\cdot\sigma\gamma})\overline{\theta}^{\gamma}.$$

For the connection $\overline{\omega}^{\alpha}{}_{\beta}$ we suppose the following conditions:

$$\begin{aligned} &\overline{D}g_{\alpha+\beta-} = \overline{D}g_{\alpha\beta} - g_{\alpha\delta}\overline{Q}^{\delta}_{\beta\gamma}(\overline{\Gamma})\overline{\theta}^{\gamma} = 0, \\ &\overline{Q}^{\alpha}{}_{\beta\alpha}(\overline{\Gamma}) = 0, \end{aligned} \tag{1.7}$$

where $\overline{D}$ is the exterior covariant derivative with respect to $\overline{\omega}^{\alpha}{}_{\beta}$ and $\overline{Q}^{\alpha}{}_{\beta\gamma}(\overline{\Gamma})$ is the torsion of $\overline{\omega}^{\alpha}{}_{\beta}$. Thus we have defined on space-time E all quantities present in Moffat's theory of gravitation (see Refs. 8–10). Let us introduce on $\underline{P}$ a frame

$$\theta^{A} = (\pi^*(\overline{\theta}^{\alpha}), \lambda\alpha = \theta^{5}). \tag{1.8}$$

Now we turn to the natural nonsymmetric metrization of the bundle $\underline{P}$. According to Refs. 1–3 we have

$$\begin{aligned} &\overline{\gamma} = \pi^*\overline{g} - \theta^5\otimes\theta^5 = \pi^*(g_{(\alpha\beta)}\overline{\theta}^{\alpha}\otimes\overline{\theta}^{\beta}) - \theta^5\otimes\theta^5, \\ &\gamma = \pi^*g = \pi^*(g_{[\alpha\beta]}\overline{\theta}^{\alpha}\wedge\overline{\theta}^{\beta}). \end{aligned} \tag{1.9}$$

From the classical Kaluza–Klein theory we know that $\lambda = 2\sqrt{G}/c^2$ (see Ref. 1). We work with a system of units such that $G = c = 1$ and $\lambda = 2$. We have

$$\gamma_{AB} = \left(\begin{array}{c|c} g_{\alpha\beta} & 0 \\ \hline 0 & -1 \end{array}\right), \tag{1.10}$$

where

$$\gamma_{AB} = \gamma_{(AB)} + \gamma_{[AB]} \tag{1.11}$$

and

$$\overline{\gamma} = \gamma_{(AB)}\theta^{A}\otimes\theta^{B}, \tag{1.12}$$

$$\gamma = \gamma_{[AB]}\theta^{A}\wedge\theta^{B} \tag{1.13}$$

(see Refs. 1–3 for more details). Now we define on $\underline{P}$ a connection $\omega^{A}{}_{B}$ such that

$$D\gamma_{A+B-} = D\gamma_{AB} - \gamma_{AD}Q^{D}_{BC}(\Gamma)\theta^{C} = 0, \tag{1.14}$$

which is invariant with respect to the action of the group U(1) on $\underline{P}$. D is the exterior covariant derivative with respect to the connection $\omega^{A}{}_{B}$ and $Q^{D}{}_{BC}(\Gamma)$ is the tensor of torsion for the connection $\omega^{A}{}_{B}$. In Refs. 1 and 2 it is shown that

$$\omega^{A}{}_{B} = \left(\begin{array}{c|c} \pi^*(\overline{\omega}^{\alpha}{}_{\beta}) + g^{\gamma\alpha}H_{\gamma\beta}\theta^{5} & H_{\beta\gamma}\theta_{\gamma} \\ \hline g^{\alpha\beta}(H_{\gamma\beta} + 2F_{\beta\gamma})\theta^{\gamma} & 0 \end{array}\right), \tag{1.15}$$

where $H_{\beta\gamma}$ is a tensor on E such that

$$g_{\delta\beta}g^{\gamma\delta}H_{\gamma\alpha} + g_{\alpha\delta}g^{\delta\gamma}H_{\beta\gamma} = 2g_{\alpha\delta}g^{\delta\gamma}F_{\beta\gamma}. \tag{1.16}$$

In order to get the usual interpretation of geodesics in the classical Kaluza–Klein theory we must assume[1–3]

$$H_{\alpha\beta} = -H_{\beta\alpha}. \tag{1.17}$$

We define on $\underline{P}$ a second connection

$$W^{A}{}_{B} = \left(\begin{array}{c|c} \pi^{*}(\overline{W}^{\alpha}{}_{\beta}) + g^{\gamma\alpha}H_{\gamma\beta}\theta^{5} & H_{\beta\gamma}\theta^{\gamma} \\ \hline g^{\alpha\beta}(H_{\gamma\beta} + 2F_{\beta\gamma})\theta^{\gamma} & 0 \end{array}\right) \tag{1.18}$$

Let us define a Moffat–Ricci curvature scalar for $W^{A}{}_{B}$. One gets[1-3]

$$R(W) = \overline{R}(\overline{W}) + (2(g^{[\mu\nu]}F_{\mu\nu})^{2} - H^{\mu\alpha}F_{\mu\alpha}), \tag{1.19}$$

where

$$\overline{R}(\overline{W}) = g^{\mu\nu}\overline{R}_{\mu\nu}(\overline{\Gamma}) + \tfrac{2}{3}g^{[\beta\mu]}\overline{W}_{[\beta,\mu]} \tag{1.20a}$$

is a Moffat–Ricci curvature scalar for the connection $\overline{W}^{\alpha}{}_{\beta}$ and $\overline{R}_{\alpha\beta}(\overline{\Gamma})$ is a Moffat–Ricci curvature for the connection $\overline{\omega}^{\alpha}{}_{\beta}$. In particular,

$$\overline{R}_{\mu\nu}(\overline{\Gamma}) = \overline{R}^{\alpha}{}_{\mu\nu\alpha}(\overline{\Gamma}) + \tfrac{1}{2}\overline{R}^{\alpha}{}_{\alpha\mu\nu}(\overline{\Gamma}), \tag{1.20b}$$

where $\overline{R}^{\alpha}{}_{\mu\nu\rho}(\overline{\Gamma})$ are the components of the ordinary curvature tensor for $\overline{\Gamma}$. In addition

$$H^{\mu\alpha} = g^{\beta\mu}g^{\gamma\alpha}H_{\beta\gamma}. \tag{1.21}$$

From Eq. (1.19) one gets the field equations[1]

$$\overline{R}_{\alpha\beta}(\overline{W}) - \tfrac{1}{2}g_{\alpha\beta}\overline{R}(\overline{W}) = 8\pi T^{\mathrm{em}}_{\alpha\beta}, \tag{1.22}$$

$$\mathbf{g}^{[\mu\nu]}{}_{,\nu} = 0, \tag{1.23}$$

$$g_{\mu\nu,\sigma} - g_{\zeta\nu}\overline{\Gamma}^{\zeta}_{\mu\sigma} - g_{\mu\zeta}\overline{\Gamma}^{\zeta}_{\sigma\nu} = 0, \tag{1.24}$$

$$\partial_{\mu}(\mathbf{H}^{\alpha\mu}) = 4\mathbf{g}^{[\alpha\beta]}\partial_{\beta}(g^{[\mu\nu]}F_{\mu\nu}), \tag{1.25}$$

where

$$T^{\mathrm{em}}_{\alpha\beta} = (1/4\pi)(g^{\gamma\mu}H_{\gamma\alpha}F_{\mu\beta} - 2g^{[\mu\nu]}F_{\mu\nu}F_{\alpha\beta} - \tfrac{1}{4}g_{\alpha\beta}(H^{\mu\nu}F_{\mu\nu} - 2(g^{[\mu\nu]}F_{\mu\nu})^{2})), \tag{1.26}$$

$$\mathbf{g}^{[\mu\nu]} = \sqrt{-g}\,g^{[\mu\nu]},$$

$$\mathbf{H}^{\mu\alpha} = \sqrt{-g}\,g^{\beta\mu}g^{\gamma\alpha}H_{\beta\gamma}. \tag{1.27}$$

The tensor $H_{\mu\nu}$ has an interpretation as a second electromagnetic field strength tensor.[1-3] We have

$$g^{\alpha\beta}T^{\mathrm{em}}_{\alpha\beta} = 0. \tag{1.28}$$

Equations (1.22)–(1.25) can be written in the form

$$\overline{R}_{(\alpha\beta)}(\overline{\Gamma}) = 8\pi T^{\mathrm{em}}_{(\alpha\beta)}, \tag{1.29}$$

$$\overline{R}_{[[\alpha\beta],\gamma]}(\overline{\Gamma}) - 8\pi T^{\mathrm{em}}_{[[\alpha\beta],\gamma]} = 0, \tag{1.30}$$

$$\overline{\Gamma}_{\mu} = 0, \tag{1.31}$$

$$g_{\mu\nu,\sigma} - g_{\zeta\nu}\overline{\Gamma}^{\zeta}_{\mu\sigma} - g_{\mu\zeta}\overline{\Gamma}^{\zeta}_{\sigma\nu} = 0, \tag{1.32}$$

$$\partial_{\mu}(\mathbf{H}^{\alpha\mu} - 4\mathbf{g}^{[\alpha\mu]}(g^{[\nu\beta]}F_{\nu\beta})) = 0, \tag{1.33}$$

where $\overline{R}_{\alpha\beta}(\overline{\Gamma})$ is a Moffat–Ricci tensor for the connection

$$\overline{\omega}^{\alpha}{}_{\beta} = \overline{\Gamma}^{\alpha}_{\beta\gamma}\overline{\theta}^{\gamma}, \qquad \overline{\Gamma}_{\mu} = \overline{\Gamma}^{\alpha}_{[\mu\alpha]}. \tag{1.34}$$

The condition (1.31) is equivalent to (1.23).

2. SPHERICALLY SYMMETRIC FIELDS IN THE NONSYMMETRIC KALUZA–KLEIN THEORY

Let us suppose that the fundamental fields in the nonsymmetric Kaluza–Klein theory possesses spherical symmetry. According to Refs. 15–23 one gets

$$g_{\mu\nu} = \begin{pmatrix} -\alpha & 0 & 0 & \omega \\ 0 & -\beta & f\sin\theta & 0 \\ 0 & -f\sin\theta & -\beta\sin^{2}\theta & 0 \\ -\omega & 0 & 0 & \gamma \end{pmatrix}, \tag{2.1}$$

where α, β, γ, f, and ω are real functions of r and t with $\alpha, \gamma > 0$. In addition

$$F_{14} = E(r,t), \quad F_{23} = B\sin\theta \tag{2.2}$$

and all other components of $F_{\mu\nu}$ vanish. For $g^{\mu\nu}$, the only nonvanishing components are

$$\begin{aligned} g^{11} &= \gamma/(\omega^{2} - \alpha\gamma), \\ g^{22} &= g^{23}\sin^{2}\theta = -\beta/(\beta^{2} + f^{2}), \\ g^{44} &= -\alpha/(\omega^{2} - \alpha\gamma), \\ g^{[14]} &= \omega/(\omega^{2} - \alpha\gamma), \\ g^{[23]}\sin\theta &= f/(\beta^{2} + f^{2}). \end{aligned} \tag{2.3}$$

We suppose that

$$\omega^{2} - \alpha\gamma \neq 0 \quad \text{and} \quad \beta^{2} + f^{2} \neq 0. \tag{2.4}$$

Let us suppose that $H_{\alpha\beta}$ is also spherically symmetrical, so that

$$H_{14} = D(r,t), \quad H_{23} = H\sin\theta \tag{2.5}$$

and the other components vanish. Using Eqs. (1.16), (2.1), and (2.3) it can be shown that

$$\begin{aligned} H_{14} &= F_{14} = E(r,t), \\ H_{23} &= F_{23} = B\sin\theta. \end{aligned} \tag{2.6}$$

The Bianchi identity equation (1.3) yields

$$B = B_{0} = \text{const}. \tag{2.7}$$

From Eq. (1.23) one gets

$$\frac{\omega^{2}}{\alpha\gamma - \omega^{2}} = \frac{l^{4}}{\beta^{2} + f^{2}}, \tag{2.8}$$

where l is a constant of integration. In Moffat's theory of gravitation this constant has an interpretation as fermion charge. From Eq. (1.33) we have

$$\frac{E}{\omega} = \frac{-(Q/l^{2})(\beta^{2} + f^{2}) + 8fB_{0}}{(\beta^{2} + f^{2} + 8f^{4})}, \tag{2.9}$$

where Q is an integration constant. In the intermediate stages of calculation we used the following expressions for $H^{\mu\alpha}$ and $\sqrt{-g}$

$$H^{14} = -\frac{H_{14}}{(\alpha\gamma - \omega^{2})} = \frac{-E}{(\alpha\gamma - \omega^{2})}, \tag{2.10}$$

$$H^{23} = \frac{B_{0}}{\beta^{2} + f^{2}}, \tag{2.11}$$

$$\sqrt{-g} = \sin\theta\,[(\alpha\gamma - \omega^{2})(\beta^{2} + f^{2})]^{1/2}. \tag{2.12}$$

Thus finally we get Eqs. (1.29)–(1.32) plus the algebraic relations (2.7)–(2.9). From Eq. (1.30) we get immediatelly

$$R_{[23]}(\overline{\Gamma}) - 8\pi T^{\mathrm{em}}_{[23]} = C_{1}\sin\theta, \tag{2.13}$$

where $C_{1} = \text{const}$ is an integration constant and

$$\frac{8\pi}{\sin\theta}T^{\rm em}_{[23]} = -\frac{7fB_0^2}{\beta^2+f^2} + \frac{fl^4}{(\beta^2+f^2)}\left(\frac{E}{\omega}\right)^2 + 4f\left(\frac{fB_0}{(\beta^2+f^2)} - \frac{l^4}{(\beta^2+f^2)}\frac{E}{\omega}\right)^2 + \frac{8B_0l^4}{(\beta^2+f^2)}\left(\frac{E}{\omega}\right). \quad (2.14)$$

Equations (1.31) and (1.32) were solved in Ref. 17 in which Pant wrote down the Ricci tensor for such a connection.

Note that the Moffat–Ricci tensor [Eq. (1.20b)] is a linear combination of the ordinary Ricci tensor and the second contraction of the curvature tensor. However, Eqs. (1.23) and (1.24) imply that[10]

$$\bar{\Gamma}^{\alpha}_{[\mu\alpha]} = 0 \quad (2.15)$$

and

$$\bar{\Gamma}^{\beta}_{\nu\beta} = [\ln((-g)^{1/2})]_{,\nu} \quad (2.16)$$

so that the second contraction

$$\bar{R}^{\alpha}_{\alpha\mu\nu} = \tfrac{1}{2}(\bar{\Gamma}^{\beta}_{(\mu\beta),\nu} - \bar{\Gamma}^{\beta}_{(\nu\beta),\nu}) = 0. \quad (2.17)$$

Consequently the Moffat–Ricci tensor in this case is identically equal to the ordinary Ricci tensor used by Pant,[17] which we shall denote by $A_{\mu\nu}(\bar{\Gamma})$.

Thus we get the equations

$$A_{(\mu\nu)}(\bar{\Gamma}) = 8\pi T^{\rm em}_{(\mu\nu)}, \quad (2.18)$$

$$A_{[23]}(\bar{\Gamma}) - 8\pi T^{\rm em}_{[23]} = C_1\sin\theta,$$

where

$$8\pi T^{\rm em}_{11} = \alpha\left(\frac{l^4}{\beta^2+f^2}\right)\frac{E^2}{\omega^2} + \frac{\alpha B_0^2}{\beta^2+f^2} - 4\alpha\left(\frac{fB_0}{\beta^2+f^2} - \frac{l^4}{(\beta^2+f^2)}\frac{E}{\omega}\right)^2. \quad (2.19)$$

Using Eq. (2.9), the last term in Eq. (2.19) can be written in the form

$$-4\alpha\left(\frac{fB_0+Ql^2}{\beta^2+f^2+8l^4}\right)^2. \quad (2.20)$$

Moreover, it can be shown that

$$8\pi T^{\rm em}_{44} = -(\gamma/\alpha)8\pi T^{\rm em}_{11}, \quad (2.21)$$

$$8\pi T^{\rm em}_{22} = \frac{8\pi}{\sin^2\theta}T^{\rm em}_{33} = \frac{\beta}{\alpha}8\pi T^{\rm em}_{11} = -\frac{\beta B_0^2}{(\beta^2+f^2)} - \frac{\beta l^4}{(\beta^2+f^2)}\left(\frac{E^2}{\omega^2}\right) - 4\beta\left(\frac{fB_0}{(\beta^2+f^2)} - \frac{l^4}{(\beta^2+f^2)}\frac{E}{\omega}\right)^2, \quad (2.22)$$

$$8\pi T^{\rm em}_{14} = -8\pi T^{\rm em}_{41} = \frac{\omega}{(\beta^2+f^2)}\left(7l^4\left(\frac{E^2}{\omega^2}\right) - 8fB_0\left(\frac{E}{\omega}\right) - B_0^2\right) - \omega\left(\frac{fB_0}{(\beta^2+f^2)} - \frac{l^4}{(\beta^2+f^2)}\left(\frac{E}{\omega}\right)\right)^2. \quad (2.23)$$

The rest of the components of $T^{\rm em}_{\mu\nu}$ vanish. The electromagnetic Lagrangian in this case is

$$\mathscr{L}_{\rm em} = \frac{1}{8\pi}(2(g^{[\mu\nu]}F_{\mu\nu})^2 - H^{\mu\nu}F_{\mu\nu}) = \frac{1}{8\pi}\left[\frac{8\omega^4}{(\alpha\gamma-\omega^2)^2}\left(\frac{fB_0}{l^4} - \frac{E}{\omega}\right)^2 - \frac{2\omega^2}{(\alpha\gamma-\omega^2)}\left(\frac{B_0^2}{l^4} - \frac{E^2}{\omega^2}\right)\right]. \quad (2.24)$$

Finally, we have the following equations:

$$A_{11}(\bar{\Gamma}) = 8\pi T^{\rm em}_{11}, \quad (2.25a)$$

$$A_{44}(\bar{\Gamma}) = 8\pi T^{\rm em}_{44}, \quad (2.25b)$$

$$A_{22}(\bar{\Gamma}) = 8\pi T^{\rm em}_{22}, \quad (2.25c)$$

$$A_{33}(\bar{\Gamma}) = 8\pi T^{\rm em}_{33}, \quad (2.25d)$$

$$A_{[23]}(\bar{\Gamma}) - 8\pi T^{\rm em}_{23} = C_1\sin\theta, \quad (2.25e)$$

$$A_{(14)}(\bar{\Gamma}) = 0. \quad (2.25f)$$

Using results from Ref. 17 and Eq. (2.22) one finds the identity (see Appendix A)

$$(A_{22}(\bar{\Gamma}) - 8\pi T^{\rm em}_{22}) \equiv (1/\sin^2\theta)(A_{33}(\bar{\Gamma}) - 8\pi T^{\rm em}_{33}), \quad (2.26)$$

so that Eq. (2.25d) is not independent. In the above

$$8\pi T^{\rm em}_{11} = \alpha\left[\frac{(8l^2fB_0 - Q(\beta^2+f^2))^2 + B_0^2(\beta^2+f^2+8l^4)^2 - (fB_0+Ql^2)(\beta^2+f^2)}{(\beta^2+f^2)(\beta^2+f^2+8l^4)^2}\right], \quad (2.27)$$

$$\frac{8\pi}{\sin\theta}T^{\rm em}_{23} = \frac{8\pi}{\sin\theta}T^{\rm em}_{[23]} = \frac{[-7fB_0(\beta^2+f^2+8l^4)^2 - f(8fB_0 - Q(\beta^2+f^2))^2]}{(\beta^2+f^2)(\beta^2+f^2+8l^4)^2} + \frac{[8B_0l^4(8B_0l^2 - Q(\beta^2+f^2))(\beta^2+f^2+8l^4) + 4f(\beta^2+f^2)(fB_0+Ql^2)^2]}{(\beta^2+f^2)(\beta^2+f^2+8l^4)^2}. \quad (2.28)$$

For $T^{\rm em}_{14}$ one finds

$$8\pi T^{\rm em}_{14} = 8\pi T^{\rm em}_{[14]} = \frac{\omega}{(\beta^2+f^2)}\frac{[7l^2(8l^2fB_0 - Q(\beta^2+f^2))^2 - 8B_0f(8l^2fB_0 - Q(\beta^2+f^2))(-l^2B_0(\beta^2+f^2+8l^4)^2]}{l^2(\beta^2+f^2+8l^4)} \quad (2.29)$$

$A_{11}(\bar{\Gamma})$, $A_{44}(\bar{\Gamma})$, $A_{33}(\bar{\Gamma})$, $A_{(14)}(\bar{\Gamma})$, $A_{[14]}(\bar{\Gamma})$, and $A_{[23]}(\bar{\Gamma})$ are given by the formulas (2.11) (see Appendix A) from Ref. 17. For $\mathscr{L}_{\rm em}$ one easily gets, using (2.24),

$$\mathscr{L}_{\rm em} = \frac{1}{4\pi}\frac{l^4}{(\beta^2+f^2)}\left[\frac{4^4}{(\beta^2+f^2)}\left(\frac{fB_0}{l^2} - \frac{(8l^2fB_0 - Q(\beta^2+f^2))}{(\beta^2+f^2+8l^4)}\right)^2 - \frac{1}{l^4}\left(B_0^2 - \frac{(8fl^2B_0 - Q(\beta^2+f^2))^2}{(\beta^2+f^2+8l^4)^2}\right)\right]. \quad (2.30)$$

3. STATIC, SPHERICALLY SYMMETRIC SOLUTION

Let us consider a spherical field configuration such that

$$B_0 = f = 0. \tag{3.1}$$

Later we suppose that

$$\beta = r^2, \tag{3.2}$$

which is simply a coordinate choice. In addition, our quantities do not depend on time (static case). One finds [see Eq. (2.9)]

$$E = -\omega \frac{Q}{l^2}\left(\frac{r^4}{r^4 + 8l^2}\right) \tag{3.3}$$

(substituting $\beta = r^2$). Equations (2.29) now read

$$\begin{aligned} &A_{11}(\bar{\Gamma}) - \frac{\alpha Q^2}{(\beta^2 + 8l^4)}\left(\frac{\beta^2 - 4l^4}{\beta^2 + 8l^4}\right) = 0, \\ &A_{44}(\bar{\Gamma}) + \frac{\gamma Q^2}{(\beta^2 + 8l^4)}\frac{(\beta^2 - 4l^4)}{(\beta^2 + 8l^4)} = 0, \\ &A_{22}(\bar{\Gamma}) - \frac{\beta Q^2}{(\beta^2 + 8l^4)}\frac{(\beta^2 - 4l^4)}{(\beta^2 + 8l^4)} = 0, \\ &A_{(14)} = 0, \\ &A_{[23]} - 8\pi T^{\text{em}}_{23} = C_1 \sin\theta, \end{aligned} \tag{3.4}$$

and we have

$$8\pi T^{\text{em}}_{[14]} = 8\pi T^{\text{em}}_{14} = \omega Q \frac{(7\beta^2 + 16l^4)}{(\beta^2 + 8l^4)^2}, \tag{3.5}$$

$$\omega = l^2/r^2. \tag{3.6}$$

One gets

$$E = -\frac{Q}{r^2}\left(\frac{r^4}{r^4 + 8l^4}\right). \tag{3.7}$$

It is easy to see that the function (3.7) is bounded

$$|E| < E_{\max} = |E(\sqrt[4]{8}l)| = |Q|/8l^2. \tag{3.7a}$$

From (3.4), using results from Ref. 17, one gets (see Appendix B)

$$\frac{d}{dr}(r\alpha^{-1}) = 1 - Q^2 r^2 \frac{(r^4 + 4l^4)}{(r^4 + 8l^4)^2}. \tag{3.8}$$

Thus we have

$$\frac{1}{\alpha} = 1 + \frac{C}{r} + \frac{Q^2}{r}K(r,l), \tag{3.9}$$

where

$$K(r,l) = -\int r^2 \frac{(r^4 + 4l^4)}{(r^4 + 8l^4)^2}\,dr \tag{3.10}$$

and C is a constant of integration. Moreover,

$$\gamma = \left(1 + \frac{C}{r} + \frac{Q^2}{r}K(r,l)\right)\left(1 + \frac{l^4}{r^4}\right) \tag{3.11}$$

[see Eqs. (B8) and (B11) in Appendix B]. Performing the integration in (3.10) one gets

$$\frac{1}{\alpha} = 1 + \frac{C}{r} + \frac{Q^2 b}{b^2 r}g\left(\frac{b}{r}\right), \tag{3.12}$$

where $b^4 = 8l^4$ and

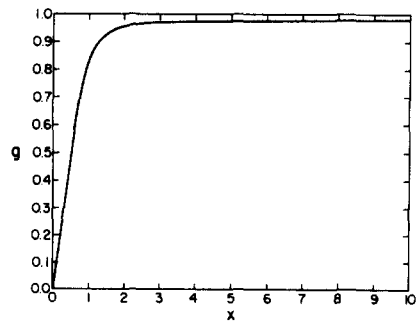

FIG. 1. The function $g = g(x)$ vs x.

$$g(x) = \frac{1}{8}\left(\frac{x}{x^4 + 1}\right) + \frac{7}{32\sqrt{2}}\left(\log\left(\frac{x^2 + \sqrt{2}x + 1}{x^2 - \sqrt{2}x + 1}\right) + 2\arctan(\sqrt{2}x + 1) + 2\arctan(\sqrt{2}x - 1)\right). \tag{3.13}$$

The function $g(x)$ is plotted in Fig. 1. Let us examine the properties of the function

$$g(b/r).$$

It can be shown that

$$\lim_{r=0} g\left(\frac{b}{r}\right) = \frac{7}{16}\frac{\pi}{\sqrt{2}}. \tag{3.14}$$

Thus for small r we get

$$\alpha^{-1} \simeq 1 + \frac{1}{r}\left(C + \frac{7}{16\sqrt{2}}\pi\left(\frac{Q^2}{b}\right)\right). \tag{3.15}$$

We can avoid a singularity in α at $r = 0$ by choosing

$$C = -\frac{7}{16\sqrt{2}}\pi\left(\frac{Q^2}{b}\right) \tag{3.16}$$

so that

$$\lim_{r=0}(\alpha^{-1}) = 1. \tag{3.17}$$

Let us examine the asymptotic properties of α and γ. One gets

$$\alpha^{-1} \underset{r\to\infty}{\to} \left(1 - \frac{[(7/16\sqrt{2}\pi)Q^2/b]}{r} + \frac{Q^2}{r^2}\right). \tag{3.18}$$

For large r, α clearly behaves like the analogous function in the Reissner–Nördström solution, with Q as the electric charge and with

$$m_{\text{N}} = \frac{7}{32\sqrt{2}}\pi\left(\frac{Q^2}{b}\right) \tag{3.19}$$

playing the role of the Newtonian mass. To summarize, we have

$$\alpha^{-1} = \left(1 - \frac{7}{8\sqrt{2}}\left(\frac{\pi}{2}\right)\frac{Q^2/b}{r} + \frac{Q^2}{r^2}\tilde{g}\left(\frac{b}{r}\right)\right), \tag{3.20}$$

where

$$\lim_{r\to\infty}\tilde{g}\left(\frac{b}{r}\right) = 1, \tag{3.21}$$

$$\tilde{g}\left(\frac{b}{r}\right) = \frac{g(b/r)}{(b/r)}, \tag{3.21a}$$

and

$$\lim_{r\to 0} \alpha^{-1} = 1. \tag{3.22}$$

In the neighborhood of $r = 0$ one gets for our metric

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & \frac{l^2}{r^2} \\ 0 & -r^2 & 0 & 0 \\ 0 & 0 & -r^2 \sin\theta & 0 \\ -\frac{l^2}{r^2} & 0 & 0 & \left(1+\frac{l^4}{r^4}\right) \end{pmatrix} \tag{3.23}$$

(for $r\to 0$). The determinant of the symmetric part of the metric is

$$(-\tilde{g})^{1/2} = (r^4 + l^4)^{1/2} \sin\theta. \tag{3.24}$$

The full determinant is

$$\sqrt{-g} = r^2 \sin\theta. \tag{3.25}$$

Thus there is a singularity at $r = 0$. It is worth noting, however, that there is no singularity in α and only one singularity in γ due to the $(1 + l^4/r^4)$ factor. ω, the skew-symmetric part of $g_{\mu\nu}$, is also singular at $r = 0$.

Let us examine properties of the electric field:

$$E = -\frac{Q}{r^2}\left(\frac{r^4}{r^4 + 8l^4}\right). \tag{3.26}$$

One easily sees that

$$E(0) = 0 \tag{3.27}$$

and

$$E \underset{r\to\infty}{\to} -\frac{Q}{r^2}. \tag{3.28}$$

Thus there is no singularity at $r = 0$. This is similar to the situation in Born–Infeld electrodynamics.[14] Let us calculate the charge distribution and total charge for the electric field. It is known that

$$4\pi\sqrt{-g}\rho = \mathbf{H}^{4i}{}_{,i} \sim \operatorname{div}\vec{D}, \tag{3.29}$$

where ρ is the charge density distribution and $\vec{D}$ is an electric induction vector. One gets

$$\mathbf{H}^{41} = \sqrt{-g}E/(\alpha\gamma - \omega^2) = \sqrt{-g}E \tag{3.30}$$

and

$$\sqrt{-g}\rho = -\frac{1}{\pi}\frac{Q}{r}\frac{(8l^4r^4)}{(r^4 + 8l^4)^2}\sin\theta. \tag{3.31}$$

The total charge is

$$Q_{\text{tot}} = \int \sqrt{-g}\rho\, d^3x = -32Ql^4 \int_0^\infty \frac{1}{r}\frac{r^4}{(r^4 + 8l^4)^2}\, dr = -Q. \tag{3.32}$$

Thus we find the following interesting feature: the total electric charge defined above is the same as the charge obtained from the asymptotic properties of the electric field E and the metric (functions α and γ). Let us pass on the calculation of the energy of the electromagnetic field. One has

$$T^4{}_4 = \frac{1}{8\pi}Q^2\left(\frac{(r^4 - 10l^4)}{(r^4 + 8l^4)^2}\right). \tag{3.33}$$

The total energy

$$E_{\text{tot}} = 4\pi\int_0^\infty r^2 \overset{\text{em}}{T}{}^4{}_4\, dr = \left(\frac{Q^2}{b}\right)\frac{\pi}{\sqrt{2}}\left(\frac{59}{64}\right), \tag{3.34}$$

where $b^4 = 8l^4$. Thus we get that the total mass is

$$m_{\text{tot}} = \left(\frac{59}{64}\right)\left(\frac{\pi}{\sqrt{2}}\right)\left(\frac{Q^2}{b}\right). \tag{3.35}$$

and the Newtonian mass is

$$m_{\text{N}} = \left(\frac{7}{32}\right)\left(\frac{\pi}{\sqrt{2}}\right)\left(\frac{Q^2}{b}\right). \tag{3.36}$$

Thus,

$$m_{\text{N}}/m_{\text{tot}} = \tfrac{14}{59}. \tag{3.37}$$

Equation (3.37) implies that asymptotically we see only $[\frac{14}{59}]$ of the total energy as a Newtonian gravitational mass. Let us divide the total energy into two parts: Newtonian and electromagnetic. That is

$$m_{\text{tot}} = m_{\text{N}} + m_{\text{em}}. \tag{3.38}$$

One gets

$$m_{\text{em}} = \frac{\pi}{\sqrt{2}}\left(\frac{Q^2}{b}\right)\left(\frac{45}{64}\right). \tag{3.39}$$

This energy could be treated as the energy of the electric field of the charge Q distributed over a sphere of radius r_0. That is,

$$c^2 m_{\text{em}} = Q^2/r_0, \tag{3.40}$$

so that

$$r_0 = b\frac{r^2}{\pi}\left(\frac{64}{45}\right) = l^4\sqrt{2}\left(\frac{128}{45\pi}\right). \tag{3.41}$$

Let us suppose that the Newtonian mass is the mass of an electron.

$$m_{\text{N}} = m_{\text{e}}. \tag{3.42}$$

One gets

$$m_{\text{e}}c^2 = \left(\frac{Q^2}{b}\right)\left(\frac{\pi}{\sqrt{2}}\frac{7}{32}\right). \tag{3.43}$$

Thus we get

$$l = \left(\frac{7}{64}\right)\left(\frac{\pi}{\sqrt{2}}\right)\left(\frac{e^2}{m_{\text{e}}c^2}\right), \tag{3.44}$$

where e is an elementary charge. For r_0 we get similarly

$$r_0 = (\tfrac{14}{45})(e^2/m_{\text{e}}c^2). \tag{3.45}$$

The classical radius of an electron is defined as

$$r_{\text{cl}} = e^2/m_{\text{e}}c^2 \simeq 2.81\times 10^{-13}\ \text{cm}. \tag{3.46}$$

Thus we get

$$r_0 = (\tfrac{14}{45})r_{\text{cl}} \simeq 10^{-13}\ \text{cm} \tag{3.47}$$

and

$$l = \left(\frac{7}{64}\right)\left(\frac{\pi}{\sqrt[4]{2}}\right)r_{\text{cl}} \simeq 10^{-13}\ \text{cm}. \tag{3.48}$$

Let us introduce the dimensionless variables

$$q \equiv Q/b = Q/\sqrt[4]{8}l, \tag{3.49}$$

$$R \equiv r/b = r/\sqrt[4]{8}l. \tag{3.50}$$

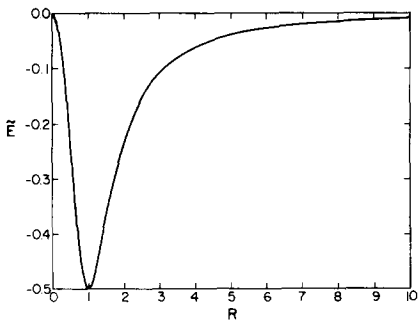

FIG. 2. The function $\tilde{E} = \tilde{E}(R)$ vs R (normalized electric field).

Using Eqs. (3.49) and (3.50) we have

$$\alpha^{-1} = \left(1 - \frac{7}{8\sqrt{2}}\left(\frac{\pi}{2}\right)\frac{q}{R} + \frac{q^2}{R^2}\tilde{g}\left(\frac{1}{R}\right)\right) = (1 - q^2 P(R)), \tag{3.51}$$

$$E = -\frac{q^2}{R^2}\left(\frac{R^4}{R^4+1}\right) = q^2\tilde{E}, \tag{3.52}$$

$$e = 4\pi T^{em}_{44} r^2 = \frac{q^2 R^2}{2}\frac{(R^4 - \frac{5}{4})}{(R^4+1)^2} = q^2\tilde{e}, \tag{3.53}$$

$$\rho_R = \frac{4\pi\rho r^2}{8l^4} = \frac{4\pi\rho r^2}{8l^4} = \frac{4\pi\rho r^2}{b^4} = -\frac{2q}{R}\left(\frac{R^4}{R^4+1}\right) = q\tilde{\rho}_R, \tag{3.54}$$

where q is a normalized charge, R is a normalized radial coordinate, and $\tilde{E}, \tilde{e}, \tilde{\rho}_R$ are normalized, electric field, radial energy distribution, and radial charge distribution, respectively. These functions are plotted in Figs. 2–4. The function

$$P(R) = \frac{1}{R}\left(-q\left(\frac{1}{R}\right) + \frac{7\pi}{16\sqrt{2}}\right) \tag{3.55}$$

is plotted in Fig. 5. It expresses the properties of the generalized Newtonian potential for our solution. Notice that the function $e \leqslant 0$ for $0 \leqslant R \leqslant \sqrt[4]{5}/\sqrt{2}$. This means that our solution corresponds to a kind of bounded system of gravitational and electromagnetic fields.

An interesting question which we can pose here concerns the existence of event horizons. This problem reduces to finding real roots for the function $\alpha^{-1} = f(R,q)$. This depends of course on the value of the parameter q. Let us consider the function

$$f(R,q) = 1 + q^2(1/R)(-7\pi/16\sqrt{2} + g(1/R)). \tag{3.56}$$

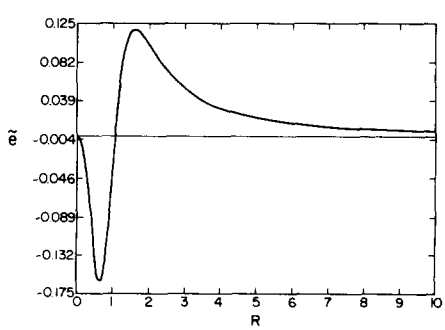

FIG. 3. The function $\tilde{e} = \tilde{e}(R)$ vs R (normalized radial energy distribution).

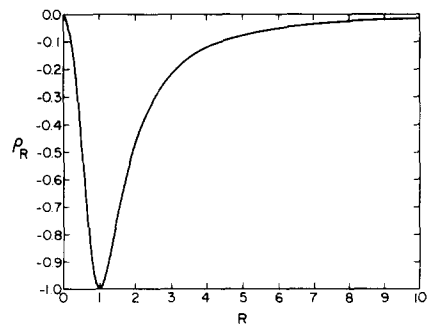

FIG. 4. The function $\tilde{\rho}_R = \tilde{\rho}_R(R)$ vs R (normalized radial charge distribution).

We have

$$f(0,q) = 1 \tag{3.57a}$$

and

$$\lim_{R\to\infty} f(R,q) = 1. \tag{3.57b}$$

Consider now the function

$$h(x) = -7\pi/16\sqrt{2} + g(x) \tag{3.58}$$

and look for a value of $x = x_1$ such that

$$h(x_1) < 0. \tag{3.59}$$

The function $g(x)$ is monotonic in the interval $(0, +\infty)$ and positive. Moreover,

$$\lim_{x=\infty} g(x) = \frac{7\pi}{16\sqrt{2}} \tag{3.60}$$

so that

$$g(1/R) < 7\pi/16\sqrt{2}. \tag{3.61}$$

Consequently,

$$h(1/R_1) < 0 \tag{3.62}$$

for every $R_1 > 0$. Let us suppose that

$$q > \frac{\sqrt{R_1}}{\sqrt{-g(1/R_1) + 7\pi/16\sqrt{2}}}. \tag{3.63}$$

It is easy to check that if (3.63) is satisfied then

$$f(q,R_1) < 0. \tag{3.64}$$

The function $f(q,R)$ changes sign in the interval $\langle 0,R_1\rangle$. This means that there exists a value $R_H \in \langle 0,R_1\rangle$ such that

$$f(q,R_H) = 0. \tag{3.65}$$

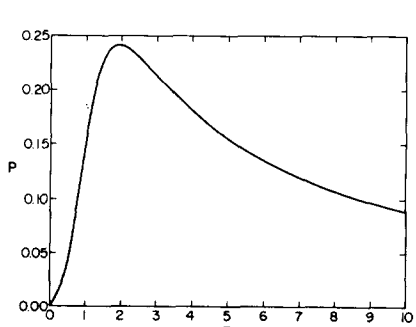

FIG. 5. The function $p = p(R)$ vs R (generalized Nördström function).

The function $f(q,R)$ changes sign in the interval $\langle R_1, +\infty)$ too. Thus there exists a value $R_{\bar{H}} \in \langle R_1, +\infty)$ such that

$$f(q,R_{\bar{H}}) = 0 \tag{3.66}$$

[if condition (3.67) is satisfied]. Hence there are two event horizons for sufficiently large q in general.

Let us examine the situation with only one event horizon. The conditions necessary for the existence of a single horizon are

$$f(q,R) = 0, \tag{3.67a}$$

$$\frac{df}{dR}(q,R) = 0. \tag{3.67b}$$

From (3.67b) one easily gets

$$\frac{1d}{R\,dr} g\left(\frac{1}{R}\right) = g\left(\frac{1}{R}\right) - \frac{7\pi}{16\sqrt{2}}. \tag{3.68}$$

Equation (3.67) is equivalent to

$$\frac{7\pi}{16\sqrt{2}} - \frac{R(R^4 + \frac{1}{2})}{(R^4+1)^2} = g\left(\frac{1}{R}\right). \tag{3.69}$$

In terms of the variable $x \equiv 1/R$ we have

$$\frac{7\pi}{16\sqrt{2}} - \frac{(x^4+2)x}{2(x^4+1)^2} = g(x). \tag{3.70}$$

The soluiton x_0 of Eq. (3.70) is

$$x_0 = 0.516\,288\,994\,64\cdots. \tag{3.71}$$

Let us solve Eq. (3.67a) with respect to q. One gets

$$q_0 = \frac{1}{\sqrt{x_0(7\pi/16\sqrt{2} - g(x_0))}} \tag{3.72}$$

or

$$q_0 = \frac{x_0(x_0^4+1)\sqrt{2x_0}}{\sqrt{x_0^4+2}}. \tag{3.73}$$

Thus there is exactly one event horizon when

$$R_H = 1/x_0 \simeq 1.9369\cdots \tag{3.74}$$

and

$$\left(\frac{r_H}{l}\right) = \frac{\sqrt[4]{8}}{x_0} \simeq 3.2575\cdots, \tag{3.73a}$$

$$q_0 = \frac{x_0(x_0^4+1)\sqrt{2x_0}}{\sqrt{x_0^4+2}} \simeq 2.038\,6231\cdots. \tag{3.75}$$

In this case we have for the Newtonian and total mass,

$$m_{\mathrm{N}}^0 = \pi^4\sqrt{2}\left(\frac{7}{16}\right)\frac{x_0^3(x_0^4+1)^2}{(x_0^4+2)}\left(\frac{c^2 l}{G}\right), \tag{3.76}$$

$$m_{\mathrm{tot}}^0 = \pi^4\sqrt{2}\left(\frac{59}{32}\right)\frac{x_0^3(x_0^4+1)^2}{(x_0^4+2)}\left(\frac{c^2 l}{G}\right), \tag{3.77}$$

or

$$m_{\mathrm{N}}^0 = 3.39\left(\frac{c^2 l}{G}\right) \cong 10^7\ \mathrm{g}, \tag{3.76a}$$

$$m_{\mathrm{tot}}^0 = 14.31\left(\frac{c^2 l}{G}\right) \cong 10^7\ \mathrm{g} \tag{3.77a}$$

for $l = 10^{-20}$ cm. The total charge is

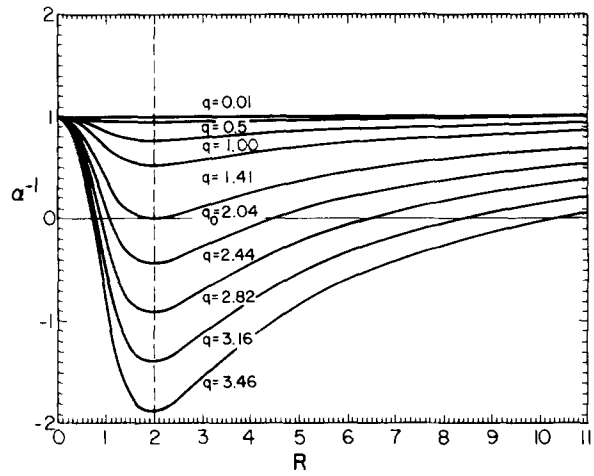

FIG. 6. The function $\alpha^{-1} = f(q,R)$ vs R for various values of parameters q. q_0 means the critical value for which we have only one event horizon for the value $R = R_H$. For the value $R = R_H$ the function $f(q,R)$ has a minimum regardless the value of q. If $q > q_0$ we have two event horizons [two real roots of $f(q,R)$, R_{H_1}, R_{H_2}]. If $q < q_0$ there are not any event horizons [no real roots for $f(q,R)$].

$$Q_0 = q_0{}^4\sqrt{8}lc^2/\sqrt{G} \simeq 2.82(lc^2/\sqrt{G})$$
$$\simeq 10^5\ \mathrm{esu} \simeq 10^{14}\ \text{elementary charges}$$
$$(\text{for } l \simeq 10^{-20}\ \mathrm{cm}). \tag{3.78}$$

It is easy to see that if $q > q_0$ we hve two horizons. This also implies that

$$m_{\mathrm{N}} > m_{\mathrm{N}}^0. \tag{3.79}$$

In other words the Newtonian mass is large enough to form event horizons. If $q = q_0$ we have only one horizon and if $q < q_0$ we have no horizons. This situation is described in Fig. 6 where we plot the function $\alpha^{-1} = f(q,R)$ for various values of the parameter q. For example for an electron one has

$$q_{\mathrm{electron}} = e\sqrt{G}/{}^4\sqrt{8}lc^2 \simeq 10^{-37} \ll q_0. \tag{3.80}$$

Thus there are no event horizons. It is worth noting that if there exists only one event horizon the solution is unstable due to pair creation and Hawking radiation. Such "black holes" are "very hot" (see Ref. 23) and decay very quickly. In the case of two event horizons the solution is unstable because of pair creation. If the Newtonian mass is sufficiently big this solution could be more stable because the Hawking effect is not important for very massive black holes (see Ref. 23). The situation without any event horizons is very interesting from a physical point of view, because it corresponds to the parameter q for electron (in general for any elementary particle). Thus we have in this case a singularity without a horizon. The structure of this singularity is different from the Nördström-like or Schwarzchild-like singularity in the nonsymmetric theory of gravitation [see Refs. 15, 16, and 23 and Eq. (3.23)].

To summarize, we have found the following exact solution (in the form suggested in Sec. 6 of Ref. 1):

$$g_{\mu\nu} = \begin{pmatrix} -\alpha & 0 & 0 & l^2/r^2 \\ 0 & -r^2 & 0 & 0 \\ 0 & 0 & -r^2\sin\theta & 0 \\ -l^2/r^2 & 0 & 0 & \gamma \end{pmatrix}, \tag{3.81}$$

$$\alpha = \left(1 - \frac{7\pi}{16\sqrt{2}}\left(\frac{Q^2}{b}\right)\frac{1}{r} + \frac{Q^2}{rb}g\left(\frac{b}{r}\right)\right)^{-1}, \quad (3.82)$$

$$\gamma = \left(1 + \frac{l^4}{r^4}\right)\left(1 - \frac{7\pi}{16\sqrt{2}}\left(\frac{Q^2}{b}\right)\frac{1}{r} + \frac{Q^2}{rb}g\left(\frac{b}{r}\right)\right), \quad (3.82a)$$

$$b^4 = 8l^4, \quad (3.83)$$

$$E = -\frac{Q}{r^2}\left(\frac{r^4}{r^4 + 8l^4}\right). \quad (3.84)$$

The function g is plotted on Fig. 1 [see Eq. (3.13)]. The solution has one horizon if

$$Q = Q_0 = 2.82(lc^2/\sqrt{G}). \quad (3.85)$$

If $Q < Q_0$ there no horizons. If $Q > Q_0$ we have two horizons (as for the Nördström solution to the Einstein–Maxwell equations).In other words, the horizons exist if the mass is sufficiently big [see Eq. (3.79)]. Finally let us calculate the ratio Q/m_N for our solution. One gets using (3.36) and (3.49)

$$Q/m_N = 32\sqrt{2G}/7\pi q. \quad (3.86)$$

However, for an electron,

$$\frac{e}{m_e} = \frac{32\sqrt{2G}}{7\pi q_{\text{electron}}} \quad (3.87)$$

so that

$$\frac{Q}{m_N} = \left(\frac{q_{\text{electron}}}{q}\right)\left(\frac{e}{m_e}\right). \quad (3.88)$$

4. CONCLUSIONS AND PROSPECTS

We have found an exact static, spherically symmetric solution for the nonsymmetric Kaluza–Klein theory.[1,3] Our solution has the following properties: The metric (symmetric part of $g_{\alpha\beta}$) behaves asymptotically like the Reissner–Nördström solution of general relativity [apart from a factor of $(1 + l^4/r^4)$ which is typical in the nonsymmetric gravitational theory[15,16]]. The most remarkable feature of this metric is that the function α is not singular at $r = 0$ and goes to 1 as $r \to 0$. We have calculated the total energy of the solution and its Newtonian mass. Both quantities are constructed from Q and l, the charge and fermion number parameters respectively.[10] The electric field in our solution asymptotically behaves like the Coulomb field generated by a charge Q. However, this field vanishes at $r = 0$ and is nonsingular for all r. We get a maximal value of this field similar to the one in Born–Infeld electrodynamics.[14] We calculated the charge distribution for such a field and showed that it is nonsingular and equal to zero at $r = 0$. Asymptotically our solution behaves similarly to the Reissner–Nördström-like solution in NGT.[16] Although asymptotically we see a Newtonian mass and an electric charge, at the origin ($r = 0$) there is no mass or electric charge (only fermion charge l). Thus it seems that we get "mass" without mass and "charge" without charge. The total charge for our solution is the same as the Coulomb charge (charge seen at infinity). The total mass, on the other hand, is not the same as the Newtonian mass (mass seen at infinity). In this sense we get a kind of finite mass renormalization. If we consider this solution to be a model for a charged particle constructed from gravitational and electromagnetic fields, this mass renormalization is understandable. The Newtonian mass is the mass of the particle and the remainder is the mass of the external electric field. For example, if we consider this solution as a model of an electron we get a connection between the classical radius of an electron and its fermion number parameter l. Note that in general relativity the total energy associated with the electric field of a pointlike electron is infinite.

Our solution possesses a singularity at $r = 0$ in the determinant of the full nonsymmetric metric. However, the (symmetric) metric seems to be less singular. There is no singularity for the function α. The function γ has a singularity only in the factor $(1 + l^4/r^4)$ and the function $\omega = l^2/r^2$ has the usual singularity at $r = 0$. The electric field is not singular. Our solution posseses one or two event horizons if the charge Q (and consequently the Newtonian mass) is sufficiently large. The solution seems to represent a bounded system of gravitational and electromagnetic fields [c.f. the behavior of the function $\tilde{e}$ (see Fig. 3)]. The radial energy density is zero at the origin, and finite everywhere. In a small region around $r = 0$ it is negative. The metric is spatially flat at the origin. For a very small value of the parameter q (see Fig. 6) the function $\alpha \simeq 1$, and $\gamma = (1 + l^4/r^4)$. If the parameter q is equal to q_{electron}, one gets

$$1 \geqslant \alpha^{-1} = (1 - q^2_{\text{electron}} P(R)) \geqslant (1 - q^2_{\text{electron}} P_{\max}) \geqslant 1 - 10^{-74} \simeq 1. \quad (4.1)$$

Thus α is almost exactly one and γ is almost exactly $(1 + l^4/r^4)$. The metric is then as follows:

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & l^2/r^2 \\ 0 & -r^2 & 0 & 0 \\ 0 & 0 & -r^2\sin^2\theta & 0 \\ -l^2/r^2 & 0 & 0 & (1 + l^4/r^4) \end{pmatrix}. \quad (4.2)$$

The symmetric part of this metric is spatially flat. It is easy to see that such behavior is valid for every elementary particle. The remarkable property of (4.2) is that it is described completely by the parameter l (fermion number) which plays the role of the second gravitational charge in the nonsymmetric theory of gravitation. It seems that the fermion number parameter should play a significant role in the unification of elementary particle theory and gravity. In Eq. (4.2) the fermion number parameter is much more important than mass. Thus the geometry of space-time on the level of elementary particles is determined by the second gravitational charge. The function α^{-1} in general relativity has the form

$$\alpha^{-1} = 1 - 2m/r. \quad (4.3)$$

This function describes the difference between the Schwarzschild solution and a Minkowski metric; in particular the curvature of a space. In the solar system at the earth's orbit one finds

$$\alpha^{-1}(1\text{ au}) \cong 1 - 3 \times 10^{-8}, \quad (4.4)$$

where $1\text{ au} = 1.45 \times 10^8$ km, is one astronomical unit (the radius of earth's orbit) and we have put into Eq. (4.3)

$$2m \cong 5\text{ km} \quad (4.5)$$

which is the Schwarzschild radius of the sun. If we compare

Eq. (4.4) with Eq. (4.1) we easily see that our solution with $q = q_{\text{electron}}$ is spatially much more flat *everywhere* than 3-space at the orbit of the earth.

Note that in Eq. (4.2) we get in a natural constant l which has the dimension of length. Some authors claim that it is impossible to get a true unification of the gravitational field and elementary particles without a new universal constant dimensions of length. In the nonsymmetric theory of gravitation there exists such a constant connected to fermion number. The nonsymmetric Kaluza–Klein theory which unifies the nonsymmetric theory of gravitation with a gauge field theory (i.e., the a electromagnetic field), possesses this constant as well.[1-7] This fact might enable these investigations to lead ultimately to a true unification of gravity and elementary particles.

Here are some prospects for further investigation:

1. Find more general spherical solutions with nonzero f and B_0, including nonstatic solutions.

2. Find axially symmetric solutions of the field equations. This is more difficult, because there is no known axially symmetric solution in the Einstein unified field theory and in NGT.

3. Extend our formalism to the nonabelian-nonsymmetric Kaluza–Klein theory (see Refs. 2 and 6), i.e., to find such a solution for the case $G = \mathrm{SU}(2)$ and $G = \mathrm{SU}(2)\times\mathrm{U}(1)$. This will offer a model of an electron or a lepton constructed from gravitational, electromagnetic, and weak interactions.

4. Extend our solution for the nonsymmetric Jordan–Thiry theory (see Ref. 4).

ACKNOWLEDGMENTS

One of us (M.W.K.) would like to thank Professor M. W. Moffat and Dr. R. B. Mann for their kind hospitality and numerous extremely valuable discussions during my stay at the Physics Department of the University of Toronto.

APPENDIX A

Using Eqs. (2.9) and (2.11) from Ref. 17 and the equation

$$\frac{\omega^2}{\alpha\gamma-\omega^2} = \frac{l^4}{\beta^2+f^2}, \tag{A1}$$

one gets

$$\begin{aligned} A_{11}(\bar{\Gamma}) = &-\frac{1}{2}\phi'' - \frac{1}{8}\{(\phi')^2+4C^2\} + \frac{\alpha'}{4\alpha}\phi' \\ &+ \frac{\omega^2}{8\gamma^2}(3(\dot\phi)^2+4D^2) + \left(\frac{\omega^2}{2\alpha\gamma}\phi' + \frac{\gamma'}{2\gamma}\right)\left(\frac{\alpha'}{2\alpha} - \frac{\omega^2}{2\alpha\gamma}\phi' - \frac{\gamma'}{2\gamma}\right) \\ &- \frac{\partial}{\partial r}\left(\frac{\omega^2}{2\alpha\gamma}+\frac{\gamma'}{2\gamma}\right) + \frac{\partial}{\partial t}\left(\frac{\omega^2}{\gamma^2}\dot\phi + \frac{\dot\alpha}{2\gamma}\right) \\ &+ \left(\frac{\omega^2}{\gamma^2}\dot\phi + \frac{\dot\alpha}{2\gamma}\right)\left(\frac{\dot\gamma}{2\gamma} - \frac{\omega^2}{2\alpha\gamma}\dot\phi - \frac{\dot\alpha}{2\alpha} + \frac{1}{2}\dot\phi\right), \end{aligned} \tag{A2}$$

$$\begin{aligned} A_{44}(\bar{\Gamma}) = &-\frac{1}{2}\ddot\phi - \frac{1}{8}((\dot\phi)^2+4D^2) + \frac{\dot\gamma}{4\gamma}\dot\phi \\ &+ \frac{\omega^2}{8\alpha^2}(3(\phi')^2+4C^2) + \left(\frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha}\right)\left(\frac{\gamma'}{2\alpha\gamma}\dot\phi - \frac{\omega^2}{2\alpha\gamma}\dot\phi\frac{\dot\alpha}{2\alpha}\right) \\ &- \frac{\partial}{\partial t}\left(\frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha}\right) + \left(\frac{\alpha'}{2\alpha} - \frac{\omega^2}{2\alpha\gamma}\phi' - \frac{\gamma'}{2\gamma}\right) + \frac{\partial}{\partial r}\left(\frac{\omega^2}{\alpha^2}\phi' + \frac{\gamma'}{2\alpha}\right), \end{aligned} \tag{A3}$$

$$\begin{aligned} A_{22}(\bar{\Gamma}) = &\left[\left(\frac{2fC-\beta\phi'}{4\alpha}\right) + \frac{(2fC-\beta\phi')}{8\alpha}\frac{\partial}{\partial r}\log(\omega^2(\beta^2+f^2)) + \frac{B(f\phi'+2\beta C)}{4\alpha} + 1 - \frac{\partial}{\partial t}\left(\frac{2fD-\beta\dot\phi}{4\gamma}\right)\right. \\ &\left. - \frac{(2fD-\beta\dot\phi)}{8\gamma}\frac{\partial}{\partial t}\log(\omega^2(\beta^2+f^2)) - \frac{D}{4\gamma}(f\dot\phi+2\beta D)\right] \\ = &\frac{1}{\sin^2\theta}A_{33}(\bar{\Gamma}), \end{aligned} \tag{A4}$$

$$\begin{aligned} A_{(14)}(\bar{\Gamma}) = &\frac{\partial}{\partial r}\left(\frac{\omega^2}{4\alpha\gamma}\dot\phi + \frac{\dot\alpha}{4\alpha} - \frac{\dot\gamma}{4\gamma} - \frac{1}{4}\dot\phi\right) + \frac{\partial}{\partial t}\left(\frac{\omega^2}{4\alpha\gamma}\phi' + \frac{\gamma'}{4\gamma} - \frac{\alpha'}{4\alpha} - \frac{1}{4}\phi'\right) \\ &+ \frac{1}{2}\phi'\left(\frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha} - \frac{1}{4}\dot\phi\right) - \left(\frac{\omega^2}{\gamma^2}\dot\phi + \frac{\dot\alpha}{2\gamma}\right)\left(\frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha}\right) \\ &+ \left[\frac{\omega^2}{2\alpha\gamma}\phi' + \frac{\gamma'}{2\gamma}\right]\left(\frac{1}{2}\dot\phi + \frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha}\right) + \frac{\omega^2}{2\alpha\gamma}\phi'\cdot\dot\phi - \frac{DC}{2l^4}\frac{(\beta^2+f^2)}{\alpha\gamma}, \end{aligned} \tag{A5}$$

$$A_{[23]}(\bar{\Gamma}) = \sin\theta\left(\left(\frac{f\phi' - 2\beta C}{4\alpha}\right)' - \frac{C}{4\alpha}(2fC - \beta\phi') + \frac{1}{8\alpha}(f\phi' + 2\beta C)\left(\frac{\alpha'}{\alpha} + \frac{\omega^2}{\alpha\gamma}\phi' + \frac{\gamma'}{\gamma}\right)\right.$$
$$\left. + \frac{1}{8\gamma}(f\dot{\phi} + 2\beta D)\left(\frac{\dot{\gamma}}{\gamma} + \frac{\omega^2}{2\alpha\gamma}\dot{\phi} + \frac{\dot{\alpha}}{2\alpha}\right) - \frac{\partial}{\partial t}\left(\frac{f\dot{\phi} + 2\beta D}{4\gamma}\right) + \frac{D}{4\gamma}(2fD - \beta\dot{\phi})\right), \tag{A6}$$

where

$$\phi = \log(\beta^2 + f^2), \tag{A7}$$

$$C = \frac{f\beta' - \beta f'}{\beta^2 + f^2}, \quad D = \frac{\dot{\beta}f - \beta\dot{f}}{\beta^2 + f^2}, \tag{A8}$$

· means derivative with respect to time t, and ′ means derivative with respect to radius r. (A9)

$$A_{[14]}(\bar{\Gamma}) = \frac{\omega}{8\alpha}((\phi')^2 + 4C^2) - \frac{\omega}{8\gamma}((\dot{\phi})^2 + 4D^2) + \frac{\omega^2}{4\alpha}\phi'(\phi' + \dot{\phi}) - \frac{1}{2}\frac{\partial}{\partial t}\left(\dot{\phi}\frac{\omega}{\gamma}\right) \tag{A10}$$

APPENDIX B

Using condition (3.1) in the static case and the following ideas from Ref. 17 we get from (A2)–(A4) and from Eqs. (3.4) in the static case,

$$-\frac{1}{\alpha}(A_{11}(\bar{\Gamma}) - 8\pi T^{\text{em}}_{11}) + \frac{2}{\beta}(A_{22}(\bar{\Gamma}) - 8\pi T^{\text{em}}_{22}) + \frac{1}{\gamma}(A_{44} - 8\pi T^{\text{em}}_{44})$$
$$= -\frac{1}{\alpha}A_{11}(\bar{\Gamma}) + \frac{2}{\beta}A_{22}(\bar{\Gamma}) + \frac{1}{\gamma}A_{44}(\bar{\Gamma}) - \frac{4}{\alpha}\frac{\beta^2}{(\beta^2 + 4l^2)}8\pi T^{\text{em}}_{11} = P. \tag{B1}$$

One gets

$$0 = \frac{1}{\alpha}(A_{11}(\bar{\Gamma}) - 8\pi T^{\text{em}}_{11}) + \frac{1}{2}P$$
$$= \frac{1}{2\alpha}A_{11}(\bar{\Gamma}) + \frac{1}{\beta}A_{22}(\bar{\Gamma}) + \frac{1}{2\gamma}A_{44}(\bar{\Gamma}) - \frac{8\pi}{\alpha}\left(\frac{\beta^2 + 4l^4}{\beta^2 + 4l^4}\right)T^{\text{em}}_{11}, \tag{B2}$$

$$0 = \frac{1}{\beta}(A_{22}(\bar{\Gamma}) - 8\pi T^{\text{em}}_{22}) + \frac{1}{2}P$$
$$= -\frac{1}{2\alpha}A_{11}(\bar{\Gamma}) + \frac{1}{\beta}A_{22}(\bar{\Gamma}) + \frac{1}{2\gamma}A_{44}(\bar{\Gamma}) - \frac{8\pi}{\alpha}\left(\frac{3\beta^2 - 4l^4}{\beta^2 + 4l^4}\right)T^{\text{em}}_{11}, \tag{B3}$$

$$0 = -\frac{1}{\gamma}(A_{44}(\bar{\Gamma} - 8\pi T^{\text{em}}_{44}) + \frac{1}{2}P$$
$$= -\frac{1}{2\alpha}A_{11}(\bar{\Gamma}) + \frac{1}{\beta}A_{22}(\bar{\Gamma}) - \frac{1}{2\gamma}A_{44}(\bar{\Gamma}) - \frac{8\pi}{\alpha}\left(\frac{\beta^2 - 4l^4}{\beta^2 + 4l^4}\right)T^{\text{em}}_{11}, \tag{B4}$$

where

$$8\pi T^{\text{em}}_{11} = \frac{\alpha Q^2}{\beta^2}\left(\frac{\beta^2 - 4l^4}{(\beta^2 + 8l^4)^2}\right). \tag{B5}$$

From Eqs. (B2)–(B4) one gets

$$(1/\alpha)A_{11}(\bar{\Gamma}) + (1/\gamma)A_{44}(\bar{\Gamma}) = 0. \tag{B6}$$

Let us substitute

$$\alpha = \exp(M), \quad \gamma = \exp(N), \tag{B7}$$

where $M = M(r)$ and $N = N(r)$ are real functions of r. From (B5) one gets

$$\frac{M' + N'}{r} + \frac{4}{r^2}H = 0, \tag{B8}$$

where

$$H = (l^4/(l^4 + \beta^2). \tag{B9}$$

Let us take

$$\beta = r^2 \tag{B10}$$

and substitute Eqs. (B8)–(B10) to Eq. (B4). One gets, using Eqs. (B5) and (B6),

$$\frac{d}{dr}(r\exp(-M)) = 1 - \frac{Q^2}{r^2}\frac{(r^2 + 4l^4)}{(r^4 + 8l^4)^2}. \tag{B11}$$

APPENDIX C

Let us calculate the connection $\bar{\Gamma}^{\alpha}_{\beta\gamma}$ and the Christoffel symbols for our solution. One gets (using results from Ref. 17)

$$\bar{\Gamma}^1_{[14]} = 2l^2/\alpha r^3, \ \bar{\Gamma}^2_{33} = -\tfrac{1}{2}\sin 2\theta, \ \bar{\Gamma}^2_{23} = \bar{\Gamma}^3_{23} = \cot\theta,$$
$$\bar{\Gamma}^1_{22} = (1/\sin^2\theta)\bar{\Gamma}^1_{33} = -r/\alpha,$$
$$\bar{\Gamma}^2_{(12)} = \bar{\Gamma}^3_{(13)} = 1/r,$$
$$\bar{\Gamma}^2_{[24]} = \bar{\Gamma}^3_{[34]} = -l^2/\alpha r^3, \tag{C1}$$
$$\bar{\Gamma}^1_{11} = \alpha'/2\alpha,$$
$$\Gamma^1_{44} = \frac{4l^4}{r^5\alpha^2} + \frac{\gamma'}{2\alpha} = \frac{7l^4}{8\alpha^2 r^5} - \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3},$$
$$\bar{\Gamma}^4_{(14)} = \frac{2l^4}{r^5\alpha\gamma} + \frac{\gamma'}{2\gamma} = \frac{3l^4}{2r^5}\left(1 + \frac{l^4}{r^4}\right)^{-1} - \frac{\alpha'}{2\alpha}.$$

The remaining $\bar{\Gamma}$'s are zero. Let us consider the symmetric part of our soluiton, i.e.,

$$g_{(\mu\nu)} = \begin{pmatrix} -\alpha & 0 & 0 & 0 \\ 0 & -r^2 & 0 & 0 \\ 0 & 0 & -r^2\sin^2\theta & 0 \\ 0 & 0 & 0 & \gamma \end{pmatrix}, \tag{C2}$$

where α and γ are given by the formulas (3.81) and (3.81a). One easily finds the determinant

$$\tilde{g} = \det[g_{(\mu\nu)}] = -(1 + l^4/r^4)r^4 \sin^2\theta. \tag{C3}$$

The determinant is not singular at $r = 0$. The inverse tensor for $g_{(\mu\nu)}$,

$$\tilde{g}^{(\mu\alpha)}g_{(\alpha\nu)} = \delta^{\mu}_{\nu}, \tag{C4}$$

is

$$\tilde{g}^{(\mu\nu)} = \begin{pmatrix} -1/\alpha & 0 & 0 & 0 \\ 0 & -1/r^2 & 0 & 0 \\ 0 & 0 & -1/r^2\sin^2\theta & 0 \\ 0 & 0 & 0 & 1/\gamma \end{pmatrix}. \tag{C5}$$

Let us calculate the Christoffel symbols for $g_{(\mu\nu)}$.

$$\begin{pmatrix} \alpha \\ \beta\gamma \end{pmatrix} = \frac{1}{2}\tilde{g}^{(\alpha\mu)}(g_{(\beta\mu),\gamma} + g_{(\gamma\mu),\beta} - g_{(\beta\gamma),\mu}). \tag{C6}$$

One easily finds

$$\begin{aligned} &\begin{pmatrix} 1 \\ 11 \end{pmatrix} = \frac{\alpha'}{2\alpha}, \\ &\begin{pmatrix} 1 \\ 22 \end{pmatrix} = \frac{r}{\alpha}, \\ &\begin{pmatrix} 1 \\ 33 \end{pmatrix} = \frac{r}{\alpha}\sin^2\theta, \\ &\begin{pmatrix} 2 \\ 33 \end{pmatrix} = -\frac{1}{2}\sin 2\theta, \quad \begin{pmatrix} 3 \\ 32 \end{pmatrix} = \cot\theta, \\ &\begin{pmatrix} 1 \\ 44 \end{pmatrix} = \frac{\gamma'}{2\alpha} = \frac{-l^4}{2\alpha^2 r^5} - \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}, \\ &\begin{pmatrix} 2 \\ 21 \end{pmatrix} = \frac{1}{r} = \begin{pmatrix} 3 \\ 31 \end{pmatrix}, \\ &\begin{pmatrix} 4 \\ 41 \end{pmatrix} = -\frac{\gamma'}{2\gamma} = \frac{\alpha'}{2\alpha} + \frac{l^4}{r^5}\left(1 + \frac{l^4}{r^4}\right)^{-1}. \end{aligned} \tag{C7}$$

The remaining Christoffel symbols are zero. Let us write equations of motion for an uncharged test particle for our solution, i.e., equation for geodesics.

$$\frac{d^2x^{\alpha}}{d\tau^2} + \overline{\Gamma}^{\alpha}_{(\beta\gamma)}\frac{dx^{\beta}}{d\tau}\frac{dx^{\gamma}}{d\tau} = 0. \tag{C8}$$

One easily finds, from (C1),

$$\begin{aligned} &\frac{d^2r}{d\tau^2} + \frac{\alpha'}{2\alpha}\left(\frac{dr}{d\tau}\right)^2 + \left(\frac{7l^4}{8\alpha^2r^5} - \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}\right) \\ &\quad\times\left(\frac{dt}{d\tau}\right)^2 - \frac{r}{\alpha}\left[\left(\frac{d\theta}{d\tau}\right)^2 + \sin^2\theta\left(\frac{d\gamma}{d\tau}\right)^2\right] = 0, \\ &\frac{d^2\theta}{d\tau^2} + \frac{2}{r}\left(\frac{dr}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) - \frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right)^2 = 0, \\ &\frac{d^2\phi}{dt^2} + \frac{2}{r}\left(\frac{dr}{d\tau}\right)\left(\frac{d\phi}{d\tau}\right) + 2\cot\theta\left(\frac{d\phi}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) = 0, \\ &\frac{d^2t}{d\tau^2} + \left(\frac{3l^4}{2r(l^4 + r^4)} - \frac{\alpha'}{2\alpha}\right)\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right) = 0. \end{aligned} \tag{C9}$$

In the nonsymmetric theory of gravitation uncharged particles move along geodesics in Riemanian geometry formed from $g_{(\mu\nu)}$ (see Ref. 13), i.e., in Christoffels' symbols

$$\frac{d^2x^{\alpha}}{d\tau^2} + \begin{pmatrix} \alpha \\ \beta\gamma \end{pmatrix}\frac{dx^{\gamma}}{d\tau}\frac{dx^{\beta}}{d\tau} = 0. \tag{C10}$$

One easily finds, from (C7),

$$\begin{aligned} &\frac{d^2r}{d\tau^2} + \frac{\alpha'}{2\alpha}\left(\frac{dr}{d\tau}\right)^2 - \left(\frac{l^4}{2\alpha^2r^5} + \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}\right)\left(\frac{dt}{d\tau}\right)^2 \\ &\quad + \frac{r}{\alpha}\left(\left(\frac{d\theta}{d\tau}\right)^2 + \sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right) = 0, \\ &\frac{d^2\theta}{d\tau^2} - \frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right)^2 + \frac{2}{r}\left(\frac{d\theta}{d\tau}\right)\left(\frac{dr}{d\tau}\right) = 0, \\ &\frac{d^2\phi}{d\tau^2} + \frac{2}{r}\left(\frac{d\phi}{d\tau}\right)\left(\frac{dr}{d\tau}\right) + 2\cot\theta\left(\frac{d\phi}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) = 0, \\ &\frac{d^2t}{d\tau^2} + \left(\frac{\alpha'}{2\alpha} + \frac{l^4}{r(l^4 + r^4)}\right)\left(\frac{dt}{d\tau}\right)\left(\frac{dr}{d\tau}\right) = 0. \end{aligned} \tag{C11}$$

Let us find equations of motion for a charged test particle. In the nonsymmetric Kaluza–Klein theory one derived such equations, (see Ref. 1)

$$\begin{aligned} &\frac{d^2x^{\alpha}}{d\tau^2} + \overline{\Gamma}^{\alpha}_{(\beta\gamma)}\frac{dx^{\gamma}}{d\tau}\frac{dx^{\beta}}{d\tau} + \left(\frac{q}{m_0}\right) \\ &\quad\times[g^{\alpha\gamma}F_{\gamma\beta} - g^{[\alpha\gamma]}H_{\gamma\beta}]\frac{dx^{\beta}}{d\tau} = 0, \end{aligned} \tag{C12}$$

where q is a charge and m_0 a rest mass of a test particle. Using (C9) and (3.7) one gets

$$\begin{aligned} &\frac{d^2r}{d\tau^2} + \frac{\alpha'}{2\alpha}\left(\frac{dr}{d\tau}\right)^2 + \left(\frac{7l^4}{8\alpha^2r^5} - \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}\right) \\ &\quad\times\left(\frac{dt}{d\tau}\right)^2 - \frac{r}{\alpha}\left(\left(\frac{d\theta}{d\tau}\right)^2 + \sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right) \\ &\quad - \left(\frac{q}{m_0}\right)\frac{Q}{\alpha r^2}\frac{(r^4 + l^4)}{(r^4 + 8l^4)}\left(\frac{dt}{d\tau}\right) = 0, \\ &\frac{d^2\theta}{d\tau^2} + \frac{2}{r}\left(\frac{dr}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) - \frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right) = 0, \\ &\frac{d^2\phi}{d\tau^2} + \frac{2}{r}\left(\frac{dr}{d\tau}\right)\left(\frac{d\phi}{d\tau}\right) + 2\cot\theta\left(\frac{d\phi}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) = 0, \\ &\frac{d^2t}{d\tau^2} + \left(\frac{3l^4}{2r(l^4 + r^4)} - \frac{\alpha'}{2\alpha}\right)\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right) \\ &\quad + \left(\frac{q}{m_0}\right)\left(\frac{r^2\alpha Q}{r^4 + 8l^4}\right)\left(\frac{dr}{d\tau}\right) = 0. \end{aligned} \tag{C13}$$

In Ref. 3 a different possibility is considered for the equations of motion for a charged test particle.

$$\begin{aligned} &\frac{d^2x^{\alpha}}{d\tau^2} + \begin{pmatrix} \alpha \\ \beta\gamma \end{pmatrix}\left(\frac{dx^{\beta}}{d\tau}\right)\left(\frac{dx^{\gamma}}{d\tau}\right) + \left(\frac{q}{m_0}\right) \\ &\quad\times[g^{\alpha\gamma}F_{\gamma\beta} - g^{[\alpha\gamma]}H_{\gamma\beta}]\frac{dx^{\beta}}{d\tau} = 0. \end{aligned} \tag{C14}$$

Using (C9) and (C11) one finds the equations

$$\begin{aligned} &\frac{d^2r}{d\tau^2} + \frac{\alpha'}{2\alpha}\left(\frac{dr}{d\tau}\right)^2 - \left(\frac{l^4}{2\alpha^2r^5} + \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}\right) \\ &\quad\times\left(\frac{dt}{d\tau}\right)^2 + \frac{r}{\alpha}\left[\left(\frac{d\theta}{d\tau}\right)^2 + \sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right] \\ &\quad - \left(\frac{q}{m_0}\right)\frac{Q}{\alpha r^2}\left(\frac{r^4 + l^4}{r^4 8l^4}\right)\frac{dt}{d\tau} = 0, \\ &\frac{d^2\theta}{d\tau^2} - \frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right)^2 + \frac{2}{r}\left(\frac{d\theta}{d\tau}\right)\left(\frac{dr}{d\tau}\right) = 0, \end{aligned}$$

$$\frac{d^2\phi}{d\tau^2}+\frac{2}{r}\left(\frac{d\phi}{d\tau}\right)\left(\frac{dr}{d\tau}\right)+2\cot\theta\left(\frac{d\phi}{d\tau}\right)\left(\frac{d\phi}{d\tau}\right)=0,$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{\alpha'}{2\alpha}+\frac{l^4}{r(l^4+r^4)}\right)\left(\frac{dt}{d\tau}\right)\left(\frac{dr}{d\tau}\right)+\left(\frac{q}{m_0}\right)\left(\frac{r^2\alpha Q}{r^4+8l^4}\right)\left(\frac{dr}{d\tau}\right)=0. \tag{C15}$$

Notice that equations for θ and ϕ are the same in (C9), (C11), (C13), and (C15) regardless of connections and whether the particle is charged or not. For α' we have

$$\alpha'=\frac{\alpha}{r}+\alpha^2\left(\frac{Q^2(r^4+4l^4)}{(r^4+8l^4)^2}-\frac{1}{r}\right), \tag{C16}$$

where α is given by formula (3.81). According to the general properties of the geodetic equations in Einstein's unified theory, nonsymmetric theory of gravitation, and in the nonsymmetric Kaluza–Klein theory, the Eqs. (C9), (C11), (C13), and (C15) have the following first integral (see Refs. 1 and 3):

$$\gamma\left(\frac{dt}{d\tau}\right)^2-\alpha\left(\frac{dr}{d\tau}\right)^2-r^2\times\left[\left(\frac{d\theta}{d\tau}\right)^2+\sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right]=\text{const}. \tag{C17}$$

We can choose const = 1 and

$$\gamma\left(\frac{dt}{d\tau}\right)^2-\alpha\left(\frac{dr}{d\tau}\right)^2-r^2\times\left(\left(\frac{d\theta}{d\tau}\right)^2+\sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right)=1. \tag{C18}$$

Let us consider equations for θ and ϕ,

$$\frac{d^2\theta}{d\tau^2}-\frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right)^2+\frac{2}{r}\left(\frac{d\theta}{d\tau}\right)\left(\frac{dr}{d\tau}\right)=0, \tag{C19}$$

$$\frac{d^2\phi}{d\tau^2}+2\cot\theta\left(\frac{d\theta}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right)+\frac{2}{r}\left(\frac{d\phi}{d\tau}\right)\left(\frac{dr}{d\tau}\right)=0.$$

One easily finds the first integral of motion of (C19),

$$r^2\left(\left(\frac{d\theta}{d\tau}\right)^2+\sin^2\theta\left(\frac{d\theta}{d\tau}\right)^2\right)=\frac{2E_0}{r^2}, \tag{C20}$$

where

$$E_0=\text{const}. \tag{C21}$$

comparing (C18) and (C20) one gets

$$\gamma\left(\frac{dt}{d\tau}\right)^2-\alpha\left(\frac{dr}{d\tau}\right)^2=1-\frac{2E_0}{r^2}. \tag{C22}$$

Let us consider the second equation of (C19). One easily finds the first integral of motion

$$\frac{d\phi}{d\tau}=\frac{L}{r^2\sin^2\theta}, \tag{C23}$$

where L = const. Comparing (C20) and (C23) one gets

$$\left(\frac{d\theta}{d\tau}\right)^2=\frac{1}{r^4}\left(2E_0-\frac{L^2}{\sin^2\theta}\right). \tag{C24}$$

The first integrals (C20) and (C22) lead to the following simplifications of our equations (C9), (C11), (C13), and (C15):

$$\frac{d^2r}{d\tau^2}+\frac{7l^4}{8r(l^4+r^4)}\left(\frac{dr}{d\tau}\right)^2+\left(\frac{7l^4}{8\alpha r(l^4+r^4)}-\frac{\alpha'}{2\alpha^2}\right)\times\left(1-\frac{2E_0}{r^2}\right)-\frac{2E_0}{\alpha r^3}=0, \tag{C9a}$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{3l^4}{2r(l^4+r^4)}-\frac{\alpha'}{2\alpha}\right)\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right)=0,$$

$$\frac{d^2r}{d\tau^2}-\frac{l^4}{2r(l^4+r^4)}\left(\frac{dr}{d\tau}\right)^2-\left(\frac{l^4}{2r\alpha(l^4+\gamma^4)}+\frac{\alpha'}{2\alpha^2}\right)\left(1-\frac{2E_0}{r^2}\right)+\frac{2E_0}{\alpha r^3}=0,$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{\alpha}{2\alpha}+\frac{l^4}{r(l^4+r^4)}\right)\left(\frac{dt}{d\tau}\right)\left(\frac{dr}{d\tau}\right)=0, \tag{C11a}$$

$$\frac{d^2r}{d\tau^2}+\frac{7l^4}{8r(l^4+r^4)}\left(\frac{dr}{d\tau}\right)^2+\left(\frac{7l^4}{8\alpha r(l^4+r^4)}-\frac{\alpha'}{2\alpha}\right)\left(1-\frac{2E_0}{r^2}\right)-\frac{2E_0}{\alpha r^3}-\left(\frac{q}{m_0}\right)\frac{Q}{\alpha r^2}\left(\frac{r^4+l^4}{r^4+8l^4}\right)\left(\frac{dt}{d\tau}\right)=0,$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{3l^4}{2r(l^4+r^4)}-\frac{\alpha'}{2\alpha}\right)\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right)+\left(\frac{q}{m_0}\right)\left(\frac{r^2\alpha Q}{r^4+8l^4}\right)\left(\frac{dr}{d\tau}\right)=0, \tag{C13a}$$

$$\frac{d^2r}{d\tau^2}-\frac{l^4}{2r(l^4+r^4)}\left(\frac{dr}{d\tau}\right)^2-\left(\frac{l^4}{2r\alpha(l^4+r^4)}+\frac{\alpha'}{2\alpha^2}\right)\left(1-\frac{2E_0}{r^2}\right)+\frac{2E_0}{\alpha r^3}-\left(\frac{q}{m_0}\right)\frac{Q}{\alpha r^2}\left(\frac{r^4+l^4}{r^4+8l^4}\right)\left(\frac{dt}{d\tau}\right)=0,$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{\alpha'}{2\alpha}+\frac{l^4}{r(l^4+r^4)}\right)\left(\frac{dt}{d\tau}\right)\left(\frac{dr}{d\tau}\right)+\left(\frac{q}{m_0}\right)\left(\frac{r^2\alpha Q}{r^4+8l^4}\right)\left(\frac{dr}{d\tau}\right)=0. \tag{C15a}$$

For angular coordinates we have for Eqs. (C9a), (C11a), (C13a), and (C15a) the same equations (C19) and the same first integral of motion (C20), (C22), and (C23).

[1] M. W. Kalinowski, "The nonsymmetric Kaluza–Klein theory," J. Math. Phys. **24**, 1835 (1983).

[2] M. W. Kalinowski, "The nonsymmetric-nonabelian Kaluza–Klein theory," J. Phys. A **16**, 1669 (1983).

[3] M. W. Kalinowski, "Material sources in the nonsymmetric Kaluza–Klein theory," University of Toronto report, September 1982 (to appear in J. Math. Phys., 1984).

[4] M. W. Kalinowski, "The nonsymmetric Jordan–Thiry theory," Can. J. Phys. **61**, 884 (1983).

[5] M. W. Kalinowski, "The nonsymmetric–nonabelian Jordan–Thiry theory," University of Toronto report, September 1982.

[6] M. W. Kalinowski, "Spontaneous symmetry breaking and Higgs' mechanism in the nonsymmetric Kaluza–Klein theory," Ann. Phys. **148**, 214 (1983).

[7] M. W. Kalinowski, "Spontaneous symmetry breaking and Higgs' mechanism in the nonsymmetric Jordan–Thiry theory," University of Toronto report, December 1982.

[8] J. W. Moffat, "New theory of Gravitation," Phys. Rev. D **19**, 3557 (1979).

[9] J. W. Moffat, "Gauge invariance and string interactions in a generalized theory of gravitation," Phys. Rev. D **23**, 2870 (1981).

[10] J. W. Moffat, "Generalized theory of gravitation and its physical consequences," in *Proceedings of the VII International School of Gravitation and Cosmology*, Erice Sicily, edited by V. de Sabatta (World Scientific Publishing, Singapore, 1982), p. 127.

[11] M. W. Kalinowski and R. B. Mann, "Linear approximation in the nonsymmetric Kaluza–Klein theory," University of Toronto report, March 1983.

[12]H. A. Hill, R. J. Bos, and P. R. Goode, "Preliminary determination of the quadrupole moment of the sun from rotational splitting of global oscillations and its relevance to tests of general relativity," Phys. Rev. Lett. **33**, 1497 (1983).
[13]J. W. Moffat, "Consequences of a new experimental determination of the quadrupole moment of the sun for gravitational theory," Phys. Rev. Lett. **50**, 709 (1983).
[14]M. Born and L. Infeld, "Foundations of the new field theory," Proc. Roy. Soc. London, Ser. A **144**, 425 (1934).
[15]J. W. Moffat and D. H. Boal, "Solutions in the nonsymmetric unified field theory," Phys. Rev. D **11**, 1375 (1975).
[16]J. W. Moffat, "Static spherically symmetric solution for the field of a charged particle in a theory of gravity," Phys. Rev. D **19**, 3562 (1978).
[17]D. N. Pant, "Spherically Symmetric Rigorous Solutions in Bonnor's Unified Field Theory," Nuovo Cimento B **25**, 175 (1975).
[18]A. Papapetrou, "Static spherically symmetric solutions in the unitary field theory," Proc. Roy. Irish Acad. **52**, 69 (1948).
[19]M. Wyman, "Unified field theory," Can. J. Math. **2**, 427 (1950).
[20]W. B. Bonnor, "The general static spherically symmetric solution in Einstein's unified field theory," Proc. Roy. Soc. **210**, 427 (1952).
[21]W. B. Bonnor, "Static spherically symmetric solutions in Einstein's unified field theory," Proc. Roy. Soc. **209**, 353 (1951).
[22]J. R. Vanstone, "The general static spherically symmetric solution of the 'weak' unified field theory," Can. J. Math. **14**, 568 (1962).
[23]L. Campbell and J. W. Moffat, "Black Holes in the Nonsymmetric Theory of Gravitation," University of Toronto Report, August 1982.

Solution of multidimensional inverse transport problems[a)]

Edward W. Larsen
University of California, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 19 May 1983; accepted for publication 29 July 1983)

Formulas are derived for energy-dependent, steady-state, and time-dependent neutron transport problems, relating the surface neutron fluxes for a convex, homogeneous, three-dimensional region to the neutron scattering laws that apply within the region. In principle, these formulas can be used to deduce information about the scattering laws.

PACS numbers: 05.60. + w, 42.68.Db

I. INTRODUCTION

In recent years, a substantial effort has been directed toward the problem of obtaining exact formulas relating incoming and exiting neutron fluxes for a homogeneous slab to the scattering laws that apply within the slab.[1-13] Such formulas have generally been obtained by directly manipulating the forward and adjoint one-dimensional slab geometry transport equations, although there are exceptions; some early work of Siewert[1,2] makes use of the Chandrasekhar X and Y functions; recent work by Sanchez and McCormick[11] uses the diffusion equation as an approximation to the transport equation; and a recent article by Siewert and Dunn[9] allows for spatial variations in the angular flux in directions parallel to the edges of the slab. Also, most of this prior work considers only monoenergetic transport problems, although Larsen[6] has considered multigroup problems.

In an effort to obtain a more general, and therefore possibly more useful theory, we shall in this article extend the domain of the previous results to the general case of time- and energy-dependent neutron transport in a three-dimensional, convex, homogeneous region. Specifically, for such transport problems we derive exact formulas relating both steady-state and time-dependent surface neutron fluxes to the neutron scattering laws that apply within the region. In principle, these formulas can be used to determine properties of the material scattering laws. However, there are limitations: a large number of neutron flux measurements generally must be made, and the theory described here is only applicable for homogeneous regions.

Our theory thus cannot be used to determine the structure of a heterogeneous solid by irradiating it with external neutrons and measuring (and processing) the incident and exiting fluxes. However, it can be used to solve the following two general problems for a homogeneous region D: (1) If D consists of a uniform mixture of known materials (with known cross sections) in unknown proportions, then determine the proportions; and (2) if the cross sections in D can be regarded as multigroup with a finite number of groups and a finite Legendre expansion in angle, then determine these cross sections.

The remainder of this article is organized as follows. In Sec. II we establish notation and derive physical interpretations for solutions of certain adjoint neutron transport problems. In Sec. III we use these results to derive the inverse theory for steady-state problems; in Sec. IV we repeat this analysis for time-dependent problems. We conclude, in Sec. V, by describing a way to simplify some of the results obtained in Secs. III and IV.

II. PRELIMINARIES

The main purpose of this section is to show that solutions of adjoint transport problems for a convex solid exist having simple interpretations at points on the surface.

To begin, let us assume that steady-state neutron transport occurs within a homogeneous convex region D according to the standard equations

$$\mathbf{\Omega}\cdot\nabla\psi(\mathbf{r},\mathbf{\Omega},E)+\sigma_T(E)\psi(\mathbf{r},\mathbf{\Omega},E)=\int\int\sigma_s(E'\to E,\mathbf{\Omega}'\cdot\mathbf{\Omega})\psi(\mathbf{r},\mathbf{\Omega}',E')d^2\Omega'\,dE', \tag{2.1}$$

$$\psi(\mathbf{r},\mathbf{\Omega},E)=f(\mathbf{r},\mathbf{\Omega},E),\quad \mathbf{r}\in\partial D,\quad \mathbf{\Omega}\cdot\mathbf{n}<0. \tag{2.2}$$

Here $\mathbf{n}$ is the unit outer normal. The solution ψ of problem (2.1), (2.2) is, physically, the neutron angular flux arising from the incident flux f on the surface of D.

To proceed, let R be the set of all phase-space points $(\mathbf{r},\mathbf{\Omega},E)$, with $\mathbf{r}\in\partial D$ and $\mathbf{\Omega}\cdot\mathbf{n}>0$. Let R_0 be any subset of R, and χ_0 the characteristic function for R_0:

$$\chi_0(\mathbf{r},\mathbf{\Omega},E)=\begin{cases}1, & (\mathbf{r},\mathbf{\Omega},E)\in R_0,\\ 0, & (\mathbf{r},\mathbf{\Omega},E)\in R-R_0.\end{cases} \tag{2.3}$$

For any neutron flux $\psi(\mathbf{r},\mathbf{\Omega},E)$ existing in D, we define

$$\int\int\int_R\mathbf{\Omega}\cdot\mathbf{n}\chi_0(\mathbf{r},\mathbf{\Omega},E)\psi(\mathbf{r},\mathbf{\Omega},E)d^2\Omega\,dE\,d^2r=\int\int\int_{R_0}\mathbf{\Omega}\cdot\mathbf{n}\psi(\mathbf{r},\mathbf{\Omega},E)d^2\Omega\,dE\,d^2r$$
$$=\text{the net current out of } R_0. \tag{2.4}$$

Now, let us consider the steady-state adjoint problem

$$-\mathbf{\Omega}\cdot\nabla\psi^*(\mathbf{r},\mathbf{\Omega},E)+\sigma_T(E)\psi^*(\mathbf{r},\mathbf{\Omega},E)=\int\int\sigma_s(E\to E',\mathbf{\Omega}\cdot\mathbf{\Omega}')\psi^*(\mathbf{r},\mathbf{\Omega}',E')d^2\Omega'\,dE', \tag{2.5}$$

$$\psi^*(\mathbf{r},\mathbf{\Omega},E)=\chi_0(\mathbf{r},\mathbf{\Omega},E),\quad r\in\partial D,\quad \mathbf{\Omega}\cdot\mathbf{n}>0. \tag{2.6}$$

We shall prove the following result:

Lemma 1: For any $\mathbf{r}\in\partial D$, $\mathbf{\Omega}\cdot\mathbf{n}<0$, and any E,

[a)] This research was performed under the auspices of the U. S. Department of Energy.

 0022-2488/84/010131-05$02.50

$\psi^*(\mathbf{r}, \mathbf{\Omega}, E)$ = the net current out of R_0 due to a unit delta incident beam at $(\mathbf{r}, \mathbf{\Omega}, E)$.

Proof: Let $\mathbf{r}_0$ be any point on ∂D, $\mathbf{\Omega}_0$ any unit vector such that $\mathbf{\Omega}_0 \cdot \mathbf{n}_0 < 0$, and E_0 any admissible value of E. Also, let ψ be the solution of the forward problem consisting of Eqs. (2.1) and (2.2), with

$$f(\mathbf{r}, \mathbf{\Omega}, E) = \delta(\mathbf{r} - \mathbf{r}_0)\delta(\mathbf{\Omega} - \mathbf{\Omega}_0)\delta(E - E_0)/|\mathbf{\Omega}_0 \cdot \mathbf{n}_0| . \tag{2.7}$$

Then $\psi(\mathbf{r}, \mathbf{\Omega}, E)$ is the angular flux at any point $(\mathbf{r}, \mathbf{\Omega}, E)$ due to the unit delta incident beam f at $(\mathbf{r}_0, \mathbf{\Omega}_0, E_0)$.

We multiply Eq. (2.1) by ψ^* and Eq. (2.5) by ψ, integrate both equations over $\mathbf{\Omega}$ and E, subtract, and then integrate the resulting single equation over all $\mathbf{r} \in D$ to obtain

$$0 = \int_{\partial D} \iint \mathbf{\Omega} \cdot \mathbf{n} \psi^* \psi d^2\Omega \, dE \, d^2r . \tag{2.8}$$

(This is just the reciprocity relation for the special case of no interior sources for the forward and adjoint transport fluxes.[14]) Next, we use Eqs. (2.2), (2.6), and (2.7) to get

$$0 = \iiint_{R_0} \mathbf{\Omega} \cdot \mathbf{n} \psi(\mathbf{r}, \mathbf{\Omega}, E) d^2\Omega \, dE \, d^2r + \int_{\partial D} \iint_{\mathbf{\Omega}\cdot\mathbf{n}<0} \mathbf{\Omega} \cdot \mathbf{n} \psi^* \frac{\delta(\mathbf{r} - \mathbf{r}_0)\delta(\mathbf{\Omega} - \mathbf{\Omega}_0)\delta(E - E_0)}{|\mathbf{\Omega}_0 \cdot \mathbf{n}_0|} \times d^2\Omega \, dE \, d^2r , \tag{2.9}$$

or

$$\psi^*(\mathbf{r}_0, \mathbf{\Omega}_0, E_0) = \iiint_{R_0} \mathbf{\Omega} \cdot \mathbf{n} \psi(\mathbf{r}, \mathbf{\Omega}, E) d^2\Omega \, dE \, d^2r . \tag{2.10}$$

This proves the result. Q.E.D.

Now let us assume that time-dependent neutron transport occurs within the homogeneous convex region D according to the standard equations

$$\frac{1}{v}\frac{\partial}{\partial t}\psi(\mathbf{r}, \mathbf{\Omega}, E, t) + \mathbf{\Omega} \cdot \nabla\psi(\mathbf{r}, \mathbf{\Omega}, E, t) + \sigma_t(E)\psi(\mathbf{r}, \mathbf{\Omega}, E, t) = \iint \sigma_s(E' \to E, \mathbf{\Omega}' \cdot \mathbf{\Omega})\psi(\mathbf{r}, \mathbf{\Omega}', E', t) d^2\Omega' \, dE' , \tag{2.11}$$

$$\psi(\mathbf{r}, \mathbf{\Omega}, E, t) = f(\mathbf{r}, \mathbf{\Omega}, E, t), \quad \mathbf{r} \in \partial D, \ \mathbf{\Omega} \cdot \mathbf{n} < 0, \ 0 < t , \tag{2.12}$$

$$\psi(\mathbf{r}, \mathbf{\Omega}, E, 0) = 0, \quad \mathbf{r} \in D . \tag{2.13}$$

The solution ψ of Eqs. (2.11)–(2.13) is, physically, the time-dependent neutron angular flux arising from the incident flux f on the surface of D. [Throughout this article, we only treat problems with initial data of the form (2.13), i.e., we assume that initially no free neutrons are present in D.]

We let R, R_0, and χ_0 be defined above, and for any neutron flux $\psi(\mathbf{r}, \mathbf{\Omega}, E, t)$ existing in D and $T > 0$, we define

$$\int_0^T \iiint_R \mathbf{\Omega} \cdot \mathbf{n} \chi_0(\mathbf{r}, \mathbf{\Omega}, E)\psi(\mathbf{r}, \mathbf{\Omega}, E, t) d^2\Omega \, dE \, d^2r \, dt = \int_0^T \iiint_{R_0} \mathbf{\Omega} \cdot \mathbf{n} \psi(\mathbf{r}, \mathbf{\Omega}, E, t) d^2\Omega \, dE \, d^2r \, dt = \text{the net current out of } R_0 \text{ up to time } T. \tag{2.14}$$

We now consider the time-dependent adjoint problem

$$-\frac{1}{v}\frac{\partial}{\partial t}\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t) - \mathbf{\Omega} \cdot \nabla\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t) + \sigma_T(E)\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t) = \iint \sigma_s(E \to E', \mathbf{\Omega} \cdot \mathbf{\Omega}')\psi^*(\mathbf{r}, \mathbf{\Omega}', E', t) d^2\Omega' \, dE' , \tag{2.15}$$

$$\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t) = \chi_0(\mathbf{r}, \mathbf{\Omega}, E), \quad \mathbf{r} \in \partial D, \ \mathbf{\Omega} \cdot \mathbf{n} > 0, \ 0 < t < T , \tag{2.16}$$

$$\psi^*(\mathbf{r}, \mathbf{\Omega}, E, T) = 0, \quad \mathbf{r} \in D . \tag{2.17}$$

We shall prove the following result:

Lemma 2: Let $0 < t < T$. Then for any $\mathbf{r} \in \partial D$, $\mathbf{\Omega} \cdot \mathbf{n} < 0$, and any E, $\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t)$ = the net current out of R_0 up to time T due to a unit delta incident beam at $(\mathbf{r}, \mathbf{\Omega}, E, t)$.

Proof: Let $\mathbf{r}_0$ be any point on ∂D, $\mathbf{\Omega}_0$ any unit vector such that $\mathbf{\Omega}_0 \cdot \mathbf{n}_0 < 0$, E_0 any admissible value of E, and $0 < t_0 < T$. Also, let ψ be the solution of the forward problem consisting of Eqs. (2.11)–(2.13), with

$$f(\mathbf{r}, \mathbf{\Omega}, E, t) = \frac{\delta(\mathbf{r} - \mathbf{r}_0)\delta(\mathbf{\Omega} - \mathbf{\Omega}_0)\delta(E - E_0)\delta(t - t_0)}{|\mathbf{\Omega}_0 \cdot \mathbf{n}_0|} . \tag{2.18}$$

Then $\psi(\mathbf{r}, \mathbf{\Omega}, E, t)$ is the time-dependent angular flux at any point $(\mathbf{r}, \mathbf{\Omega}, E, t)$ due to the unit delta incident beam at $(\mathbf{r}_0, \mathbf{\Omega}_0, E_0, t_0)$.

We multiply Eq. (2.11) by ψ^*, Eq. (2.15) by ψ, integrate both equations over $\mathbf{\Omega}$ and E, and subtract to obtain the single equation

$$0 = \frac{\partial}{\partial t}\iint \frac{1}{v}\psi\psi^* d^2\Omega \, dE + \nabla \cdot \iint \mathbf{\Omega}\psi\psi^* \, d^2\Omega \, dE . \tag{2.19}$$

Next, we operate on Eq. (2.19) by

$$\int_0^T \int_D (\cdot) d^3r \, dt , \tag{2.20}$$

and use the initial conditions, Eqs. (2.13), (2.17), and the boundary conditions, Eqs. (2.12), (2.16), and (2.18) to easily obtain

$$\psi^*(\mathbf{r}_0, \mathbf{\Omega}_0, E_0, t_0) = \int_0^T \iiint_{R_0} \mathbf{\Omega} \cdot \mathbf{n} \psi \, d^2\Omega \, dE \, d^2r \, dt . \tag{2.21}$$

This proves the result. Q.E.D.

The main purpose of Lemmas 1 and 2 is to establish the following: (1) there exist solutions ψ^* of the steady-state adjoint transport Eq. (2.5) for which $\psi^*(\mathbf{r}, \mathbf{\Omega}, E)$ is physically measurable for all $\mathbf{r} \in \partial D$, all $\mathbf{\Omega}$, and all E; and (2) there exist solutions ψ^* of the time-dependent adjoint transport Eq. (2.15) and initial condition Eq. (2.17) for which $\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t)$ is physically measurable for all $\mathbf{r} \in \partial D$, all $\mathbf{\Omega}$, all E, and all $t < T$. Such solutions will play a key role in the remainder of this article.

III. STEADY-STATE THEORY

Let ψ be any solution of Eq. (2.1) and ψ^* any solution of Eq. (2.5). We multiply Eq. (2.1) by $\nabla\psi^*$, Eq. (2.5) by $\nabla\psi$, integrate over $\mathbf{\Omega}$ and E, and then add the two resulting equa-

tions, obtaining

$$\iint [(\nabla\psi^*)(\mathbf{\Omega}\cdot\nabla\psi) - (\nabla\psi)(\mathbf{\Omega}\cdot\nabla\psi^*)]d^2\Omega\, dE + \nabla\iint \sigma_T(E)\psi(\mathbf{r},\mathbf{\Omega},E)\psi^*(\mathbf{r},\mathbf{\Omega},E)d^2\Omega\, dE = \nabla\iiiint \sigma_s(E'\to E,\mathbf{\Omega}'\cdot\mathbf{\Omega})\psi(\mathbf{r},\mathbf{\Omega}',E') \times \psi^*(\mathbf{r},\mathbf{\Omega},E)d^2\Omega'\, dE'\, d^2\Omega\, dE\,. \tag{3.1}$$

However, elementary operations give

$$\begin{aligned}(\nabla\psi^*)(\mathbf{\Omega}\cdot\nabla\psi) - (\nabla\psi)(\mathbf{\Omega}\cdot\nabla\psi^*) &= \nabla(\psi^*\mathbf{\Omega}\cdot\nabla\psi) - \mathbf{\Omega}\cdot\nabla(\psi^*\nabla\psi)\\ &= \mathbf{\Omega}\cdot\nabla(\psi\nabla\psi^*) - \nabla(\psi\mathbf{\Omega}\cdot\nabla\psi^*)\,.\end{aligned} \tag{3.2}$$

Introducing Eq. (3.2) into Eq. (3.1) and integrating over $\mathbf{r}$, we obtain

$$\mathbf{S} + \int_{\partial D}\mathbf{n}\iint \sigma_T(E)\psi(\mathbf{r},\mathbf{\Omega},E)\psi^*(\mathbf{r},\mathbf{\Omega},E)d^2\Omega\, dE\, d^2r = \int_{\partial D}\mathbf{n}\iiiint \sigma_s(E'\to E,\mathbf{\Omega}'\cdot\mathbf{\Omega})\psi(\mathbf{r},\mathbf{\Omega}',E') \times \psi^*(\mathbf{r},\mathbf{\Omega},E)d^2\Omega'\, dE'\, d^2\Omega\, dE\, d^2r\,, \tag{3.3}$$

where, using a standard vector identity,[15] we have

$$\mathbf{S} = \int_{\partial D}\iint \psi^*[\mathbf{n}(\mathbf{\Omega}\cdot\nabla\psi) - (\mathbf{\Omega}\cdot\mathbf{n})(\nabla\psi)]d^2\Omega\, dE\, d^2r = \int_{\partial D}\iint \psi^*[\mathbf{\Omega}\times(\mathbf{n}\times\nabla\psi)]d^2\Omega\, dE\, d^2r\,, \tag{3.4a}$$

or

$$\mathbf{S} = \int_{\partial D}\iint \psi[(\mathbf{\Omega}\cdot\mathbf{n})(\nabla\psi^*) - \mathbf{n}(\mathbf{\Omega}\cdot\nabla\psi^*)]d^2\Omega\, dE\, d^2r = \int_{\partial D}\iint \psi[\mathbf{\Omega}\times(\nabla\psi^*\times\mathbf{n})]d^2\Omega\, dE\, d^2r\,, \tag{3.4b}$$

However, if ∇_T denotes the gradient operator in the plane tangent to ∂D, then for any point on ∂D we may use

$$\nabla\psi = \mathbf{n}(\mathbf{n}\cdot\nabla\psi) + \nabla_T\psi \tag{3.5a}$$

in Eq. (3.4a), and

$$\nabla\psi^* = \mathbf{n}(\mathbf{n}\cdot\nabla\psi^*) + \nabla_T\psi^* \tag{3.5b}$$

in (3.4b). Making these substitutions (and noting that $\mathbf{n}\times\mathbf{n} = \mathbf{0}$) we obtain

$$\mathbf{S} = \int_{\partial D}\iint \psi^*[\mathbf{\Omega}\times(\mathbf{n}\times\nabla_T\psi)]d^2\Omega\, dE\, d^2r\,, \tag{3.6a}$$

or

$$\mathbf{S} = \int_{\partial D}\iint \psi[\mathbf{\Omega}\times(\nabla_T\psi^*\times\mathbf{n})]d^2\Omega\, dE\, d^2r\,. \tag{3.6b}$$

Our result is Eq. (3.3) and Eq. (3.6). Each of the terms in these equations consists only of a surface integral involving ψ, ψ^*, $\nabla_T\psi$, or $\nabla_T\psi^*$. Since boundary conditions for ψ and ψ^* have not yet been imposed, we can choose these boundary conditions so that both ψ and ψ^* are physically measurable on ∂D. Doing this, then $\nabla_T\psi$ and $\nabla_T\psi^*$ can also be obtained, and the vector equation (3.3) reduces (for general three-dimensional geometry) to three linear scalar constraints involving σ_T and σ_s. For different combinations of ψ and ψ^*, different constraints are derived, and one can use these constraints to determine properties of σ_s and σ_T, such as described above in Sec. I.

To obtain new constraints on σ_T and σ_s, one does not have to determine new values of both ψ and ψ^*. For instance, one could experimentally determine a specific, unique ψ^*, and then three new constraints are determined by each different value of ψ. Alternatively, one could determine a unique ψ and then derive three different constraints using each different value of ψ^*. [This can easily be done if in evaluating the "first" ψ^* using the theory in Sec. II, one determines the exiting angular fluxes for all points $(\mathbf{r},\mathbf{\Omega},E)\in R_0$. Then, the "first" ψ^* arises from R_0, and arbitrarily many other solutions ψ^* arise from arbitrary subsets of R_0.]

Whichever way one chooses to determine different constraints, it is clear that the experimental determination of the necessary data will require a large number of measurements. In addition, because the problem under consideration is truly inverse in nature, it is likely that our set of constraints will be sensitive to errors in neutron flux measurements. However, only experiment can determine just how accurately the fluxes need to be determined so that errors in measurements of ψ do not lead to unacceptable errors in σ_T or σ_s.

IV. TIME-DEPENDENT THEORY

Let ψ be any solution of Eqs. (2.11) and (2.13), and ψ^* any solution of Eqs. (2.15) and (2.17). We multiply Eq. (2.11) by $\nabla\psi^*$, Eq. (2.15) by $\nabla\psi$, integrate over $\mathbf{\Omega}$ and E, and then add the two resulting equations, obtaining

$$\iint \frac{1}{v}\left[(\nabla\psi^*)\frac{\partial\psi}{\partial t} - (\nabla\psi)\frac{\partial\psi^*}{\partial t}\right]d^2\Omega\, dE + \iint [(\nabla\psi^*)(\mathbf{\Omega}\cdot\nabla\psi) - (\nabla\psi)(\mathbf{\Omega}\cdot\nabla\psi^*)]d^2\Omega\, dE + \nabla\iint \sigma_T\psi\psi^*\, d^2\Omega\, dE = \nabla\iiiint \sigma_s\psi\psi^*\, d^2\Omega'\, dE'\, d^2\Omega\, dE\,. \tag{4.1}$$

Equation (3.2) can be used to rewrite the second term on the left side of Eq. (4.1), while the first term can be rewritten using

$$\begin{aligned}(\nabla\psi^*)\frac{\partial\psi}{\partial t} - (\nabla\psi)\frac{\partial\psi^*}{\partial t} &= \nabla\left(\psi^*\frac{\partial\psi}{\partial t}\right) - \frac{\partial}{\partial t}(\psi^*\nabla\psi)\\ &= \frac{\partial}{\partial t}(\psi\nabla\psi^*) - \nabla\left(\psi\frac{\partial\psi^*}{\partial t}\right).\end{aligned} \tag{4.2}$$

Introducing Eqs. (3.2) and (4.2) into Eq. (4.1), operating by

$$\int_0^T\int_D (\cdot)d^3r\, dt\,,$$

and using the initial conditions (2.13) and (2.17) and the formulas (3.5), we obtain

$$\mathbf{U} + \mathbf{V} + \int_0^T\int_{\partial D}\mathbf{n}\iint \sigma_T(E)\psi(\mathbf{r},\mathbf{\Omega},E,t) \times \psi^*(\mathbf{r},\mathbf{\Omega},E,t)d^2\Omega\, dE\, d^2r\, dt = \int_0^T\int_{\partial D}\mathbf{n}\iiiint \sigma_s(E'\to E,\mathbf{\Omega}'\cdot\mathbf{\Omega})\psi(\mathbf{r},\mathbf{\Omega}',E',t) \times \psi^*(\mathbf{r},\mathbf{\Omega},E,t)d^2\Omega'\, dE'\, d^2\Omega\, dE\, d^2r\, dt\,, \tag{4.3}$$

where

$$\mathbf{U} = \int_0^T \int_{\partial D} \mathbf{n} \iint \frac{1}{v} \psi^* \frac{\partial \psi}{\partial t} \, d^2\Omega \, dE \, d^2r \, dt \tag{4.4a}$$

or

$$\mathbf{U} = -\int_0^T \int_{\partial D} \mathbf{n} \iint \frac{1}{v} \psi \frac{\partial \psi^*}{\partial t} \, d^2\Omega \, dE \, d^2r \, dt \tag{4.4b}$$

and

$$\mathbf{V} = \int_0^T \int_{\partial D} \iint \psi^* [\boldsymbol{\Omega} \times (\mathbf{n} \times \nabla_T \psi)] \, d^2\Omega \, dE \, d^2r \, dt \tag{4.5a}$$

or

$$\mathbf{V} = \int_0^T \int_{\partial D} \iint \psi [\boldsymbol{\Omega} \times (\nabla_T \psi^* \times \mathbf{n})] \, d^2\Omega \, dE \, d^2r \, dt \, . \tag{4.5b}$$

Our result consists of Eqs. (4.3)–(4.5). As with the steady-state analysis, each of the expressions in these equations involving ψ or ψ^* can, in principle, be determined by a suitable interpretation of ψ^* (see Sec. II) together with suitable measurements of surface neutron fluxes. The comments at the end of Sec. III regarding (1) the likely sensitivity of our equations to experimental errors, and (2) the effort that appears necessary to determine acceptable measurements, apply here to an even greater degree than in Sec. III. This is because one must now make accurate measurements for each value of t; therefore, the dimensionality of the space in which measurements must be made, recorded, and processed, is increased by one.

To conclude this section, we note that there is a simple instance in which time-dependent results can be analyzed directly by the steady-state results of Sec. III. This occurs for the case of a subcritical medium and $T = \infty$. Then, assuming that a source of neutrons is beamed onto D for only a finite amount of time, the angular flux ψ will tend to zero as $t \to \infty$. Thus, one can integrate Eq. (2.11) from $t = 0$ to $t = \infty$ and define

$$\psi(\mathbf{r}, \boldsymbol{\Omega}, E) = \int_0^\infty \psi(\mathbf{r}, \boldsymbol{\Omega}, E, t) dt$$

to obtain exactly Eq. (2.1) for the steady-state ψ. The boundary condition is just the time-integrated boundary condition for the time-dependent ψ. Sanchez and McCormick have discussed this (and more general) procedure for slab geometry problems.[10]

V. ADDITIONAL RESULTS

In the previous sections of this article we have considered the problem of forward (and adjoint) transport with boundary conditions that are as general as possible, constrained only by the requirement that ψ and ψ^* are both measurable for all $\mathbf{r} \in \partial D$, all $\boldsymbol{\Omega}$, all E, and all suitable t if the problem is time dependent. In this section, we show that by placing additional constraints on these boundary conditions, a simplification of our results can occur. For brevity and simplicity, we only consider the case of steady-state transport as described in Sec. III.

To be specific, we prove that for certain types of boundary conditions on ψ and ψ^*, the expressions (3.6) for $\mathbf{S}$ simplify to line integrals involving only ψ and ψ^* (not their tangential derivatives) over simple closed curves on ∂D. This makes the resulting constraint (3.3) on σ_T and σ_s substantially simpler and almost certainly less prone to experimental error, because errors in measurements of $\nabla_T \psi$ or $\nabla_T \psi^*$ are likely to be much greater than errors in ψ or ψ^*. We shall not attempt to discuss the most general boundary conditions for which this simplification occurs; we just show that it can occur in special cases.

To describe a special case, let Σ_1 and Σ_2 be simply connected subsets of the boundary ∂D of D with the following properties: (1) the boundaries of Σ_1 and Σ_2 are simple closed curves, Γ_1 and Γ_2, having piecewise continuous tangent vectors; and (2) Σ_1 is sufficiently small in diameter that there exists a unit vector $\hat{\boldsymbol{\Omega}}$ with the property that $\hat{\boldsymbol{\Omega}} \cdot \mathbf{n} < 0$ for all unit outer normal vectors $\mathbf{n}$ corresponding to points in Σ_1. (Thus, $\hat{\boldsymbol{\Omega}}$ points into D at all points in Σ_1. If Σ_1 happens to consist of a planar part of ∂D; then $\hat{\boldsymbol{\Omega}}$ exists and can be any unit vector pointing into D through this plane. In general, $\hat{\boldsymbol{\Omega}}$ exists if Σ_1 is "small" enough that $\mathbf{n}_1 \cdot \mathbf{n}_2 > 0$ for all unit outer normals $\mathbf{n}_1$ and $\mathbf{n}_2$ corresponding to points on Σ_1.) Finally, let $\chi_n(\mathbf{r})$, $n = 1,2$, be the characteristic functions for Σ_1 and Σ_2:

$$\chi_n(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} \in \Sigma_n \, , \\ 0, & \mathbf{r} \in \partial D - \Sigma_n \, . \end{cases} \tag{5.1}$$

We now consider the forward transport problem consisting of Eq. (2.1) and the boundary condition

$$\psi(\mathbf{r}, \boldsymbol{\Omega}, E) = \chi_1(\mathbf{r}) \delta(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}), \quad \mathbf{r} \in \partial D, \ \boldsymbol{\Omega} \cdot \mathbf{n} < 0 \, . \tag{5.2}$$

(This equation describes a uniform, monodirectional beam incident on Σ_1.) Also, we consider the adjoint problem consisting of Eq. (2.5) and

$$\psi^*(\mathbf{r}, \boldsymbol{\Omega}, E) = \chi_2(\mathbf{r}), \quad \mathbf{r} \in \partial D, \ \boldsymbol{\Omega} \cdot \mathbf{n} > 0 \, . \tag{5.3}$$

(The physical interpretation of ψ^* with this boundary condition is given in Sec. II.)

To proceed, we use Eqs. (5.2) and (5.3) in Eq. (3.6b) [use of Eq. (3.6a) leads to the same result] and write

$$\mathbf{S} = \mathbf{S}^+ + \mathbf{S}^- \, , \tag{5.4}$$

where

$$\mathbf{S}^+ = \int_{\partial D} \iint_{\boldsymbol{\Omega} \cdot \mathbf{n} > 0} \psi [\boldsymbol{\Omega} \times (\nabla_T \psi^* \times \mathbf{n})] \, d^2\Omega \, dE \, d^2r \tag{5.5}$$

and

$$\begin{aligned} \mathbf{S}^- &= \int_{\partial D} \iint_{\boldsymbol{\Omega} \cdot \mathbf{n} < 0} \psi [\boldsymbol{\Omega} \times (\nabla_T \psi^* \times \mathbf{n})] \, d^2\Omega \, dE \, d^2r \\ &= \iint_{\Sigma_1} \hat{\boldsymbol{\Omega}} \times [\nabla_T \psi^*(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) \times \mathbf{n}] \, dE \, d^2r \, . \end{aligned} \tag{5.6}$$

If we define

$$d(\mathbf{r}, \Gamma_2) = \text{the distance from } \mathbf{r} \text{ to } \Gamma_2 \, , \tag{5.7}$$

then by Eq. (5.3), for $\boldsymbol{\Omega} \cdot \mathbf{n} > 0$,

$$\nabla_T \psi^* = -\delta[d(\mathbf{r}, \Gamma_2)] \mathbf{m} \, , \tag{5.8}$$

where δ is the usual delta function and $\mathbf{m}$ is the unit outer normal to Γ_2 in the plane of ∂D. Introducing Eq. (5.8) into

Eq. (5.5), we obtain

$$\mathbf{S}^{+} = -\int_{\Gamma_2}\iint_{\mathbf{\Omega}\cdot\mathbf{n}>0} \psi[\mathbf{\Omega}\times(\mathbf{m}\times\mathbf{n})]d^{2}\Omega\, dE\, d^{1}r\,. \quad (5.9)$$

Finally, we note that

$$-\mathbf{m}\times\mathbf{n} = \mathbf{t}\,, \quad (5.10)$$

where $\mathbf{t}$ is the unit tangent vector pointing in the direction of the transverse of Γ_2. (This direction is right handed with respect to the outer normals of Σ_2.) Equation (5.9) thus reduces to

$$\mathbf{S}^{+} = \int_{\Gamma_2}\iint_{\mathbf{\Omega}\cdot\mathbf{n}>0} (\mathbf{\Omega}\times\mathbf{t})\psi\, d^{2}\Omega\, dE\, d^{1}r\,, \quad (5.11)$$

which is the desired simplification of Eq. (5.5).

To simplify Eq. (5.6), it is necessary to use vector indicial notation and Stokes' theorem.[15] Then, with

$$\hat{\psi}^{*}(\mathbf{r}, E) \equiv \psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)\,, \quad (5.12)$$

we have

$$\hat{\mathbf{\Omega}}\times[\nabla_T\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)\times\mathbf{n}] = \hat{\mathbf{\Omega}}\times(\nabla\hat{\psi}^{*}\times\mathbf{n}) = \epsilon_{ijk}\hat{\Omega}_j\epsilon_{klm}\hat{\psi}^{*}_{,l}n_m$$
$$= -\epsilon_{mlk}[\epsilon_{ijk}\hat{\Omega}_j\hat{\psi}^{*}]_{,l}n_m\,. \quad (5.13)$$

Thus, by Stokes' theorem,

$$\int_{\Sigma_1}\hat{\Omega}\times[\nabla_T\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)\times\mathbf{n}]d^{2}r$$
$$= -\int_{\Sigma_1}\epsilon_{mlk}[\epsilon_{ijk}\hat{\Omega}_j\hat{\psi}^{*}]_{,l}n_m\, d^{2}r$$
$$= -\int_{\Gamma_1}[\epsilon_{ijk}\hat{\Omega}_j\hat{\psi}^{*}]t_k\, d^{1}r$$
$$= -\int_{\Gamma_1}(\hat{\mathbf{\Omega}}\times\mathbf{t})\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)d^{1}r\,. \quad (5.14)$$

Using this result in Eq. (5.6), we obtain

$$\mathbf{S}^{-} = -\int_{\Gamma_1}\int(\hat{\mathbf{\Omega}}\times\mathbf{t})\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)dE\, d^{1}r\,, \quad (5.15)$$

which is the desired simplification. Combining Eqs. (5.4), (5.11), and (5.15), we obtain the final result

$$\mathbf{S} = \int_{\Gamma_2}\iint_{\mathbf{\Omega}\cdot\mathbf{n}>0}(\mathbf{\Omega}\times\mathbf{t})\psi(\mathbf{r}, \mathbf{\Omega}, E)d^{2}\Omega\, dE\, d^{1}r$$
$$-\int_{\Gamma_1}\int(\hat{\mathbf{\Omega}}\times\mathbf{t})\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)dE\, d^{1}r\,, \quad (5.16)$$

which consists of line integrals of just ψ and ψ^{*}.

Other boundary conditions for ψ and ψ^{*} also lead to expressions of the form (5.16) for $\mathbf{S}$. For example, one could replace the delta function in $\mathbf{\Omega}$ in Eq. (5.2) by a characteristic function in $\mathbf{\Omega}$ over a subset of the cone of directions pointing into D through all of Σ_1. ($\hat{\mathbf{\Omega}}$ belongs to this cone.) However, we shall not consider this topic further here.

ACKNOWLEDGMENTS

I would like to thank Norman McCormick and Richard Sanchez for their interest, encouragement, and helpful suggestions.

[1] C. E. Siewert, "On a possible experiment to establish the validity of the one-speed or constant cross-section model of the neutron transport equation," J. Math. Phys. **19**, 1587 (1978).
[2] C. E. Siewert, "On Establishing a Two-Term Scattering Law in the Theory of Radiative Transfer," Z. Angew. Math. Phys. **30**, 522 (1979).
[3] N. J. McCormick, "Transport scattering coefficients from reflection and transmission measurements," J. Math. Phys. **20**, 1504 (1979).
[4] C. E. Siewert, "On the Inverse Problem for a Three-Term Phase Function," J. Quant. Spectrosc. Radiat. Transfer **22**, 441 (1979).
[5] C. E. Siewert and J. R. Maiorino, "The Inverse Problem for a Finite Rayleigh-Scattering Atmosphere," Z. Angew. Math. Phys. **31**, 767 (1980).
[6] E. W. Larsen, "Solution of the inverse problem in multigroup transport theory," J. Math. Phys. **22**, 158 (1981).
[7] N. J. McCormick and R. Sanchez, "Inverse problem transport calculations for anisotropic scattering coefficients," J. Math. Phys. **22**, 199 (1981).
[8] R. Sanchez and N. J. McCormick, "General solutions to inverse transport problems," J. Math. Phys. **22**, 847 (1981).
[9] C. E. Siewert and W. L. Dunn, "On inverse problems for plane-parallel media with nonuniform source illumination," J. Math. Phys. **23**, 1376 (1982).
[10] R. Sanchez and N. J. McCormick, "Numerical Evaluation of Optical Single-Scattering Properties Using Multiple-Scattering Inverse Transport Methods," J. Quant. Spectrosc. Radiat. Transfer **28**, 169 (1982).
[11] R. Sanchez and N. J. McCormick, "Inverse Problem Calculations for Multigroup Diffusion Theory," Nucl. Sci. Eng. **83**, 63 (1983).
[12] C. E. Siewert, "Solutions to an Inverse Problem in Radiative Transfer with Polarization-I," J. Quant. Spectrosc. Radiat. Transfer (in press).
[13] N. J. McCormick and R. Sanchez, "Solutions to an Inverse Problem in Radiative Transfer with Polarization-II," J. Quant. Spectrosc. Radiat. Transfer (in press).
[14] G. I. Bell and S. Glasstone, *Nuclear Reactor Theory* (Van Nostrand Reinhold, New York, 1970), p. 258.
[15] R. Aris, *Vector, Tensors, and the Basic Equations of Fluid Mechanics* (Prentice-Hall, Englewood Cliffs, NJ, 1965).

Symmetric Hadamard series

M. R. Brown
Department of Astrophysics, South Parks Road, Oxford, OX1 3RQ, United Kingdom

(Received 20 July 1982; accepted for publication 23 December 1982)

In a general curved space–time, the requirements that the Feynman Green's function be symmetric and have the Hadamard form are shown to result in specific constraints on the local behavior of the function. These constraints are solved yielding a general form for the function.

PACS numbers: 11.10.Cd, 02.30.Bi

I. INTRODUCTION

The Feynman Green's function, or time-ordered, two-point function, is a quantity of central importance in the study of quantum field theory in curved, or flat, space–time. In Minkowski space–time there is, for a given field, exactly one such function. When space–time is curved, there are often many candidates for the title. In this paper I wish to discuss the structure of these functions that is required by the two constraints: that they have the Hadamard[1] form and that they be symmetric functions of the two space–time points involved in their definition. I shall not discuss whether they ought to have the Hadamard form, although there is fast growing support for this idea,[2] nor shall I discuss boundary conditions or Cauchy problems. They must be symmetric functions, and it is how this condition affects the Hadamard form that I shall investigate. I shall use the example of the massless, conformally invariant, scalar field in an arbitrary curved space–time. The analysis will be seen to be applicable to more general fields.

Although in writing this paper I have in mind the application to quantum field theory, it is exclusively concerned with properties of the classical wave equation; Planck's constant enters only in spirit. This is an important point: Much of the subsequent analysis is about finding a missing length. In quantum field theory this length might find expression as an arbitrary renormalization length or the Planck length. Here, with a massless, classical field theory, it is a length that can only be constructed from the curvature of space–time itself.

II. THE SYMMETRIC HADAMARD SERIES

In this section I shall derive a necessary condition for the Green's function $G(x,x')$ to be a symmetric solution to the inhomogeneous wave equation,

$$(\Box - \tfrac{1}{6}R)G(x,x') = -g^{-1/2}(x)\delta^4(x-x') \tag{2.1}$$

having the Hadamard form,

$$G(x,x') = i(8\pi^2)^{-1}[\Delta^{1/2}(\sigma + i\epsilon)^{-1} + v\ln(\sigma + i\epsilon) + w]. \tag{2.2}$$

First, note some well-known features of Eq. (2.2): The factors $i\epsilon$ are included to give G the singularity structure that is appropriate for a Feynman Green's function. $2\sigma(x,x')$ is the square of the length along the geodesic joining x and x'. (One can require that x and x' belong to a "simple region"[3]; this ensures that it is meaningful to speak of their being joined by a unique geodesic.) $\Delta(x,x')$ is the symmetric biscalar constructed from the Van Vleck–Morette determinant, viz.,

$$\Delta(x,x') \equiv -g^{-1/2}(x)\,g^{-1/2}(x')\det(-\sigma_{;ab'}). \tag{2.3}$$

Δ satisfies the equation

$$\sigma^{;a}(\ln\Delta)_{;a} = 4 - \Box\sigma. \tag{2.4}$$

The functions $v(x,x')$ and $w(x,x')$ can be represented as the uniformly convergent power series,[1]

$$v(x,x') = \sum_{n=0}^{\infty} v_n(x,x')\sigma^n(x,x'), \tag{2.5}$$

$$w(x,x') = \sum_{n=0}^{\infty} w_n(x,x')\sigma^n(x,x'), \tag{2.6}$$

where the coefficients v_n and w_n satisfy the differential recursion relations

$$(n+1)(n+2)v_{n+1} + (n+1)v_{n+1;c}\sigma^{;c} - (n+1)v_{n+1}\Delta^{-1/2}\Delta^{1/2}_{;c}\sigma^{;c} + \tfrac{1}{2}(\Box - \tfrac{1}{6}R)v_n = 0, \tag{2.7}$$

and

$$\begin{aligned}(n+1)(n+2)w_{n+1} &+ (n+1)w_{n+1;c}\sigma^{;c} \\ &- (n+1)w_{n+1}\Delta^{-1/2}\Delta^{1/2}_{;c}\sigma^{;c} \\ &+ \tfrac{1}{2}(\Box - \tfrac{1}{6}R)w_n + (2n+3)v_{n+1} + v_{n+1;c}\sigma^{;c} \\ &- v_{n+1}\Delta^{-1/2}\Delta^{1/2}_{;c}\sigma^{;c} = 0;\end{aligned} \tag{2.8}$$

the biscalar $v(x,x')$ is completely determined by Eq. (2.7), and the boundary condition

$$v_0 + v_{0;c}\sigma^{;c} - v_0\Delta^{-1/2}\Delta^{1/2}_{;c}\sigma^{;c} + \tfrac{1}{2}(\Box - \tfrac{1}{6}R)\Delta^{1/2} = 0. \tag{2.9}$$

$v(x,x')$ is a solution to the homogeneous wave equation. The functions $v(x,x')$ and $v_n(x,x')$ are known to be symmetric.[2] v and v_1 have the covariant Taylor series expansions

$$v(x,x') = \tfrac{1}{2}v_{ab}(x)\sigma^{;a}\sigma^{;b} - \tfrac{1}{4}v_{ab;c}(x)\sigma^{;a}\sigma^{;b}\sigma^{;c} + O(\sigma^2), \tag{2.10}$$

and

$$v_1(x,x') = v_1(x) - \tfrac{1}{2}v_{1;a}(x)\sigma^{;a} + O(\sigma), \tag{2.11}$$

where

$$\begin{aligned}v^{ab} &= \tfrac{1}{120}(C^{c(ab)d}R_{cd} + 2C^{c(ab)d}{}_{;cd}), \\ &= \tfrac{1}{240}g^{-1/2}\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2}C_{abcd}C^{abcd},\end{aligned} \tag{2.12}$$

and

$$v_1(x) = \tfrac{1}{720}(R_{abcd}R^{abcd} - R_{ab}R^{ab} + \Box R). \tag{2.13}$$

Equations (2.12) and (2.13) are immediate consequences of the formulae given in the Appendix.

Less is known about the biscalar $w(x,x')$. Clearly it must be symmetric if G is to be symmetric. w (and hence G) is completely determined by the recursion relations once the biscalar $w_0(x,x')$ is specified. Thus the requirement that $w(x,x')$ be symmetric can be seen as a condition on $w_0(x,x')$. w, unlike v, is not a solution to the homogeneous wave equation; it is a simple matter to show that it has to satisfy the equation

$$(\Box - \tfrac{1}{6}R)w(x,x') = -6v_1(x) + 2v_{1;a}(x)\sigma^{;a} + O(\sigma). \tag{2.14}$$

w is unlike v in another important respect: The biscalar $v(x,x')$ has a covariant Taylor series expansion, the first few terms of which are given by Eq. (2.10). The complete expansion has the property that the coefficients $v_{ab}(x)$, etc., are polynomial functions of the curvature tensor and its covariant derivatives. One might ask if one should expect the same property to hold for the covariant Taylor seris expansion of w, when, as in often the case, one seeks to find a purely geometrical solution to equation (2.2). The answer is that, in general, one should not: In Eq. (2.2) the argument of the logarithm is a dimensional quantity. Thus w must supply a term $-v(x,x')\ln L(x,x')$, where $L(x,x')$ is a function having the dimensions of area. The requirement that G be geometrical implies that L must be some function of the curvature tensor. I shall return to this point in the next section where I shall be able to specify further $L(x,x')$.

Let me now determine a condition on $w_0(x,x')$ that must be satisfied if $G(x,x')$ is to be symmetric. I begin with some observations on covariant Taylor series: Let A be a biscalar possessing a covariant Taylor series expansion in a neighborhood of the point x, namely,

$$A(x,x') = A(x) + A_a(x)\sigma^{;a} + \tfrac{1}{2}A_{ab}(x)\sigma^{;a}\sigma^{;b} + \tfrac{1}{6}A_{abc}(x)\sigma^{;a}\sigma^{;b}\sigma^{;c} + O(\sigma^2), \tag{2.15}$$

where $A_{ab} = A_{(ab)}$ and $A_{abc} = A_{(abc)}$, etc. The expansion coefficients, A_{ab} etc., can be expressed as coincidence limits of covariant derivatives of the biscalar $A(x,x')$ by means of the equations[4]

$$\begin{aligned} A(x) &= [A], \\ A_a(x) &= [A_{;a}] - [A]_{;a}, \\ A_{ab}(x) &= [A_{;(ab)}] - 2[A_{;(a}]_{;b)} + [A]_{;(ab)}, \\ A_{abc}(x) &= [A_{;(abc)}] - 3[A_{;(ab}]_{;c)} + 3[A_{;(a}]_{;bc)} - [A]_{;(abc)}, \end{aligned} \tag{2.16}$$

where I use the standard notation

$$[A] \equiv \lim_{x' \to x} A(x,x').$$

Using these equations, it is easy to compute the Taylor series for the function $A(x',x)$. The requirement that $A(x,x')$ equal $A(x',x)$ results in the conditions

$$2A_a(x) = -A_{;a}(x), \tag{2.17}$$

$$4A_{abc}(x) = -6A_{(ab;c)}(x) + A_{;(abc)}(x), \tag{2.18}$$

and so on. More generally, the requirement of symmetry determines the odd coefficients, A_a, A_{abc}, A_{abcde}, etc. However, I shall need only Eqs. (2.17) and (2.18) in what follows and shall not record the higher order constraints.

$w(x,x')$ is a symmetric biscalar that, it is supposed, possesses a Taylor series expansion. Therefore, by the above argument, it can be written

$$w(x,x') = w(x) - \tfrac{1}{2}w_{;a}(x)\sigma^{;a} + \tfrac{1}{2}w_{ab}(x)\sigma^{;a}\sigma^{;b} - \tfrac{1}{4}\{w_{ab;c}(x) - \tfrac{1}{6}w_{;abc}(x)\}\sigma^{;a}\sigma^{;b}\sigma^{;c} + O(\sigma^2), \tag{2.19}$$

where $w(x) = [w]$ and $w_{ab} = [w_{;ab}]$.

At this point there are several ways to proceed. Perhaps the most direct is to require that $w(x,x')$, as given by Eq. (2.19), satisfy Eq. (2.14). So doing, one obtains the equations

$$w^a{}_a(x) = \tfrac{1}{6}Rw(x) - 6v_1(x), \tag{2.20}$$

and

$$\{w^a{}_b(x) - \tfrac{1}{2}\delta^a{}_b w^c{}_c(x)\}_{;a} = 2v_{1;b}(x) + \tfrac{1}{4}(\Box w(x))_{;b} + \tfrac{1}{2}R^a{}_b w_{;a}(x) - \tfrac{1}{12}Rw_{;b}(x). \tag{2.21}$$

Next one has to relate these equations to $w_0(x,x')$. This is done as follows: $w_0(x,x')$ has a Taylor series expansion

$$w_0(x,x') = w_0(x) - \tfrac{1}{2}w_{0;a}(x)\sigma^{;a} + \tfrac{1}{2}w_{0ab}(x)\sigma^{;a}\sigma^{;b} + O(\sigma^{3/2}), \tag{2.22}$$

where $w_0(x) = w(x)$. [The form of the second term in Eq. (2.22) is required by the symmetry of $w(x,x')$; it must not be supposed that $w_0(x,x')$ has any particular symmetry property.] $w_1(x,x')$ has a Taylor series

$$w_1(x,x')\sigma = \tfrac{1}{2}w_{1ab}(x)\sigma^{;a}\sigma^{;b} + O(\sigma^{3/2}), \tag{2.23}$$

where, by Eq. (2.8) and (2.6),

$$w_{1ab}(x) = g_{ab}[w_1(x,x')] \tag{2.24}$$

and

$$[w_1(x,x')] = \tfrac{1}{24}Rw_0(x) - \tfrac{1}{4}w_0{}^a{}_a(x) - \tfrac{3}{2}v_1(x). \tag{2.25}$$

Combining Eqs. (2.22) and (2.23) with (2.6), one sees that

$$w(x,x') = w_0(x) - \tfrac{1}{2}w_{0;a}(x)\sigma^{;a} + \tfrac{1}{2}\{w_{0ab}(x) + w_{1ab}(x)\}\sigma^{;a}\sigma^{;b} + O(\sigma^{3/2}). \tag{2.26}$$

Comparing this equation with Eq. (2.19) yields the result

$$w_{ab}(x) = w_{0ab}(x) + w_{1ab}(x). \tag{2.27}$$

Now Eqs. (2.20) and (2.21) can be written in terms of $w_0(x,x')$. The first of these equations is identically satisfied; in other words, it is not a constraint on $w_0(x,x')$. The second is more interesting and becomes

$$\{w_0{}^a{}_b(x) - \tfrac{1}{4}\delta^a{}_b w_0{}^c{}_c(x)\}_{;a} = \tfrac{1}{2}v_{1;b}(x) + \tfrac{1}{4}(\Box w_0(x))_{;b} + \tfrac{1}{2}R^a{}_b w_{0;a}(x) + \tfrac{1}{24}\{R_{;b}w_0(x) - Rw_{0;b}(x)\}. \tag{2.28}$$

Equation (2.28) must be satisfied by the coefficients in the Taylor series expansion of $w_0(x,x')$ if G is to be a symmetric Hadamard solution to Eq. (2.1). Of course, there will be additional constraints on the higher order Taylor series coefficients. These would require some dedication to compute; fortunately, one needs only those terms up to $w_{0ab}(x)$ to understand quantum field theoretic energy densities.[5] In this context, notice that $w_0(x,x') = 0^6$ is not a solution to Eq. (2.28) unless $v_1(x)$ is constant. $v_1(x)$ [Eq. (2.13)] is a function that is commonly known[7] as the "trace anomaly."

In the next section I shall describe the geometrical solutions to Eq. (2.28).

III. THE FORM OF $w(x,x')$

I shall regard Eq. (2.28) as a constraint on $w_{0ab}(x)$ for some given w_0 in a general curved space–time. It can be solved as follows:

Let me write

$$w_{0ab} = s_{ab} + t_{ab}, \tag{3.1}$$

where s_{ab} satisfies

$$(s^a{}_b - \tfrac{1}{4}\delta^a{}_b s^c{}_c)_{;a} = \tfrac{1}{4}(\Box w_0)_{;b} + \tfrac{1}{2}R^a{}_b w_{0;a} + \tfrac{1}{24}(R_{;b} w_0 - R w_{0;b}), \tag{3.2}$$

and t_{ab} satisfies

$$(t^a{}_b - \tfrac{1}{4}\delta^a{}_b t^c{}_c)_{;a} = \tfrac{1}{2}v_{1;b}. \tag{3.3}$$

A solution to Eq. (3.2) is provided by

$$s_{ab} = \tfrac{1}{6}(w_0 R_{ab} - \tfrac{1}{4}g_{ab} w_0 R) + \tfrac{1}{3}(w_{0;ab} - \tfrac{1}{4}g_{ab}\Box w_0). \tag{3.4}$$

This is geometrical, provided, of course, that w_0 is a function of the curvature. That it satisfies Eq. (3.2) is easily checked: One uses the Bianchi identity

$$R^a{}_{b;a} = \tfrac{1}{2}R_{;b}, \tag{3.5}$$

and the differential identity

$$\Box(w_{0;b}) = (\Box w_0)_{;b} + R^a{}_b w_{0;a}. \tag{3.6}$$

Finding a solution to Eq. (3.3) is not so easy. I first gave a solution to this equation some years ago.[8] However, the method I then used is inappropriate in the present context. I think that the following is a more interesting way to proceed.

I define the tensor T:

$$T_{ab} \equiv t_{ab} - \tfrac{1}{4}g_{ab} t^c{}_c - \tfrac{1}{2}v_1 g_{ab}. \tag{3.7}$$

Then Eq. (3.3) implies that

$$T^{ab}{}_{;a} = 0 \tag{3.8}$$

and

$$T^a{}_a = -2v_1. \tag{3.9}$$

Thus one has a geometrical solution to Eq. (3.3) if one can find a geometrical tensor T^{ab} that is conserved [Eq. (3.8)] and whose trace is proportional to the trace anomaly [Eq. (3.9)]. The clue to finding such a tensor is provided by the conservation equation: suppose that T^{ab} is the variation with respect to the metric of an invariant action. In other words, let

$$T^{ab} = 2g^{-1/2}\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2} A(g_{cd}). \tag{3.10}$$

Equation (3.8) is the statement that A be a scalar under general coordinate transformations. Equation (3.9) can be treated as a statement about the scaling behavior of A; more precisely,

$$\frac{\delta}{\delta\omega}\int d^4x\, \hat{g}^{1/2} A(\hat{g}_{cd})\Big|_{\omega=0} = -g^{1/2} T^a{}_a, \tag{3.11}$$

where $\hat{g}$ is related g by the equation

$$\hat{g}_{ab} \equiv e^{-2\omega} g_{ab}.$$

Equation (3.11) suggests that a suitable action can be found by integrating the functional differential equation

$$\frac{\delta}{\delta\omega}\int d^4x\, \hat{g}^{1/2} A(\hat{g}_{cd}) = -\hat{g}^{1/2} T^a{}_a(\hat{g}_{cd}). \tag{3.12}$$

Equation (3.12) clearly reduces to (3.11) in the limit $\omega = 0$. The variation with respect to ω is taken holding the metric g_{ab} fixed. In this sense, Eq. (3.12) is a partial, functional differential equation. Bearing this in mind, it is remarkably simple to integrate it.

Using the formulae in the Appendix, $\hat{g}^{1/2} T^a{}_a(\hat{g}_{cd})$ can be written

$$\begin{aligned}\hat{g}^{1/2} T^a{}_a(\hat{g}_{cd}) &= -2\hat{g}^{1/2} v_1(\hat{g}_{cd}) \\ &= -\tfrac{1}{360}\hat{g}^{1/2}(\hat{R}_{abcd}\hat{R}^{abcd} - \hat{R}_{ab}\hat{R}^{ab} + \hat{\Box}\hat{R}) \\ &= -\tfrac{1}{360}g^{1/2}\{R_{abcd}R^{abcd} - R_{ab}R^{ab} + \Box R \\ &\quad + 2R\,\Box\omega + 2R_{;a}\omega^{;a} + 6\Box(\Box\omega) + 8[(\Box\omega)^2 \\ &\quad - \omega_{;ab}\omega_{;ab} - R_{ab}\omega^{;a}\omega^{;b} \\ &\quad - \omega^{;c}\omega_{;c}\Box\omega - 2\omega_{;ab}\omega^{;a}\omega^{;b}]\}.\end{aligned} \tag{3.13}$$

It is straightforward to see that Eq. (3.12) can be functionally integrated to give

$$\hat{g}^{1/2} A(\hat{g}_{cd}) = g^{1/2} C(\omega; g_{cd}) + g^{1/2} F(g_{cd}), \tag{3.14}$$

where F is a function of the metric (but not ω) and

$$\begin{aligned}C(\omega; g_{cd}) \equiv \tfrac{1}{360}[&(R_{abcd}R^{abcd} - R_{ab}R^{ab} + \Box R)\omega + 3(\Box\omega)^2 \\ &- 2R_{ab}\omega^{;a}\omega^{;b} - 4\omega^{;c}\omega_{;c}\Box\omega + 2(\omega^{;c}\omega_{;c})^2].\end{aligned} \tag{3.15}$$

C is determined uniquely up to total divergences. Equation (3.14) must hold for $\omega = 0$. This implies that

$$F(g_{cd}) = A(g_{cd}). \tag{3.16}$$

Equation (3.14) may now be seen to determine the scaling behavior of the function A:

$$\hat{g}^{1/2} A(e^{-2\omega} g_{cd}) - g^{1/2} A(g_{cd}) = g^{1/2} C(\omega; g_{cd}). \tag{3.17}$$

Thus the problem of finding a tensor satisfying equations (3.8) and (3.9) has been reduced to finding a scalar A that satisfies the scaling equation (3.17).

The solutions to Eq. (3.17) can be found by choosing ω to be a function of the curvature that has the scaling law

$$\omega(e^{-2\chi} g_{ab}) = \omega(g_{ab}) - \chi. \tag{3.18}$$

Equation (3.17) then has the solution $A^*(g_{cd})$, where

$$A^*(g_{ab}) = -C(\omega(g_{ab}); g_{ab}). \tag{3.19}$$

This is clearly a solution since

$$A^*(e^{-2\chi} g_{ab}) = -C(\omega - \chi; e^{-2\chi} g_{ab}). \tag{3.20}$$

Setting $\chi = \omega$ in Eq. (3.20) yields

$$A^*(e^{-2\omega} g_{ab}) = 0. \tag{3.21}$$

More general solutions to Eq. (3.17) are obtained by adding to a solution $g^{1/2}A^*$ any conformal invariant. It is worth noting that, when ω satisfies Eq. (3.18), C has the scaling property

$$C(\omega(e^{-2\chi} g_{ab}); e^{-2\chi} g_{ab}) = C(\omega(g_{ab}); g_{ab}) - C(\chi; g_{ab}). \tag{3.22}$$

Thus, if ω_1 and ω_2 both satisfy Eq. (3.18), the difference $\{C(\omega_1; g_{ab}) - C(\omega_2; g_{ab})\}$ is a conformal invariant.

To summarize these results: I have shown that a solution to Eqs. (3.8) and (3.9) is provided by

$$T^{ab} = -2g^{-1/2}\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2}C(\omega(g_{cd});g_{cd}), \quad (3.23)$$

where ω is a scalar function of the curvature satisfying Eq. (3.18) and $C(\omega;g)$ is given by Eq. (3.15). There exists the freedom to add to T^{ab} any conserved, trace-free tensor. In terms of the function $w(x,x')$, this freedom corresponds to the freedom to add a symmetric solution to the homogeneous wave equation that has zero coincidence limit. {The function $v(x,x')$ provides a particular example. Recall that $v(x,x)$ is zero and $v_{ab}(x)$ is the variation of a conformally invariant action [Eq. (2.12)].}

It now remains to show that there exist scalar functions of the curvature, ω, that satisfy Eq. (3.18). These functions do indeed exist; they are more or less difficult to construct, depending upon whether or not the Weyl curvature of the space–time is zero.

When the space–time is not conformally flat ($C_{abcd} \neq 0$),

$$\omega = -\tfrac{1}{4}\ln C_{abcd}C^{abcd} \quad (3.24)$$

is the simplest to construct. Of course, it may be that C_{abcd} is not zero, but the particular invariant $C_{abcd}C^{abcd}$ is. In this case one can select any other, nonvanishing, invariant. One could take ω to be proportional to the logarithm of the sum of the squares of the independent invariants of the Weyl tensor; this would have some advantages. However, it still may not be the most natural choice. To see what might be more natural, it is necessary to see how T_{ab} contributes to the Green's function $G(x,x')$. It does this through the function $w(x,x')$. Combining Eqs. (2.26), (3.1), (3.4), and (3.7), $w(x,x')$ is now seen to have the form,

$$\begin{aligned} w(x,x') = {} & w_0(x) - \tfrac{1}{2}w_{0;a}(x)\sigma^{;a} \\ & + \tfrac{1}{2}\big[T_{ab} - v_1 g_{ab} + \tfrac{1}{6}R_{ab}w_0 \\ & + \tfrac{1}{3}(w_{0;ab} - \tfrac{1}{4}g_{ab}\Box w_0)\big]\sigma^{;a}\sigma^{;b} \\ & + O(\sigma^{3/2}). \end{aligned} \quad (3.25)$$

In the previous section I made the point that it was artificial to write G in the form of Eq. (2.2); in particular, $w(x,x')$ had to provide a term $-v(x,x')\ln L$. The Taylor series expansion of this term about the point x is provided by Eq. (2.10), and the necessary assumption that $L(x,x)$ is not zero. A term having exactly this structure is indeed provided by $w(x,x')$. Whatever the actual choice for ω, its scaling behavior is characteristic of a function that is the logarithm of a length; Eq. (3.24) is an example.

Consider taking the variation in Eq. (3.23) to obtain an explicit form for the tensor T^{ab}. It is easy to see that the only place where the logarithmic nature of ω survives is in the term

$$-\tfrac{1}{180}g^{-1/2}\omega\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2}(R_{abcd}R^{abcd} - R_{ab}R^{ab} + \Box R). \quad (3.26)$$

Elsewhere ω appears differentiated, either functionally or covariantly. Using the formulae in the Appendix, the term (3.26) can be shown to be equal to

$$-\tfrac{1}{120}g^{-1/2}\omega\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2}(C_{abcd}C^{abcd}). \quad (3.27)$$

(The variation of the other terms vanishes identically.)

Recalling Eqs. (2.12) and (3.25), one sees that the term (3.27) contributes to $w(x,x')$ an amount:

$$-\tfrac{1}{2}v_{ab}(x)\sigma^{;a}\sigma^{;b}\ln L^2(x), \quad (3.28)$$

where $\omega \equiv \ln L$. In short, ω provides the length that is missing in Eq. (2.2).

It would seem natural to choose ω to be a function of the biscalar $v(x,x')$, insofar as it is the existence of v that requires the existence of ω. The tensor v_{ab} is well suited to this purpose: its scaling behavior can be inferred from Eq. (2.12) and is given by

$$v^a{}_b(e^{-2\chi}g_{cd}) = e^{4\chi}v^a{}_b(g_{cd}). \quad (3.29)$$

The eigenvalues v_i of $v^a{}_b$ scale in the same way. Thus it is possible to choose for ω,

$$\omega = -(4d)^{-1}\ln h_d(v_i), \quad (3.30)$$

where $h_d(v_i)$ is any homogeneous function of degree d.

When the space–time is conformally flat v_{ab} vanishes. Indeed it can be shown that $v(x,x')$ vanishes.[9] If this is the case, then there is no pressing need to construct an ω satisfying Eq. (3.18). However, solutions do still exist and can be defined implicitly. For example,

$$\omega = -\ln\psi, \quad (3.31)$$

where ψ is a geometrical solution to the wave equation

$$(\Box - \tfrac{1}{6}R)\psi = 0,$$

which has the scaling behavior

$$\psi(e^{-2\chi}g_{ab}) = e^{\chi}\psi(g_{ab}).$$

Of course, functions of the type (3.31) will continue to provide solutions to Eq. (3.18) when $C_{abcd} = 0$. It can be shown[8] that by proceeding in this way one obtains for the tensor T^{ab} the polynomial expression[10]

$$\begin{aligned} T^{ab} = {} & \tfrac{1}{720}\big[6R^{ac}R^{b}{}_{c} + 2R^{;ab} \\ & - 6RR^{ab} - g^{ab}(2\Box R - 2R^2 + 3R_{cd}R^{cd})\big]. \end{aligned} \quad (3.32)$$

(One chooses for ω the solution that is conformal to a constant, the solution in flat space–time.)

The simple form for T^{ab} in Eq. (3.32) is essentially a feature of the conformal flatness. The variation of Eq. (3.15) is easy to compute because

$$\delta C = \int d^4x\,\{\hat{g}^{1/2}T^a{}_a(\hat{g}_{cd})\delta\omega - 2g^{1/2}T^{ab}\delta g_{ab}\}, \quad (3.33)$$

and, for the above choice of ω, the coefficient of $\delta\omega$ vanishes; one does not have to compute further the variation of ω with respect to the metric.

In general, when the Weyl tensor is nonzero, one can arrange for a similar simplification to take place: Require that $\omega(g_{cd})$ is determined by the condition

$$T^a{}_a(e^{-2\omega}g_{cd}) = 0. \quad (3.34)$$

This equation implies that ω satisfies (3.18) and has some interesting solutions.[11] A nongeometrical solution worth mentioning is provided by

$$\omega = -\tfrac{1}{2}\ln(K^a g_{ab}K^b), \quad (3.35)$$

where K^a is any curl-free, Killing vector field of the Ricci flat metric g_{ab}.[12]

IV. CONCLUSION

To some extent it is artificial to look for more or less natural functions ω that satisfy Eq. (3.18): for a given problem with prescribed boundary conditions an ω will be automatically provided. But, as I said in the Introduction, I was interested in how far the requirements of symmetry and having the Hadamard form determine the local structure of Feynman Green's functions. In this spirit, the hard conclusions of this paper are those contained in Eq. (3.25), (3.23), and (3.15). The rest is more speculative but, I hope, not without interest.

APPENDIX

The conventions used in this paper are consistent with Ref. 13. The following formulae were used in the derivation of the equations appearing in the text:

$$\hat{R}^{ab}{}_{cd} = e^{2\omega}(R^{ab}{}_{cd} + \delta^{[a}{}_{[c}\omega^{b]}{}_{d]}), \tag{A1}$$

$$\hat{R}^{b}{}_{d} = e^{2\omega}[R^{b}{}_{d} + \tfrac{1}{4}(2\omega^{b}{}_{d} + \delta^{b}{}_{d}\omega^{a}{}_{a})], \tag{A2}$$

$$\hat{R} = e^{2\omega}(R + \tfrac{3}{2}\omega^{a}{}_{a}), \tag{A3}$$

$$\hat{\Box}\phi = e^{2\omega}(\Box\phi - 2\omega^{;a}\phi_{;a}), \tag{A4}$$

where

$$\hat{R}_{abcd} \equiv R_{abcd}(e^{-2\omega}g_{ef}),$$

$$\omega_{ab} \equiv 4(\omega_{;ab} + \omega_{;a}\omega_{;b}) - 2g_{ab}\omega^{;c}\omega_{;c},$$

and a semicolon denotes covariant differentiation with respect to the metric g_{ab};

$$\begin{aligned}\sigma_{;ab}(x,x') = {} & g_{ab}(x) - \tfrac{1}{3}R_{acbd}(x)\sigma^{;c}\sigma^{;d} \\ & + \tfrac{1}{12}R_{acbd;e}\sigma^{;c}\sigma^{;d}\sigma^{;e} \\ & - (\tfrac{1}{60}R_{acbd;ef} + \tfrac{1}{45}R_{acgd}R_{be}{}^{g}{}_{f}) \\ & \times\sigma^{;c}\sigma^{;d}\sigma^{;e}\sigma^{;f} + O(\sigma^{5/2}),\end{aligned} \tag{A5}$$

$$\begin{aligned}\Delta^{1/2}(x,x') = {} & 1 + \tfrac{1}{12}R_{ab}\sigma^{;a}\sigma^{;b} - \tfrac{1}{24}R_{ab;c}\sigma^{;a}\sigma^{;b}\sigma^{;c} \\ & + (\tfrac{1}{288}R_{ab}R_{cd} + \tfrac{1}{360}R^{e}{}_{a}{}^{f}{}_{b}R_{ecfd} \\ & + \tfrac{1}{180}R_{ab;cd})\sigma^{;a}\sigma^{;b}\sigma^{;c}\sigma^{;d} + O(\sigma^{5/2}),\end{aligned} \tag{A6}$$

$$\begin{aligned}\Delta^{1/2}{}_{;ab}(x,x') = {} & \tfrac{1}{6}R_{ab} + \tfrac{1}{12}(2R_{c(a;b)} - R_{ab;c})\sigma^{;c} \\ & + (\tfrac{1}{40}R_{ab;cd} + \tfrac{1}{40}R_{cd;(ab)} - \tfrac{1}{15}R_{c(a;b)d} \\ & + \tfrac{1}{72}R_{ab}R_{cd} + \tfrac{1}{36}R_{ac}R_{bd} \\ & + \tfrac{1}{180}R_{e(a}R_{b)c}{}^{e}{}_{d} + \tfrac{1}{90}R_{aebf}R^{e}{}_{c}{}^{f}{}_{d} \\ & - \tfrac{1}{90}R^{e}{}_{cf(a}R_{b)}{}^{f}{}_{ed} \\ & - \tfrac{1}{90}R^{e}{}_{cf(a}R_{b)e}{}^{f}{}_{d} + \tfrac{11}{180}R_{ce}R^{e}{}_{(ab)d}) \\ & \times\sigma^{;c}\sigma^{;d} + O(\sigma^{3/2}),\end{aligned} \tag{A7}$$

[formulae (A5), (A6), and (A7) are taken from Ref. 14]

$$V_{a;[bc]} = \tfrac{1}{2}V_{d}R^{d}{}_{abc}, \tag{A8}$$

$$C_{abcd} = R_{abcd} + g_{a[d}R_{b]} - g_{b[d}R_{c]a} + \tfrac{1}{3}Rg_{a[c}g_{d]b}, \tag{A9}$$

$$C^{a}{}_{bcd;a} = R_{b[d;c]} - \tfrac{1}{6}g_{b[d}R_{;c]}, \tag{A10}$$

$$C_{abcd}C^{abcd} = R_{abcd}R^{abcd} - 2R_{ab}R^{ab} + \tfrac{1}{3}R^{2}, \tag{A11}$$

$$C_{acde}C_{b}{}^{cde} = \tfrac{1}{4}g_{ab}C_{efgh}C^{efgh}, \tag{A12}$$

$$\delta g^{1/2} = \tfrac{1}{2}g^{1/2}g^{ab}\delta g_{ab}, \tag{A13}$$

$$\delta\Gamma_{ab}{}^{c} = \tfrac{1}{2}g^{cd}(\delta g_{ad;b} + \delta g_{bd;a} - \delta g_{ab;d}), \tag{A14}$$

$$\delta R_{abc}{}^{d} = (\delta\Gamma_{ca}{}^{d})_{;b} - (\delta\Gamma_{cb}{}^{d})_{;a}, \tag{A15}$$

$$\delta R_{ab} = g^{cd}(\delta g_{c(a;b)d} - \tfrac{1}{2}\delta g_{ab;cd} - \tfrac{1}{2}\delta g_{cd;ab}), \tag{A16}$$

$$\delta R = g^{ab}g^{cd}(\delta g_{ac;db} - \delta g_{ab;cd}) - R^{ab}\delta g_{ab}. \tag{A17}$$

[1] J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations* (Yale U. P., New Haven, 1923); B. S. DeWitt and R. W. Brehme, "Radiation damping in a gravitational field," Ann. Phys. (N.Y.) **9**, 220 (1960).

[2] S. A. Fulling, M. Sweeny, and R. M. Wald, "Singularity structure of the two point function in quantum field theory in curved space–time," Commun. Math. Phys. **63**, 259 (1978); S. A. Fulling, F. J. Narcovwich, and R. M. Wald, "Singularity structure of the two point function in quantum field theory in curved space–time II," Ann. Phys. (N.Y.) **136**, 243 (1981).

[3] R. Penrose, *Techniques of Differential Topology in Relativity* (SIAM, Philadelphia, 1972).

[4] S. L. Adler, J. Lieberman, and Y. J. Ng, "Regularization of the stress–energy tensor for vector and scalar particles propagating in a general background metric," Ann. Phys. (N.Y.) **106**, 279 (1977); R. M. Wald, Phys. Rev. D **17**, 1477 (1978).

[5] M. R. Brown, A. C. Ottewill, and S. T. C. Siklos, "Comments on conformal Killing vector fields and quantum field theory," Phys. Rev. D **26**, 1881 (1982).

[6] S. L. Adler *et al.*, Ref. 4.

[7] M. J. Duff, "Observations on conformal anomalies," Nucl. Phys. B **125**, 334 (1977).

[8] M. R. Brown, "Actions and anomalies," preprint, University of Texas at Austin, 1978.

[9] F. G. Friedlander, *The Wave Equation on a Curved Space–Time* (Cambridge U.P., Cambridge, 1975).

[10] L. S. Brown and J. Cassidy, Phys. Rev. D **16**, 1712 (1977).

[11] M. R. Brown, "Quantum field theory and conformal transformations," preprint, Oxford University, 1981.

[12] D. N. Page, Phys. Rev. D **25**, 1499 (1982).

[13] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge U.P., Cambridge, 1973).

[14] S. M. Christensen, "Vacuum expectation value of the stress tensor in an arbitrary curved background: The covariant point separation method," Phys. Rev. D **14**, 2490 (1976).

Properties of the Schwinger model

Anton Z. Capri and Ruggero Ferrari[a)]
Theoretical Physics Institute, University of Alberta, Edmonton, Alberta, T6G 2J1, Canada

(Received 31 March 1983; accepted for publication 5 August 1983)

We present all the Wightman functions for an explicit operator solution of the Schwinger model. To understand these better, we study the algebra of fields of this model, representations of this algebra as well as the Hamiltonian. The latter turns out to elucidate the "confinement" of the fermion field. In addition we comment on the renormalization of the theory as well as on the analyticity of the amplitudes in terms of the coupling constant.

PACS numbers: 11.10.Mn

I. INTRODUCTION

The Schwinger model has proved to be a rich source of theoretical results for further conjectures as well as for testing conjectures. This makes it worthwhile to examine this model, in as much detail, and from as many perspectives, as possible. In a previous paper,[1] we presented an explicit operator solution of the Schwinger model for an arbitrary covariant gauge. The solution was local, Lorentz-covariant, chirally invariant, and the gauge transformations of the first kind were implementable.

In this paper we further examine properties of these solutions. In particular, we list all the Wightman functions, construct the Hamiltonian, and examine its spectrum. Finally we also comment briefly on the renormalization of the theory and the analyticity of the Wightman functions with respect to the coupling constant.

Throughout we use the same notation as in Ref. 1. To introduce this notation, we briefly review the results obtained in Ref. 1. When necessary, we use the following explicit conventions:

$$g^{00}=1, \quad \epsilon^{01}=1,$$

$$\gamma^0=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad \gamma^1=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}, \quad \gamma^5=\gamma^0\gamma^1.$$

We also define $F=F^{(+)}+F^{(-)}$ for any free quantized field F to be, respectively, the annihilation and creation parts of F.

The Schwinger model, as considered by us in Ref. 1, is defined by the formal Lagrangian

$$\mathscr{L}=-\tfrac{1}{4}(F_{\mu\nu})^2-\tfrac{1}{2}\alpha(\partial\cdot A)^2+\bar{\phi}\,(i\gamma\cdot\partial-e\gamma\cdot A)\phi. \tag{1}$$

The solutions for the Heisenberg fields ϕ, A_μ are given in terms of certain free "building block" fields as follows:

$$\phi(x)=Z^{-1/2}\exp[-ie\Omega^{(-)}(x)]\psi(x)\exp[-ie\Omega^{(+)}(x)], \tag{2}$$

$$A_\mu(x)=\partial_\mu c(x)+\epsilon_{\mu\nu}\,\partial^\nu d(x), \tag{3}$$

where

$$\Omega(x)=c(x)+\gamma^5 d(x), \tag{4}$$

$$c(x)=a(x)+\beta\rho(x),$$

$$d(x)=(\sqrt{\pi}/e)[\Sigma(x)+\sigma(x)]-(\alpha\pi/e^2)\tilde{b}(x). \tag{5}$$

Here β is a real parameter and the other quantities are free fields defined as follows:

$$\gamma\cdot\partial\psi(x)=0, \tag{6}$$

$$\Box a=b, \quad \Box b=0, \tag{7}$$

$$(\Box+e^2/\pi)\Sigma=0, \tag{8}$$

$$:\bar{\psi}\gamma_\mu\psi:(x)=(1/\sqrt{\pi})\partial_\mu\rho=(1/\sqrt{\pi})\epsilon_{\mu\nu}\,\partial^\nu\sigma, \tag{9}$$

and

$$\partial_\mu b=\epsilon_{\mu\nu}\,\partial^\nu\tilde{b}. \tag{10}$$

The relevant two-point functions for these fields are

$$\langle 0|\psi_a(x)\bar{\psi}_\beta(0)|0\rangle=-i(i\gamma\cdot\partial)_{\alpha\beta}D^{(+)}(x), \tag{11}$$

$$\langle 0|a(x)a(0)|0\rangle=-(i/\alpha)I^{(+)}(x)+i(\beta^2+2\beta\sqrt{\pi}/2)D^{(+)}(x), \tag{12}$$

$$\langle 0|\Sigma(x)\Sigma(0)|0\rangle=-i\Delta^{(+)}(x), \tag{13}$$

$$\langle 0|\rho(x)\rho(0)|0\rangle=\langle 0|\sigma(x)\sigma(0)|0\rangle=-iD^{(+)}(x), \tag{14}$$

$$\langle 0|\rho(x)\sigma(0)|0\rangle=-i\tilde{D}^{(+)}(x), \tag{15}$$

where

$$D^{(+)}(x)=(4\pi i)^{-1}\ln\mu^2(-x^2+i\epsilon x^0), \tag{16}$$

$$I^{(+)}(x)=(16\pi i)^{-1}x^2\ln\mu^2(-x^2+i\epsilon x^0). \tag{17}$$

$\Delta^{(+)}$ is the solution of $(\Box+e^2/\pi)\Delta^{(+)}=0$ with the normalization that yields

$$\partial_0\Delta^{(+)}(x)|_{x^0=0}=\delta(x^1)$$

and finally

$$\tilde{D}^{(+)}(x)=(4\pi i)^{-1}\ln[(x^0-i\epsilon+x^+)/(x^0-i\epsilon-x^1)]. \tag{18}$$

The finite normalization constant Z is given by

$$Z=(\sqrt{\pi}\mu/e)^{1/2}\exp[-\tfrac{1}{2}(\gamma-\ln 2)], \tag{19}$$

where γ is Euler's constant and μ is an arbitrary mass scale.

For further details, regarding properties of these solutions, the reader is referred to Ref. 1.

2. THE WIGHTMAN FUNCTIONS

Since the solution given in Ref. 1 conserves fermion number, the only nontrivial Wightman function involving only Fermi fields was already listed there and is given by

[a)] Permanent address: Istituto di Fisica, Università di Pisa, Piazza Torricelli 2, Pisa, Italy.

$$\langle 0|\phi(x_1)\cdots\phi(x_n)\phi^*(y_1)\cdots\phi^*(y_n)|0\rangle$$
$$= W_n(x,y)$$
$$= Z^{-n}\exp[\mathscr{F}^{(+)}(x,y)]w_0^{2n}(x,y), \tag{20}$$

where

$$\mathscr{F}^{+}(x,y) = \sum_{i,j=1}^{n} F^{(+)}(x_i,y_j) - \sum_{i<j=1}^{n}(F^{(+)}(x_i,x_j)+F^{(+)}(y_i,y_j)) \tag{21}$$

and

$$F^{(+)}(x,y) = e^2\left\{-\frac{i}{\alpha}I^{(+)}(x-y) - \frac{i\pi}{e^2}\gamma_x^5\gamma_y^5[\Delta^{(+)}(x-y)-D^{(+)}(x-y)]\right\}. \tag{22}$$

Here

$$w_0^2(x,y) = \frac{1}{2\pi i}\frac{\gamma\cdot(x-y)\gamma_0}{[(x-y)^2 - i\epsilon(x^0-y^0)]} = \langle 0|\psi(x)\psi^*(y)|0\rangle \tag{23}$$

is the free fermion two-point function and

$$w_0^{2n}(x,y) = \sum_{\text{Perm}}(-1)^{\delta p}\prod_{i=1}^{n} w_0^2(x_i,y_{ip}) = \det w_0^2(x_i,y_i) \tag{24}$$

is the free fermion $2n$-point function. It is worth noting that the two-point function can also be written as

$$\langle 0|\psi(x)\psi^*(y)|0\rangle = (\mu/2\pi)\exp 2\pi i \times\{\gamma^5\tilde{D}^{(+)}(x-y) - D^{(+)}(x-y)\}. \tag{25}$$

The vector potential A_μ is just a sum of free fields, as stated by Eq. (2). Thus all n-point functions of A_μ are just given in terms of the following two-point function:

$$\langle 0|A_\mu(x)A_\nu(y)|0\rangle = iH_{\mu\nu}^{(+)}(x-y) = (i/\alpha)\,\partial_\mu\partial_\nu I^{(+)}(x-y) + (i\pi/e^2)\,\partial_\mu\partial_\nu[\Delta^{(+)}(x-y) - D^{(+)}(x-y)] + ig_{\mu\nu}\Delta^{(+)}(x-y). \tag{26}$$

The result is

$$\langle 0|A_{\mu_1}(x_1)\cdots A_{\mu_{2n}}(x_{2n})|0\rangle = \sum\prod_{j_k}[iH_{\mu_{j_1}\mu_{j_2}}^{(+)}(x_{j_1}-x_{j_2})], \tag{27}$$

where the sum is over all partitions of $2n$ into n disjoint two-element subsets

$$(j_1j_2)(j_3j_4)\cdots(j_{2n-1}j_{2n}) \text{ with } j_{2k-1}<j_{2k}.$$

We next compute the simplest of the mixed Wightman functions, namely

$$\langle 0|A_\mu(z)\phi(x)\phi^*(y)|0\rangle = \langle 0|A_\mu^{(+)}(z)\phi(x)\phi^*(y)|0\rangle. \tag{28}$$

The computation is facilitated by using Eqs. (72)–(74) of Ref. 1, namely,

$$\phi(x) = \exp[-ie\Xi^{(-)}(x)]\zeta(x)\exp[-ie\Xi^{(+)}(x)] \tag{29}$$

with

$$\zeta(x) = \exp i\,P^{(-)}(x)\psi(x)\exp i\,P^{(+)}(x),$$
$$P(x) = \sqrt{\pi}(\rho(x) - \gamma^5\sigma(x)) \tag{30}$$

and

$$\Xi(x) = a(x) + \gamma^5(\sqrt{\pi}/2)\Sigma(x) - (\alpha\pi/e^2)\tilde{b}(x). \tag{31}$$

This is just a rewriting of the solution given by Eq. (2). We then obtain,

$$[A_\mu^{(+)}(z),\Xi^{(-)}(x)] = (i/e)G_\mu^{(+)}(z,x) = -i\{(1/\alpha)\partial_\mu I^{(+)}(z-x) + (\pi/e^2)\gamma^5\epsilon_{\mu\nu}\times\partial^\nu[\Delta^{(+)}(z-x) - D^{(+)}(z-x)]\}. \tag{32}$$

Now using the identity (for $[A,B]$ a c-number)

$$Ae^B = e^B(A + [A,B]),$$

we obtain

$$[A_\mu^{(+)}(z),\phi(x)] = G_\mu^{(+)}(z,x)\phi(x),$$
$$[A_\mu^{(+)}(z),\phi^*(x)] = -G_\mu^{(+)}(z,x)\phi^*(x), \tag{33}$$

where we used that $G^{(-)*}(x,y) = G^{(+)}(x,y)$.

Combining these results yields the desired Wightman function

$$\langle 0|A_\mu(z)\phi(x)\phi^*(y)|0\rangle = [-G_\mu^{(+)}(z,y) + G_\mu^{(+)}(z,x)]\langle 0|\phi(x)\phi^*(y)|0\rangle. \tag{34}$$

This generalizes immediately to

$$\langle 0|A_\mu(z)\phi(x_1)\cdots\phi(x_n)\phi^*(y_1)\cdots\phi^*(y_n)|0\rangle = \sum_{i=1}^{n}[G_\mu^{(+)}(z,x_i) - G_\mu^{(+)}(z,y_i)]W_n(x,y). \tag{35}$$

Further combining this result with Eq. (27), we find

$$\langle 0|A_{\mu_1}(z_1)\cdots A_{\mu_l}(z_l)\phi(x_1)\cdots\phi(x_n)\phi^*(y_1)\cdots\phi^*(y_n)|0\rangle = \sum_{r=1}^{n}\sum_{P(l)}\langle 0|A_{\mu_{i_{k+1}}}(z_{i_{k+1}})\cdots A_{\mu_{i_{k+l}}}(z_{i_{k+l}})\cdot|0\rangle \times\prod_{j=1}^{k}[G_{\mu_{i_j}}^{(+)}(z_{i_j},x_r) - G_{\mu_{i_j}}^{(+)}(z_{i_j},y_r)]W_n(x,y), \tag{36}$$

where the sum over $P(l)$ is over all partitions of l indices into two disjoint sets, with $i_j<i_k$ for $j<k$.

This completes the evaluation of all the Wightman functions. Before turning to the Hamiltonian, it is convenient to examine the operators $\zeta(x)$, $\zeta^*(x)$ given by Eq. (30). As we show later, they do not belong to the algebra of fields, but are nevertheless useful objects.

3. THE ζ-REPRESENTATION

We begin by considering the vacuum expectation value

$$Z(x,y) = \langle 0|\zeta(x_1)\cdots\zeta(x_n)\zeta^*(y_1)\cdots\zeta^*(y_n)|0\rangle. \tag{37}$$

To evaluate Z, we need

$$\zeta(x)\zeta(y)^* = \exp i[P^{(-)}(x) - P^{(-)}(y)]\psi(x)\psi^*(y) \times\exp i[P^{(+)}(x) - P^{(+)}(y)] \times\exp H^{(+)}(x,y), \tag{38}$$

$$\zeta(x_1)\zeta(x_2) = \exp i[P^{(-)}(x_1) + P^{(-)}(x_2)]\psi(x_1)\psi(x_2) \times\exp i[P^{(+)}(x_1) + P^{(+)}(x_2)] \times\exp[-H^{(+)}(x_1,x_2)], \tag{39}$$

$$\zeta^*(y_1)\zeta^*(y_2) = \exp - i[P^{(-)}(y_1) - P^{(-)}(y_2)]\psi^*(y_1)\psi^*(y_2)$$
$$\times\exp\{ - i[P^{(+)}(y_1) + P^{(+)}(y_2)]\}$$
$$\times\exp[- H^{(+)}(y_1, y_2)], \quad (40)$$

where

$$H^{(+)}(x, y) = i\pi[(1 + \gamma_x^5\gamma_y^5)D^{(+)}(x - y)$$
$$- (\gamma_x^5 + \gamma_y^5)\tilde{D}^{(+)}(x - y)]. \quad (41)$$

The subscripts x, y, etc. on γ_x^5 indicate on which side of a quantity γ^5 is to be multiplied. Thus $(\gamma_x^5 F(x, y))_{\alpha\beta}$ $= \gamma_{\alpha\gamma}^5 F_{\gamma\beta}(x, y)$ whereas $(\gamma_y^5 F(x, y))_{\alpha\beta} = F_{\alpha\gamma}(x, y)\gamma_{\gamma\beta}^5$. It then follows that

$$Z(x, y) = \exp \mathscr{H}(x, y) W_0^{2n}(x, y), \quad (42)$$

where

$$\mathscr{H}(x, y) = \sum_{i,j=1}^{n} H^{(+)}(x_i, y_j)$$
$$- \sum_{i<j=1}^{n} [H^{(+)}(x_i, x_j) + H^{(+)}(y_i, y_j)]. \quad (43)$$

To further evaluate this expression, we notice that both $\mathscr{H}(x, y)$ and $w_0^{2n}(x, y)$ are diagonal in the spinor indices. If we now consider all spinor indices to have the value 1, we find

$$w_{011}^2 = (2\pi i)^{-1}[x^0 - y^0 - (x^1 - y^1) - i\epsilon]^{-1} \quad (44)$$

and

$$w_{011\cdots11}^{2n}(x, y) = \det\{(2\pi i)^{-1}[x_i^- - y_j^- - i\epsilon]^{-1}\}$$
$$= (2\pi i)^{-n}\frac{\Pi_{i<j}(x_i^- - x_j^-)(y_i^- - y_j^-)}{\Pi_{i,j}(x_i^- - y_j^- - i\epsilon)}. \quad (45)$$

The last step above is proven in Refs. 2 and 3.

To complete the computation, we write out $\exp \mathscr{H}^{(+)}$ for $\gamma_x^5 = \gamma_y^5 = 1$ and use Eq. (25) to get

$$\exp \mathscr{H}^{(+)}(\gamma_x^5 = \gamma_y^5 = 1)$$
$$= \mu^n i^n \frac{\Pi_{i,j}(x_i^- - y_i^-)}{\Pi_{i<j}(x_i^- - x_j^- - i\epsilon)(y_i^- - y_j^- - i\epsilon)}, \quad (46)$$

and hence

$$\langle 0|\zeta_1(x_1)\cdots\zeta_1(x_n)\zeta_1^*(y_1)\cdots\zeta_1^*(y_n)|0\rangle = (\mu/2\pi)^n.$$

A similar computation for general spurion indices yields the following result:

$$\langle 0|\zeta_1(x_1)\cdots\zeta_1(x_n)\zeta_2(y_1)\cdots\zeta_2(y_m)\zeta_1^*(z_1)$$
$$\cdots\zeta_1^*(z_{n'})\zeta_2^*(w_1)\cdots\zeta_2^*(w_{m'})|0\rangle$$
$$= \delta_{n,n'}\delta_{m,m'}(-)^{n\cdot m}(\mu/2\pi)^{n+m}. \quad (47)$$

Thus we see that the algebra specified by Eq. (75) of Ref. 1 is represented on a Hilbert space with an orthonormal basis

$$|n,m\rangle = \left(\frac{\mu}{2\pi}\right)^{-(|n|+|m|)/2}(\zeta_1^*)^{(n+|n|)/2}\zeta_1^{(n-|n|)/2}$$
$$\times(\zeta_2^*)^{(m+|m|)/2}\zeta_2^{(m-|m|)/2}|0\rangle \quad (48)$$

for $n, m = 0, \pm 1, \pm 2, \pm 3, \ldots$ and $\langle n,m|n',m'\rangle$ $= \delta_{n,n'}\delta_{m,m'}$.

In this representation one has

$$\zeta_1\zeta_1^* = \zeta_2\zeta_2^* = \mu/2\pi, \quad (49)$$

and two charges q and q_5 can be defined by

$$[q, \zeta] = \zeta, \quad [q_5, \zeta] = \gamma^5\zeta, \quad (50)$$
$$q|0\rangle = q_5|0\rangle = 0.$$

It is worth noting that the algebra given by Eq. (75) of Ref. 1 has many representations, not just the one given above. The representations can even be finite if the charges q, q_5 are omitted from the algebra. An example of such a finite representation is

$$\zeta_1 = \zeta_1^* = \left(\frac{\mu}{2\pi}\right)^{1/2}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$$
$$\zeta_2 = \zeta_2^* = \left(\frac{\mu}{2\pi}\right)^{1/2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \quad (51)$$

4. THE HAMILTONIAN

The solutions of the Schwinger model in Ref. 1 were constructed to preserve the gauge invariance of the underlying formal Lagrangian equation (1). Thus, in regularizing the terms appearing in the Hamiltonian, we must respect this invariance. A straightforward calculation shows that the required condition is

$$\tau(\Lambda)P_\mu\tau^{-1}(\Lambda) = P_\mu + \alpha\int dx^1\, \partial\cdot A\, \ddot{\partial}_0\partial_\mu\Lambda, \quad (52)$$

where $\Box\Lambda = 0$ and P_μ is the generator of space-time translations. The operator $\tau(\Lambda)$ is explicitly given by

$$\tau(\Lambda) = \exp i\alpha\int dx^1\, \partial\cdot A\, \ddot{\partial}_0\Lambda \quad (53)$$

and has the properties that

$$\tau(\Lambda)A_\mu(x)\tau^{-1}(\Lambda) = A_\mu - \partial_\mu\Lambda,$$
$$\tau(\Lambda)\phi(x)\tau^{-1}(\Lambda) = \exp(ie\Lambda)\phi(x). \quad (54)$$

To see how the condition implied by Eq. (52) is implemented, we consider the *classical*, unsymmetrized energy–momentum tensor

$$K^{\mu\nu} = i\bar{\phi}\gamma^\mu\partial^\nu\phi - F^{\mu\rho}A_{\rho,\nu} - \alpha g^{\mu\rho}A_\rho{}^{,\nu}\,\partial\cdot A$$
$$- g^{\mu\nu}[- \tfrac{1}{4}F_{\rho\sigma}F^{\rho\sigma}$$
$$- \tfrac{1}{2}\alpha(\partial\cdot A)^2 + \bar{\phi}(i\gamma\cdot\partial - e\gamma\cdot A)\phi]. \quad (55)$$

The classical momentum operator is given by

$$P^\nu = \int dx^1\, K^{0\nu}. \quad (56)$$

Examining the individual terms (appearing in P^ν) under a gauge transformation, we find that Eq. (52) is valid if the following transformations hold:

$$\tau(\Lambda)\int(-\alpha A^{0,\nu}\,\partial\cdot A)\,dx^1\,\tau^{-1}(\Lambda)$$
$$= -\alpha\int A^{0,\nu}\,\partial\cdot A\,dx^1 + \alpha\int\partial\cdot A\,\partial_0\partial^\nu\Lambda\,dx^1, \quad (57)$$

$$\tau(\Lambda)i\int dx^1\,\bar{\phi}\gamma^0\gamma^\nu\phi\tau^{-1}(\Lambda)$$
$$= i\int dx^1\,\bar{\phi}\gamma^0\,\partial^\nu\phi - e\int dx^1\,\bar{\phi}\gamma^0\phi\,\partial^\nu\Lambda, \quad (58)$$

$$\tau(\Lambda)\int dx^1\, F^{0\rho}A_{\rho,}{}^{,\nu}\tau^{-1}(\Lambda)$$
$$=\int dx^1\, F^{0\rho}A_{\rho,}{}^{,\nu}-\int dx^1\,\partial^\rho F_{\rho 0}\,\partial^\nu\Lambda. \tag{59}$$

The most problematic term is the term $i\bar\phi\gamma^\mu\partial^\nu\phi$ encountered in Eq. (58). To define this term, we use the gauge-invariant point splitting given by

$$\exp\left\{-ie\int_x^y d\xi^\mu A_\mu^{(-)}(\xi)[i\bar\phi(y)\gamma^\mu\partial^\nu\phi(x)]\right\}$$
$$\times\exp\left[-ie\int_x^y d\xi^\mu A_\mu^{(+)}(\xi)\right]. \tag{60}$$

We expand this expression in a power series in $\eta=y-x$, subtract the singular parts, and verify that the result is compatible with Eq. (58) as well as the general properties of P^ν. The procedure is not covariant but can be made so by an averaging over all η_μ.[4] This averaging is discussed in Appendix A and has the effect of replacing any product of η_μ's as follows:

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_{2n}}}=\frac{(\eta^2)^n}{2n!!}\sum g_{\mu_i\mu_j}\cdots g_{\mu_k\mu_l}, \tag{61}$$

where the sum runs over all partitions of 1,2,...,2n into pairs such that $i<j$. Also

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_{2n+1}}}=0.$$

Now using

$$\phi^*(y)\phi(x)=Z^{-1}\exp F^{(+)}(x,y)$$
$$\times\exp ie[\Omega^{(-)}(y)-\Omega^{(-)}(x)]:\psi^*(y)\psi(x):$$
$$\times\exp ie[\Omega^{(+)}(y)-\Omega^{(+)}(x)]+Z^{-1}$$
$$\times\exp F^{(+)}(x,y)\exp ie[\Omega^{(-)}(y)-\Omega^{(-)}(x)]$$
$$\times\langle 0|\psi^*(y)\psi(x)|0\rangle$$
$$\times\exp ie[\Omega^{(+)}(y)-\Omega^{(+)}(x)]. \tag{62}$$

We find on inserting this result in the expression (60) that in the first term the limit $\eta\to 0$ can be taken immediately to yield

$$i:\bar\psi\gamma^\mu\partial^\nu\psi:(x)+e:\psi\gamma^\mu\partial^\nu\Omega\psi:(x)$$
$$=i:\bar\psi\gamma^\mu\partial^\nu\psi:(x)+m[:\gamma^\mu\rho\partial^\nu c:(x)-:\partial^\mu\sigma\partial^\nu d:(x)]. \tag{63}$$

To obtain this, we have used that $\gamma^\mu\gamma^5=-\epsilon^{\mu\nu}\gamma_\nu$ and that $\lim_{\eta\to 0}\exp F^{(+)}(x,x+\eta)=Z$. The second term requires more work and when inserted in (60) yields

$$iZ^{-1}\exp F(x,y)\exp ie\left[\Omega(y)-\Omega(x)-\int_x^y d\xi^\rho A_\rho(\xi)\right]^{(-)}$$
$$\times\{[\partial_x^\nu F(x,y)-ie\,\partial^\nu\Omega(x)]\langle 0|\psi^*(y)\gamma^0\gamma^\mu\psi(x)|0\rangle$$
$$+\langle 0|\psi^*(y)\gamma^0\gamma^\mu\partial^\nu\psi(x)|0\rangle\}$$
$$\times\exp ie\left[\Omega(y)-\Omega(x)-\int_x^y d\xi^\rho A_\rho(\xi)\right]^{(+)}. \tag{64}$$

Now

$$\Omega(y)-\Omega(x)-\int_x^y d\xi^\rho A_\rho(\xi)$$
$$=\gamma_x^5[d(y)-d(x)]-\int_x^y d\xi^\rho\,\xi_{\rho\sigma}\,\partial^\sigma d(\xi). \tag{65}$$

Therefore, we obtain, after expanding in η and retaining only the nonvanishing contributions for $\eta\to 0$, that

$$:\exp ie\left[\Omega(y)-\Omega(x)-\int_x^y d\xi^\rho A_\rho(\xi)\right]:$$
$$\to 1+ie(\gamma_x^5\eta^\rho\partial_\rho d-\eta^\rho\epsilon_{\rho\sigma}\partial^\sigma d)+\tfrac{1}{2}\eta^\alpha\eta^\beta$$
$$\times:-e^2(\gamma_x^5\partial_\alpha d-\epsilon_{\alpha\rho}\partial^\rho d)(\gamma_x^5\partial_\beta d-\epsilon_{\beta\sigma}\partial^\sigma d)$$
$$+ie(\gamma_x^5\partial_\alpha\partial_\beta d-\epsilon_{\alpha\sigma}\partial^\sigma\partial_\beta d):. \tag{66}$$

Combining this with the remaining expressions in (64), we find, after a straightforward computation using the averaging process described in Appendix A and Eq. (61) that the fermion kinetic term when properly defined yields

$$i\bar\phi\gamma^\mu\partial^\nu\phi\to\tfrac{1}{2}i[:\bar\psi\gamma^\mu\partial^\nu\psi:-\partial^\nu:\bar\psi\gamma^\mu\psi:]$$
$$+m:(\partial^\mu\rho-m\epsilon^{\mu\alpha}\partial_\alpha d)\partial^\nu c:$$
$$-m:\partial^\mu\sigma\partial^\nu d:$$
$$+\tfrac{1}{2}m^2g^{\mu\nu}:\partial^\rho d\,\partial_\rho d:, \tag{67}$$

where $m^2=e^2/\pi$. Evaluating the rest of the terms in $K^{\mu\nu}$ and combining all the results, we obtain

$$K^{\mu\nu}$$
$$=\tfrac{1}{2}i[:\bar\psi\gamma^\mu\partial^\nu\psi:-\partial^\nu:\bar\psi\gamma^\mu\psi:]+\alpha:\partial^\nu c\,\partial^\mu b-\partial^\mu c\partial^\nu b:$$
$$-m:\epsilon^{\mu\alpha}\partial_\alpha\Sigma\,\partial^\nu c+\Sigma\epsilon^{\mu\rho}\partial_\rho\partial^\nu c$$
$$+(\partial^\mu\sigma-m\,\partial^\mu d)\partial^\nu d:$$
$$+m^2:\tfrac{1}{2}g^{\mu\nu}\partial_\rho d\,\partial^\rho d-\partial^\mu d\,\partial^\nu d:$$
$$-m:\Sigma\,\partial^\mu\partial^\nu d+(\alpha/m)b\epsilon^{\mu\rho}\partial_\rho\partial^\nu d:$$
$$+g^{\mu\nu}:-\tfrac{1}{2}m^2\Sigma^2+\tfrac{1}{2}\alpha b^2:. \tag{68}$$

After some rewriting we then obtain the Hamiltonian

$$H=\int dx^1:\frac{i}{2}[\bar\psi\gamma^0\partial^0\psi-\partial^0\bar\psi\gamma^0\psi]$$
$$+\alpha\left(\partial^0a\,\partial^0b+\partial^1a\,\partial^1b-\frac{b^2}{2}\right)$$
$$+\alpha\beta(\partial^0\rho\,\partial^0b+\partial^1\rho\partial^1b)-\frac{1}{2}\left[\left(\partial^0\left(\sigma-\frac{\alpha}{m}\tilde b\right)\right)^2\right.$$
$$\left.+\left(\partial^1\left(\sigma-\frac{\alpha}{m}\tilde b\right)\right)^2\right]$$
$$+\frac{1}{2}[(\partial^0\Sigma)^2+(\partial^1\Sigma)^2+m^2\Sigma^2]:. \tag{69}$$

This can also be rewritten as

$$H=H_0+H_1 \tag{70}$$

with

$$H_0=\int dx^1:\frac{i}{2}[\bar\psi\gamma^0\partial^0\psi-\partial^0\bar\psi\gamma^0\psi]$$
$$+\alpha\left(\partial^0a\,\partial^0b+\partial^1a\,\partial^1b-\frac{b^2}{2}\right)$$
$$+\frac{\alpha^2}{2}\left(\beta^2+\frac{2\beta}{m}\right)[(\partial_0b)^2+(\partial_1b)^2]$$
$$+\frac{1}{2}[(\partial_0\Sigma)^2+(\partial_1\Sigma)^2+m^2\Sigma^2]: \tag{71}$$

and

$$H_1 = -\frac{1}{2}\int dx^1 : \left[\partial^0\left(\rho - \alpha\left(\beta + \frac{1}{m}\right)b\right)\right]^2 + \left[\partial^1\left(\rho - \alpha\left(\beta + \frac{1}{m}\right)b\right)\right]^2 : . \tag{72}$$

H_0 is clearly the Hamiltonian of the free building-block fields a, b, Σ, and ψ. Similarly starting from equation (68) we find that the momentum operator is given by

$$P = \int dx^1 K^{01} = P_0 + P_1, \tag{73}$$

where

$$P_0 = \int dx^1 : i\bar{\psi}\gamma^0\partial^1\psi + \alpha(\partial^1 a\partial^0 b + \partial^0 a\partial^1 b) + \alpha^2\left(\beta^2 + \frac{2\beta}{m}\right)\partial^0 b\partial^1 b + \partial^0\Sigma\partial^1\Sigma : \tag{74}$$

and

$$P_1 = \int dx^1 : -\partial^0\left[\rho - \alpha\left(\beta + \frac{1}{m}\right)b\right]\partial^1\left[\rho - \alpha\left(\beta + \frac{1}{m}\right)b\right] : . \tag{75}$$

Here again P_0 is the momentum operator for the free building-block fields a, b, Σ, and ψ. The Lagrangian corresponding to H_0 and P_0 is

$$\begin{aligned}\mathscr{L}_0 = {} & :\bar{\psi}i\gamma\cdot\partial\psi: + \alpha:\partial^\mu a\,\partial_\mu b - \tfrac{1}{2}b^2: \\ & + \tfrac{1}{2}\alpha(\beta^2 + 2\beta/m):(\partial_\mu b)^2: \\ & + \tfrac{1}{2}:(\partial_\mu\Sigma)^2 + m^2\Sigma^2:\end{aligned} \tag{76}$$

and involves only the free fields ψ, $\bar{\psi}$, a, b, and Σ.

A straightforward application of the equal-time commutation relations listed in Appendix B yields

$$\begin{aligned}&[H_0, \psi] = -i\dot{\psi}, \quad [H_0, \Sigma] = -i\dot{\Sigma}, \\ &[H_0, a] = -i\dot{a}, \quad [H_0, b] = -i\dot{b}, \\ &[H_0, \tilde{b}] = -i\dot{\tilde{b}}, \quad [H_0, \rho] = -i\dot{\rho}, \\ &[H_0, \sigma] = -i\dot{\sigma}.\end{aligned} \tag{77}$$

Using the equal-time commutation relations listed in Appendix B once more, we then find

$$\begin{aligned}&[H, \psi] = -i\dot{\psi} + \sqrt{\pi}\{[\dot{\rho} - \alpha(\beta + 1/m)\dot{b}] - [\dot{\sigma} - \alpha(\beta + 1/m)\dot{\tilde{b}}]\gamma^5\}\psi, \\ &[H, \psi] = -i\dot{\Sigma}, \\ &[H, a] = -i\dot{a} - i(\beta + 1/m)[\dot{\rho} - \alpha(\beta + 1/m)\dot{b}], \\ &[H, b] = -i\dot{b}, \quad [H, \tilde{b}] = -i\dot{\tilde{b}}, \\ &[H, \rho] = -i\alpha(\beta + 1/m)\dot{b}, \quad [H, \sigma] = -i\alpha(\beta + 1/m)\dot{\tilde{b}}.\end{aligned} \tag{78}$$

The sets of equations (77) clearly show that H_0 provides a time evolution operator for all the building-block fields and hence for the full algebra of fields $\mathfrak{A}(A_\mu, \phi, \bar{\phi})$. This is in fact what one would naively expect.

In addition to H_0, we have, however, the full Hamiltonian H, and, although it is not obvious from the set of equations (78), this Hamiltonian H is also a time evolution operator for the algebra of fields $\mathfrak{A}(A_\mu, \phi, \bar{\phi})$. This point will be clarified after we discuss the field algebra in the next section.

What is the role of these two Hamiltonians? H_0 provides the time evolution of $\mathfrak{A}$ but does not provide the full physical content of the theory. The subtle message obtained from the full operator H is that in addition to the obvious spectrum obtained from H_0 there are infinitely many Poincaré-invariant states so that we have infinitely many copies of the spectrum of H_0 (excluding the fermions) built up on these translation invariant states. The details of this will be expounded in Secs. 6 and 7 and will reveal just how subtle the confinement of fermions (exclusion from the spectrum of H) is.

5. THE ALGEBRA OF FIELDS

We consider those objects which are local relative to the fields ϕ, $\bar{\phi}$, and A_μ. Some of the useful properties of this algebra $\mathfrak{A}$ are

(i) $b = \partial^\mu A_\mu \in \mathfrak{A}$.

(ii) Since $\tilde{b}$ is not local relative to ϕ we have $\tilde{b} \notin \mathfrak{A}$. However, $\partial_\mu\tilde{b} = \epsilon_{\mu\nu}\partial^\nu b \in \mathfrak{A}$.

(iii) $\Sigma = -(1/m)\epsilon^{\mu\nu}F_{\mu\nu} \in \mathfrak{A}$.

(iv) The dipole field is not local relative to ϕ and hence $a \notin \mathfrak{A}$. However,

$$\partial_\mu[a + (\beta + 1/m)\rho] = A_\mu + (\alpha/m^2)\,\partial_\mu\partial\cdot A - (1/m^2)\,\partial^\nu F_{\mu\nu} \in \mathfrak{A}.$$

The next property of the algebra requires a proof and is thus stated as a lemma.

Lemma:

(v) $\rho, j_{f\mu} = (1/\sqrt{\pi})\partial_\mu\rho$, σ, and $\partial_\mu\sigma$ are *not* elements of $\mathfrak{A}$ if A_μ, ϕ, and $\bar{\phi}$ are irreducibly represented.

Proof: Suppose they belong to $\mathfrak{A}$ and choose $\beta = -1/m$; then they commute with ϕ, $\bar{\phi}$, and A_μ and should be c-numbers. This is contradicted by

$$\langle 0|\rho(x)\rho(y)|0\rangle = -iD^{(+)}(x-y),$$
$$\langle 0|\sigma(x)\sigma(y)|0\rangle = -iD^{(+)}(x-y),$$
$$\langle 0|\sigma(x)\rho(y)|0\rangle = -i\tilde{D}^{(+)}(x-y).$$

For a general value of β consider

$$\rho_0 = \rho - \alpha(\beta + 1/m)b,$$
$$\sigma_0 = \sigma - \alpha(\beta + 1/m)\tilde{b}.$$

Then ρ_0 and σ_0 again commute with ϕ, $\bar{\phi}$, and A_μ and the same argument applies.

(vi) Combining the results of (iv) and (v), we find that for $\beta \neq -1/m$

$$\partial_\mu a = \partial_\mu[a + (\beta + 1/m)\rho] - (\beta + 1/m)\,\partial_\mu\rho \notin \mathfrak{A}.$$

(vii) Since ζ_a and ζ_a^* are not local with respect to ϕ and $\bar{\phi}$, they also do not belong to $\mathfrak{A}$.

In view of these results, it is convenient to use instead of the original building-block fields a and ψ the compound fields

$$a_0 = a - (\alpha/m^2)b + (\beta + 1/m)\rho \tag{79}$$

and

$$\zeta = :\exp[i\sqrt{\pi}(\rho - \sigma\gamma^5)]\psi:. \tag{80}$$

These fields commute with each other and satisfy the same field equations as a and ψ, respectively. Furthermore, we can express A_μ and ϕ in terms of these fields

$$A_\mu = \partial_\mu a_0 + \frac{1}{m}\epsilon_{\mu\nu}\partial^\nu \Sigma, \tag{81}$$

$$\phi = :\exp\{-ie[a_0 + (\alpha/m^2)b + (1/m)(\Sigma - (\alpha/m)\tilde{b})\gamma^5]:\zeta\}, \tag{82}$$

and we see that the parameter β has completely disappeared. Next we find that

$$[H, \zeta] = -i\dot{\zeta}, \tag{83}$$

$$[H, a_0] = -i\dot{a}_0, \tag{84}$$

$$[H, b] = -i\dot{b}, \tag{85}$$

$$[H, \tilde{b}] = -i\dot{\tilde{b}}, \tag{86}$$

$$[H, \Sigma] = -i\dot{\Sigma}. \tag{87}$$

It is still true that ζ, a_0, and $\tilde{b}$ are not elements of $\mathfrak{A}$. If, however, we choose test functions $f_0 \in \mathscr{S}$ vanishing at $p_\mu = 0$, then both $a_0(f_0)$ and $\tilde{b}(f_0)$ belong to $\mathfrak{A}$, the algebra of fields. It would now be easy to read off the spectrum of H except that we find a host of Poincaré-invariant states in addition to the obvious vacuum. We examine these next.

6. POINCARÉ-INVARIANT STATES

To find translation-invariant states, we begin by "undressing" the fermion field ϕ. To do this requires exponentiating certain elements of the algebra $\mathfrak{A}$. We define such exponentials using the triple-dot-product. Thus for any free field $A \in \mathfrak{A}$, we define

$$:\exp A: \equiv \exp A^{(-)} \exp A^{(+)}. \tag{88}$$

For convenience we also choose the value $\beta = -1/m$ in this section. Since $\Sigma \in \mathfrak{A}$, we can "remove" Σ from ϕ and obtain

$$\begin{aligned}\phi_0(x) &\equiv \exp(ie/m)\gamma^5\Sigma^{(-)}(x)\phi(x)\exp(ie/m)\gamma^5\Sigma^{(+)}(x)\\ &= Z^{-1}\exp[-ie(a - (\alpha/m^2)\gamma^5\tilde{b})^{(-)}](x)\zeta(x)\\ &\quad\times\exp[-ie\{-(\alpha/m^2)\tilde{b}\gamma^5\}^{(+)}](x).\end{aligned} \tag{89}$$

Since $a \notin \mathfrak{A}$ but $\partial_\mu a$ is, we cannot "undress" ϕ_0 any further. For this reason we consider bilocal fields which can be undressed as far as the field a is concerned. Due to the presence of γ^5, the $\tilde{b}$ cannot be removed. Thus we consider

$$\phi_0(x)\phi_0^*(y)\exp\left[-ie\int_x^y d\xi^\mu\,\partial_\mu a(\xi)\right]$$

and multiply by the necessary c-number factors to obtain the bilocal field

$$\begin{aligned}B(x,y) &= \exp[(ie\alpha/m^2)(\tilde{b}(x)\gamma_x^5 - \tilde{b}(y)\gamma_y^5)^{(-)}]\zeta(x)\zeta^*(y)\\ &\quad\times\exp[(ie\alpha/m^2)(\tilde{b}(x)\gamma_x^5 - \tilde{b}(y)\gamma_y^5)^{(+)}],\end{aligned} \tag{90}$$

which belongs to the algebra $\mathfrak{A}$. This field has the following local properties:

$$\begin{aligned}[A_\mu(z), B(x,y)] &= (e/m^2)[\partial_\mu\tilde{D}(z-x)\gamma_x^5\\ &\quad - \partial_\mu\tilde{D}(z-y)\gamma_y^5]B(x,y),\end{aligned} \tag{91}$$

$$\begin{aligned}&[\phi(z), B(x,y)]\\ &\quad= \{\exp i\pi[\tilde{D}(z-x) - \tilde{D}(z-y)] - 1\}B(x,y)\phi(z)\\ &\qquad\text{for } \zeta(x) = \zeta(y)\\ &\quad= -\{\exp i\pi[\tilde{D}(z-x) + \tilde{D}(z-y)] + 1\}B(x,y)\phi(z)\\ &\qquad\text{for } \zeta(x) \neq \zeta(y),\end{aligned} \tag{92}$$

and we see that both commutators vanish whenever z is spacelike with respect to both x and y. Thus $B(x,y)$ is truly bilocal.

We next consider the vacuum expectation value

$$\langle 0|\phi(x)\phi^*(y)B(z,w)|0\rangle.$$

Then (keeping always $\beta = -1/m$) using the commutator

$$\begin{aligned}K^{(+)}(x,y) &\equiv [\Xi^{(+)}(x), \Xi^{(-)}(y)]\\ &= -(i/\alpha)I^{(+)}(x,y) - (i/m^2)D^{(+)}(x-y)\\ &\quad + (i/m^2)(\gamma_x^5 + \gamma_y^5)\tilde{D}^{(+)}(x-y)\\ &\quad - (i/m)\gamma_x^5\gamma_y^5\Delta^{(+)}(x-y)\end{aligned} \tag{93}$$

and the identity given by Eq. (25), we find

$$\begin{aligned}&\langle 0|\phi(x)\phi^*(y)B(z,w)|0\rangle\\ &\quad= Z^{-1}\exp[e^2K^{(+)}(x,y)](\mu/2\pi)^2\\ &\qquad\times\{\delta xw\,\delta xy\exp i\pi\gamma_w^5\\ &\qquad\times[\tilde{D}^{(+)}(y-z) - \tilde{D}^{(+)}(x-z) + \tilde{D}^{(+)}(x-w)\\ &\qquad - \tilde{D}^{(+)}(y-w)] + \delta xw\,\delta yz[1 - \delta wz]\\ &\qquad\times\exp i\pi\gamma_w^5[-\tilde{D}(y-z) + \tilde{D}^{(+)}(x-z)\\ &\qquad + \tilde{D}^{(+)}(x-w) - \tilde{D}^{(+)}(y-w)]\},\end{aligned} \tag{94}$$

where the Kronecker delta refers to the Lorentz indices.

Next we take the limit $w \to z$ and consider the three components B_{11}, B_{22}, and B_{12} separately to find

$$\begin{aligned}&\lim_{w\to z}\langle 0|\phi(x)\phi^*(y)B_{11}(z,w)|0\rangle\\ &\quad= \lim_{w\to z}\langle 0|\phi(x)\phi^*(y)B_{22}(z,w)|0\rangle\\ &\quad= Z^{-1}\exp e^2[K^{(+)}(x,y)](\mu/2\pi)^2\end{aligned} \tag{95}$$

so that

$$\lim_{w\to z}B_{11}(z,w) = \lim_{w\to z}B_{22}(z,w) = \mu/2\pi. \tag{96}$$

On the other hand we obtain

$$\begin{aligned}&\lim_{w\to z}\langle 0|\phi(x)\phi^*(y)B_{12}(z,w)|0\rangle\\ &\quad= Z^{-1}(\mu/2\pi)^2\exp e^2K^{(+)}(x,y)\\ &\qquad\times\exp 2\pi i[\tilde{D}^{(+)}(y-z) - \tilde{D}^{(+)}(x-z)]\\ &\quad= -(\mu/2\pi)\langle 0|\phi(x)\phi^*(y)\sigma_+(z)|0\rangle.\end{aligned} \tag{97}$$

We now look for translation-invariant states by Fourier transforming the relevant part of Eq. (97) with respect to z. For convenience we also choose $x = 0$. The relevant expression is

$$\begin{aligned}&\int d^2z\,e^{-ip\cdot z}\exp 2\pi i[\tilde{D}^{(+)}(y-z) - \tilde{D}^{(+)}(-z)]\\ &\quad= \frac{1}{2}\int dz^+\,dz^-\exp\left[-\frac{i}{2}(p^+z^- + p^-z^+)\right]\\ &\qquad\times\left(\frac{y^+ - z^+ - i\epsilon}{y^- - z^- - i\epsilon}\right)^{1/2}\left(\frac{-z^- - i\epsilon}{-z^+ - i\epsilon}\right)^{1/2}.\end{aligned} \tag{98}$$

If we consider the integral over z^+ we find

$$\int_{-\infty}^{\infty} dz^+ \, e^{-ip^- z^+/2}\left[\left(\frac{y^+ - z^+ - i\epsilon}{-z^+ - i\epsilon}\right)^{1/2} - 1 + 1\right]$$
$$= 2\pi\delta\left(\frac{p^-}{2}\right) + 2\pi i \cdot 2\theta(p^-)$$
$$\times \int_0^{y^+} \left(\frac{y^+ - z^+}{z^+}\right)^{1/2} e^{-ip^- z^+/2}\, dz^+. \tag{99}$$

The second term is an analytic function of p^- since it is the Fourier transform of a function with compact support.

Thus Eq. (98) becomes

$$\int d^2 z \, e^{-ip\cdot z} \exp 2\pi i[\tilde{D}^{(+)}(y - z) - \tilde{D}^{(+)}(-z)]$$
$$= (2\pi)^2 \delta^{(2)}(p) + \text{terms in } \theta(p^-)\delta(p^+),\ \theta(p^+)\delta(p^-)$$
and $\theta(p^-)\theta(p^+)$ multiplied by analytic functions in p^+ and p^-. (100)

From this we conclude that the state $(2\pi/\mu)\xi_2^* \xi_1|0\rangle$ is a normalized, Poincaré-invariant state. With these preliminaries out of the way we can finally discuss the spectrum of the Hamiltonian H.

7. THE SPECTRUM OF THE HAMILTONIAN

Using the results of the previous section, we see that we have the normalized Poincaré-invariant states

$$|n\rangle = ((2\pi/\mu)\xi_2^* \xi_1)^{(|n| + n)/2}((2\pi/\mu)\xi_1^* \xi_2)^{(|n| - n)/2}|0\rangle,$$
$$n = 0, \pm 1, \pm 2, \cdots. \tag{101}$$

Each of these states can be used as a cyclic vacuum with regard to the fields Σ, a_0, and b, where the field a_0 is not allowed to carry zero frequencies. In this way we build up a Fock space G_n of $\Sigma(f)$, $a_0(f_0)$, $b(f)$, where $f \in \mathscr{S}(\mathbb{R}^2)$ and $f_0 \in \mathscr{S}_0(\mathbb{R}^2) \subset \mathscr{S}(\mathbb{R}^2)$ is the space of test functions whose support excludes the origin $p_\mu = 0$. The Hilbert space G of asymptotic states is then the direct sum over the individual Fock spaces G_n:

$$G = \bigoplus_{n=-\infty}^{\infty} G_n. \tag{102}$$

It is now clear that each space G_n contains the same spectrum as H_0 if the fermion term is dropped from H_0. Thus in each space G_n no vestige of the fermions remains. The fermions are confined. Nevertheless, a hint of their existence is manifested by the infinite degeneracy of the spectrum.

Another comment is in order. Since a_0 is a dipole field (except for the Landau gauge, $\alpha = 0$), neither H_0 nor H can be diagonalized.[5,6]

8. RENORMALIZATION

In defining the electromagnetic current in Ref. 1 we used a split-point regularization and gauge invariance. The current was then defined by

$$j_\mu(x) = \lim_{\substack{\epsilon \to 0 \\ \epsilon^2 \neq 0}} \left\{\bar{\phi}(x + \epsilon)\gamma_\mu \phi(x) \right.$$
$$\left. \times \exp\left[-ie\int_0^\epsilon A_\nu(x + \xi)\, d\xi^\nu - \langle\ \rangle_0\right]\right\}. \tag{103}$$

It is, however, possible to maintain gauge invariance by using a different definition. Thus we now consider the current $\bar{j}_\mu$ defined by

$$\bar{j}_\mu(x) = \lim_{\substack{\epsilon \to 0 \\ \epsilon^2 \neq 0}} \left\{\bar{\phi}(x + \epsilon)\gamma_\mu \phi(x) \right.$$
$$\left. \times \exp\left[-ie\int_0^\epsilon V_\nu(x + \xi)\, d\xi^\nu - \langle\ \rangle_0\right]\right\}, \tag{104}$$

where

$$V_\nu(x) = A_\nu(x) + u\partial_\nu \partial\cdot A(x) + v\partial^\lambda F_{\lambda\nu}(x) \tag{105}$$

and u, v are two real parameters. This definition also leads to a viable current. As we now show, the net effect of replacing j_μ by $\bar{j}_\mu$ is to renormalize the Lagrangian (1).

Using the same procedure as in Ref. 1 to triple dot order the terms in (104), we find for small ϵ that

$$\bar{j}_\mu(x) = j_\mu(x) - (eu/2\pi)\,\partial^\mu \partial\cdot A(x) - (ev/2\pi)\,\partial^\lambda F_{\lambda\nu}(x), \tag{106}$$

where we have also made extensive use of the Dirac equation

$$(i\partial\!\!\!/ - eA\!\!\!/)\phi = 0. \tag{107}$$

With the above result we find that the equation of motion for A_ν is

$$(1 + e^2 v/2\pi)\,\partial^\lambda F_{\lambda\mu} + (\alpha + e^2 u/2\pi)\,\partial_\mu \partial\cdot A = e j_\mu. \tag{108}$$

Equations (107) and (108) describe the new equations of motion due to using $\bar{j}$ instead of j. They can be considered to be the equations of motion arising from the renormalized formal Lagrangian

$$\mathscr{L}_R = -\tfrac{1}{4} Z_3 (F_{\mu\nu})^2 - \tfrac{1}{2}\alpha Z_\alpha (\partial\cdot A)^2 + \bar{\phi}(i\partial\!\!\!/ - eA\!\!\!/)\phi \tag{109}$$

with

$$Z_3 = 1 + e^2 v/2\pi, \quad Z_\alpha = 1 + e^2 u/2\pi\alpha. \tag{110}$$

By rescaling the fields we can rewrite this Lagrangian as

$$\mathscr{L}_0 = -\tfrac{1}{4}(F_{\mu\nu})^2 - \tfrac{1}{2}\bar{\alpha}(\partial\cdot A)^2 - \bar{\phi}(i\partial\!\!\!/ - \bar{e}A\!\!\!/)\phi, \tag{111}$$

which is of the same form as the original Lagrangian (1) except that we have replaced α by

$$\bar{\alpha} = \alpha \frac{Z_\alpha}{Z_3} = \alpha \frac{1 + e^2 u/2\pi\alpha}{1 + e^2 v/2\pi} \tag{112}$$

and e by

$$\bar{e} = eZ_3^{-1/2} = e(1 + e^2 v/2\pi)^{-1/2}. \tag{113}$$

The mass arising from $\mathscr{L}_0$ is

$$\bar{m}^2 = \frac{\bar{e}^2}{\pi} = \frac{e^2}{\pi}\left(1 + \frac{e^2 v}{2\pi}\right)^{-1} = m^2\left(1 + \frac{e^2 v}{2\pi}\right)^{-1}. \tag{114}$$

Thus various choices of the parameters u, v lead to equivalent theories.

As a final item we consider the analytic properties of the Wightman functions with respect to the coupling constant.

9. ANALYTICITY IN THE COUPLING CONSTANT

Schwinger's[7] original solution of the Schwinger model was given in terms of perturbation theory. Since then there have been other perturbation theoretic considerations of this model.[8,9] As we now have all the Wightman functions of this model explicitly displayed, it is feasible to examine their

analyticity properties with respect to the coupling constant e.

We begin by considering the $e\to 0$ limit of the various Wightman functions. To accomplish this, we need only consider the two-point function for A_μ, the fermion $2n$-point function, and the mixed three-point function. From Eq. (26) we see that due to the presence of the term $(i\pi/e^2)\,\partial_\mu\partial_\nu \times[\Delta^{(+)} - D^{(+)}]$ the limit $e\to 0$ exists only if we take test functions which vanish for $p_\mu = 0$. In that case we obtain

$$\lim_{e\to 0}\langle 0|A_\lambda(x)A_\nu(0)|0\rangle = \lim_{m\to 0} ig_{\mu\nu}\Delta^{(+)}(m^2, x). \tag{115}$$

On the other hand, using Eqs. (15) and (20) of Ref. 1, we find that

$$\lim_{k\to 0}[\Delta^{(+)}(m^2, x) + (1/\pi i)\ln Z\,] = D^{(+)}(x) \tag{116}$$

so this limit exists.

We next consider the $e\to 0$ limit for the $2n$-point fermion functions given by Eq. (20), namely,

$$W_n(x, y) = Z^{-n}\exp[\mathscr{F}^{(1)}(x, y)]\,w_0^{2n}(x, y).$$

Using Eq. (116), we obtain

$$\lim_{e\to 0} Z^{-n}\exp\mathscr{F}^{+}(x, y)$$
$$= \exp\Bigg\{ -n\ln Z + \ln Z\Bigg[\sum_{i,j=1}^{n}\gamma_{x_i}^5\phi_{y_j}^5 - \sum_{i<j=1}^{n}(\gamma_{x_i}^5\gamma_{x_j}^5 + \gamma_{y_i}^5\gamma_{y_j}^5)\Bigg]\Bigg\}. \tag{117}$$

If we now take the spinor indices of the first $k\leqslant n$ fermion fields to be 1 and the spinor indices of the remaining $n-k$ fermion fields to be 2, then we can evaluate the sums over the γ^5 matrices to get

$$\sum_{i,j=1}^{n}\gamma_{x_i}^5\gamma_{y_j}^5 = k^2 + (n-k)^2 - 2k(n-k) = (n-2k)^2, \tag{118}$$

$$\sum_{i,j=1}^{n}\gamma_{x_i}^5\gamma_{y_j}^5 = \tfrac{1}{2}k(k-1) + \tfrac{1}{2}(n-k)(n-k-1) + k(n-k). \tag{119}$$

Combining these results, Eq. (117) becomes

$$\exp\ln Z\{-n + (n-2k)^2 - k(k-1) - (n-k)(n-k-1) - 2k(n-k)\} = 1. \tag{120}$$

Thus

$$\lim_{e\to 0} W_n(x, y) = w_0^{2n}(x, y). \tag{121}$$

Finally we consider the limit $e\to 0$ for the mixed three-point function given by Eq. (34). To obtain this limit, we must simply consider the limit of the function $G_\mu^{(+)}(x, y)$ given by Eq. (32). Again using Eq. (116), we easily obtain that

$$\lim_{e\to 0} G_\mu^{(+)}(x, y) = 0. \tag{122}$$

Thus the limits of all these Wightman functions exist in the sense of distributions in $\mathscr{S}_0'$ whose test functions are Fourier transforms of functions in $\mathscr{S}$ with their support excluding the origin $p_\mu = 0$. In spite of this, the Wightman functions are not analytic in e. To see this, consider the fermion two-point function

$$\langle 0|\phi(x)\phi^*(y)|0\rangle = Z^{-1}\exp e^2\{-(i/\alpha)I^{(+)}(x-y) - (i\pi/e^2) \times[\Delta^{(+)}(x-y) - D^{(+)}(x-y)]\}w_0^2(x, y), \tag{123}$$

where we have used that $\gamma_x^5\gamma_y^5 = 1$ for this case.

Now for small $m^2 = e^2/\pi$ we have

$$\Delta^{(+)}(m, x) = -\frac{1}{4}H^{(1)}(im(-x^2 + i\epsilon x^0)^{1/2})$$
$$\equiv -\frac{1}{4}H^{(1)}(y)$$
$$= -\frac{1}{4}\Bigg\{J_0(y)\Bigg[1 + \frac{2i}{\pi}\Bigg(\gamma + \ln\frac{y}{2}\Bigg)\Bigg] - \frac{2i}{\pi} \times\sum_{k=0}^{\infty}\frac{(-1)^n(y/2)^{2k}}{(k!)^2}\Bigg(1 + \frac{1}{2} + \cdots + \frac{1}{k}\Bigg)\Bigg\}. \tag{124}$$

This clearly shows that

$$\exp\{-i\pi[\Delta^{(+)}(m, x) + (1/i\pi)\ln Z\,]\} \underset{m\to 0}{\simeq} \exp[-\tfrac{1}{2}J_0(y)\ln(y/2) - \ln Z\,]$$

and has a cut in m.

Thus we find that the coupling constant e is not a suitable expansion parameter around zero. In spite of this, when such an expansion is summed, the correct analytic properties in e are obtained.

10. CONCLUSIONS

We have studied certain properties of the Schwinger model. In particular, we have obtained all the Wightman functions for this model. We have also studied the algebra of fields and representations of this algebra. A particularly interesting object of this model turns out to be the Hamiltonian. It does not consist simply of the Hamiltonian H_0 for the building block fields, although this one does yield the correct time evolution for the algebra of fields. The full Hamiltonian H reflects the "confinement" of the quarks in that the only vestige of the fermions that remains are zero-energy (actually Poincaré-invariant) states in its spectrum.

We also briefly discuss renormalization of the theory and analyticity of the amplitudes in terms of the coupling constant.

ACKNOWLEDGMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. One of us (R. F.) would also like to thank the theoretical Physics Institute of the University of Alberta for support during his visit there.

APPENDIX A: AVERAGING OF POINT SPLITTING

In computing the energy–momentum tensor regularized by point splitting, one obtains expressions of the form

$\eta_{\mu_1}\cdots\eta_{\mu_n}$ in the point splitting parameter η_μ. We prescribe an averaging method over "all directions" of n_μ.

Since the Lorentz group is noncompact, we go to the Euclidean region $\eta_0 \to i\eta_0$ to perform our averaging. In this case keeping the length of η_μ fixed is no problem.

Thus we define the average in the Euclidean region by

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_n}} = \frac{(i)^{\delta_{\mu_1 0}+\cdots+\delta_{\mu_n 0}}}{\pi}\int d^2\eta\,\delta(\eta^2-a^2)\,\eta_{\mu_1}\cdots\eta_{\mu_n}. \tag{A1}$$

Letting $\eta_0 = R\cos\theta$, $\eta_1 = R\sin\theta$, the integral becomes

$$I_n = \int_0^{2\pi} d\theta\,\frac{1}{2}\int R^n\,dR^2\,\delta(R^2-a^2)\,(\cos\theta)^k(\sin\theta)^{(n-k)}, \tag{A2}$$

where we have assumed that k of the μ_i values have the value 0 and the rest have the value 1. It is easy to see that unless n and k are even I_n vanishes. For n, k even we obtain

$$I_n = \frac{a^n}{2}\cdot 2\pi\,\frac{(k-1)!!(n-k-1)!!}{n!!}. \tag{A3}$$

These results now immediately yield:

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_{2n}}}_{\text{Euclidean}} = \frac{a^{2n}(i)^{\delta_{\mu_1 0}+\cdots+\delta_{\mu_{2n}0}}}{(2n)!!}\sum_{\substack{\text{partitions}\\ \text{in } n \text{ pairs}}}\delta_{\mu_i\mu_j}\cdots\delta_{\mu_k\mu_l}, \tag{A4}$$

which in Minkowsky space becomes

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_{2n}}} = \frac{(\eta^2)^n}{(2n)!!} + \sum_{\text{partitions}} g_{\mu_i\mu_j}\cdots g_{\mu_k\mu_l}, \tag{A5}$$

where the sum is over the partitions of the $2n$ indices into pairs (μ_i, μ_j) with $i<j$. Furthermore, we immediately find that the "average" of an odd product of η_μ's vanishes.

APPENDIX B: EQUAL-TIME COMMUTATORS FOR BUILDING-BLOCK FIELDS

Using the various commutators for the building-block fields, one easily finds the following useful equal-time (anti-) commutation relations:

$$[\dot a(x), b(0)]_{x^0=0} = -i/\alpha\delta(x^1), \tag{B1}$$

$$[\dot a(x), a(0)]_{x^0=0} = i(\beta^2 + 2\beta/m)\delta(x^1), \tag{B2}$$

$$[\dot b(x), a(0)]_{x^0=0} = -(i/\alpha)\delta(x^1), \tag{B3}$$

$$[\Sigma(x), \Sigma(0)]_{x^0=0} = -i\delta(x^1), \tag{B4}$$

$$[\dot\rho(x), \rho(0)]_{x^0=0} = -i\delta(x^1), \tag{B5}$$

$$[\partial^1\sigma(x), \rho(0)]_{x^0=0} = i\delta(x^1), \tag{B6}$$

$$[\partial^1\rho(x), \sigma(0)]_{x^0=0} = i\delta(x^1), \tag{B7}$$

$$[\dot\sigma(x), \sigma(0)]_{x^0=0} = -i\delta(x^1), \tag{B8}$$

$$\{\psi(x), \bar\psi(0)\}_{x^0=0} = \gamma^0\delta(x^1), \tag{B9}$$

$$\{\bar\psi(x), \psi(0)\}_{x^0=0} = \gamma^0\delta(x^1), \tag{B10}$$

$$[\dot\rho(x), \psi(y)]_{x^0=y^0} = -\sqrt{\pi}\psi(y)\delta(x^1), \tag{B11}$$

$$[\partial^1\rho(x), \psi(y)]_{x^0=y^0} = -\sqrt{\pi}\,\gamma^5\psi(y)\delta(x^1), \tag{B12}$$

$$[\dot\sigma(x), \psi(y)]_{x^0=y^0} = \sqrt{\pi}\,\gamma^5\psi(y)\delta(x^1), \tag{B13}$$

$$[\partial^1\sigma(x), \psi(y)]_{x^0=y^0} = \sqrt{\pi}\,\psi(y)\delta(x^1). \tag{B14}$$

Moreover, we find that

$$[:\bar\psi\gamma^1\partial^1\psi:(y), \rho(x)]_{y^0=x^0} = -\dot\rho(x)\delta(x^1-y^1) \tag{B15}$$

and

$$[:\bar\psi\gamma^1\partial^1\psi:(y), \sigma(x)]_{y^0=x^0} = -\dot\sigma(x)\delta(x^1-y^1). \tag{B16}$$

APPENDIX C: THE LANDAU GAUGE

We briefly consider the Landau gauge here. It is obtained as the $\alpha\to\infty$ limit of our solutions if one keeps

$$b_0(x) \equiv \alpha b(x) \tag{C1}$$

fixed.

It then follows from Eq. (79) that

$$a_0(x) = a(x) - b_0(x)/m^2 + (\beta + 1/m)\rho(x) \tag{C2}$$

and

$$\langle 0|a_0(x)a_0(0)|0\rangle = (i/m^2)D^{(+)}(x), \tag{C3}$$

whereas

$$\langle 0|a_0(x)b_0(0)|0\rangle = -iD^{(+)}(x) \tag{C4}$$

and

$$\langle 0|b_0(x)b_0(0)|0\rangle = 0, \tag{C5}$$

$$\langle 0|a_0(x)\tilde b_0(0)|0\rangle = -i\tilde D^{(+)}(x). \tag{C6}$$

Moreover, both a_0 and b_0 are scalar fields:

$$\Box a_0 = 0, \quad \Box b_0 = 0. \tag{C7}$$

The fields ϕ and A_μ are now given by

$$\phi(x) = :\exp\Big\{-ie\Big[a_0(x) + \frac{b_0(x)}{m^2} + \frac{1}{m}\Big(\Sigma(x) - \frac{\tilde b_0(x)}{m}\Big)\gamma_x^5\Big]:\zeta(x)\Big\}, \tag{C8}$$

$$A_\mu = \partial_\mu a_0 + (1/m)\epsilon_{\mu\nu}\partial^\nu\Sigma, \tag{C9}$$

where ζ is still given by Eq. (80). The content of these equations is clarified if instead of the two massless scalar fields a_0, b_0 we introduce two commuting *massless* scalar fields

$$a_1 = ma_0 + (1/m)b_0, \quad a_2 = ma_0, \tag{C10}$$

$$b_0 = m(a_1 - a_2), \quad a_0 = (1/m)a_2. \tag{C11}$$

We then find

$$\langle 0|a_1(x)a_1(0)|0\rangle = -iD^{(+)}(x), \tag{C12}$$

$$\langle 0|a_1(x)a_2(0)|0\rangle = 0, \tag{C13}$$

$$\langle 0|a_2(x)a_2(0)|0\rangle = +iD^{(+)}(x). \tag{C14}$$

Thus the field a_2 carries a *negative* norm.

[1] A. Z. Capri and R. Ferrari, Nuovo Cimento A **62**, 273 (1981).
[2] B. Klaiber, *Boulder Lectures in Theoretical Physics* (Gordon and Breach, New York, 1967), Vol. XA, p. 141.
[3] N. Nakanishi, Prog. Theor. Phys. **57**, 1025 (1977).
[4] C. M. Sommerfield, Ann. Phys. (N. Y.) **26**, 1 (1964).
[5] R. Ferrari, Nuovo Cimento A **19**, 204 (1974).
[6] A. Z. Capri, G. Grübl, and R. Kobes, Ann. Phys. (N. Y.) **147**, 140 (1983).
[7] J. Schwinger, Phys. Rev. **128**, 2425 (1962).
[8] P. Becher and H. Joos, "(1 + 1)-Dimensional quantum electrodynamics," DESY Preprint 77/43, 1977.
[9] I. O. Stamatescu and T. T. Wu, Nucl. Phys. B **143**, 503 (1978).

A novel mass-eigenvalue problem for spinors in deSitter space

Edward H. Kerner
Sharp Physics Laboratory, University of Delaware, Newark, Delaware 19711

(Received 6 May 1983; accepted for publication 26 August 1983)

It is shown that an unambiguous quantum theory of spinors in positively curved deSitter space, based on distinguished coordinates in a Hamiltonian framework, leads to a set of spinors corresponding to unsharp energy but sharp mass defined in a family of novel eigenvalue problems. An example is given in which partly real and partly complex discrete mass spectra come forth.

PACS numbers: 11.10.Qr, 04.90. + e

Spinors in spaces of constant curvature [deSitter spaces of O(3,2) or O(4,1) symmetry] have received continuing attention[1] for nearly fifty years. Their structure is of interest not only in its own right since deSitter spaces are the physically distinguished ones having maximal (tenfold) symmetry, but also because they are local osculating spaces[2] (rather than mere tangent spaces) to more generally curved Riemann spaces, attaining thereby a prototypical role. Further, they form background spaces for supersymmetry,[3] and have been broached[4] as closed up "microuniverses" for considering particle confinement at a basic geometrical level.

In the present paper an unusual family of eigenvalue problems is brought out for the mass of a spinning particle running along a geodesic of O(3,2) deSitter space. This results from a well-set and essentially unique Hamiltonian formulation of the motion developed in recent years,[5] in contrast to the formal spinor theories usually invoked.[1]

In the latter, governed by general convariance considerations, Klein–Gordon equations are typically factorized to curved-space Dirac equations $(\gamma^{\mu}(x)\nabla_{\mu} + m)\psi = 0$ as a matter of formal prescription (∇_{μ} = covariant derivative). The coordinates remain ambiguous, and of course commutation rules are renounced. The Hamitonian formulation, on the other hand, relies on distinguished coordinates and proceeds through clear commutation rules to a quite unambiguous statement of quantum theory. The basis here is a specialized subgroup of the projective transformations $x_i' = \Lambda_i(x,a)/\Delta(x,a) \equiv \Gamma_i(x)$, with Λ_i and Δ inhomogeneous linear functions of space Cartesians $x_1,x_2,x_3 = \mathbf{r}$ and time $x_0 = t$, and a = a universal length. These are isomorphic to the deSitter group of pseudorotations O(3,2) in the five-space of homogeneous coordinates $X_i, U(x_i \equiv X_i/U)$. What is notable is that x' and x are in the relationship of coordinates of inertial frames, since $d^2\mathbf{r}'/dt'^2 = 0$ is sent into $d^2\mathbf{r}/dt^2 = 0$ and conversely, making these coordinates clearly distinguished above all others. While the appropriate invariant line element indeed describes constant curvature $1/a^2$, the geodesics one and all are the global free-particle motions $d^2\mathbf{r}/dt^2 = 0$. Given this order of simplicity, general covariance is rendered irrelevant, and only the automorphism of space-time under $x' = \Gamma(x)$ is consequential, as with the automorphism of Minkowski space under the Poincaré group. Coordinate ambiguities and equivocal quantization recipes may then be set aside, and instead the usual commutation rules $(x_i, p_j^0) = i\hbar\delta_{ij}$, etc. $(i,j = 1,2,3)$ tenably introduced as the primary physical hypothesis for the quantum dynamics of a free particle, in accord with all physical experience.

Useful coordinate transformations can now (*post* settlement of the physical basis) be performed, such as $\boldsymbol{\rho}(\mathbf{r},t)$ and $\tau(t)$ described earlier,[5] that rephrases the straights $d^2\mathbf{r}/dt^2 = 0$ as the harmonic-oscillator geodesics $d^2\boldsymbol{\rho}/d\tau^2 + (c^2/a^2)\boldsymbol{\rho} = 0$ otherwise familiar in deSitter space, and that gives a ladder spectrum of Klein–Gordon energy eigenvalues. The further transformation $\mathbf{R} = \boldsymbol{\rho}/(1 - \rho^2/a^2)^{1/2}$ brings the Hamiltonian-squared

$$\frac{H^2}{c^2} = \left[\mathbf{P}^2 + \frac{\mathbf{L}^2}{a^2} + \frac{\hbar^2}{a^2}\right] + \kappa^2\frac{\hbar^2}{a^2}\left[1 + \frac{\mathbf{R}^2}{a^2}\right] \equiv H_1^2 + \kappa^2 H_2^2, \tag{1}$$

$$\kappa^2 = m^2c^2a^2/\hbar^2 - \tfrac{1}{4}, \quad \mathbf{P} \equiv \tfrac{1}{2}(\hat{I} + \mathbf{RR}/a^2)\cdot\mathbf{P}_c + \text{h.c.},$$

where $\mathbf{P}_c$ is canonical mate $-i\hbar\nabla_R$ to $\mathbf{R}$, and $\mathbf{L}$ is $\mathbf{R}\times\mathbf{P}$, with $\hat{I}$ the unit dyadic.

This reduction forces into particularly clear view the issue of linearization to determine H upon the primary physical basis, an issue distinct from generally covariant factorization of $\nabla^{\mu}\nabla_{\mu} + m^2$. As has been remarked,[6] there does not exist any ordinary matrix squareroot of $H_1^2 + \kappa^2 H_2^2$ in Dirac matrices or otherwise (except for $\kappa = 0$). Since this point is central to any consideration of spinor theory on a Hamiltonian base, the proof will be briefly reviewed.

Taking $\hbar,c,a = 1$ from here on, the one-dimensional form of Eq. (1) already reveals the difficulty:

$$H^2 = P^2 + \kappa^2(1 + X^2),$$

(where both the terms $\mathbf{L}^2/a^2$ and $\hbar^2/a^2$ are to be dropped in one dimension). If H is $F(x)P + G(X)$, it is then required that

$$F^2 = 1,$$

$$FG + GF = iFF',$$

$$G^2 - iFG' = \kappa^2(1 + X^2),$$

be identically satisfied in X, where F' means $(1 + X^2)dF/dX$ and similarly for G'. Multiply the second, right and left, by F. This brings $FF' = F'F$, while the first states that $FF' + F'F = 0$. Hence $FF' = 0 = F'F$, so that $F' = 0$ and $FG + GF = 0$. Now multiply the third, right and left, by F, producing $FG' = G'F$. But $(FG + GF)' = FG' + G'F = 0$, whence $G'F = 0 = FG'$. Consequently $G' = 0$, and then G = const, cannot satisfy the third (except for $\kappa = 0$).

In short, while H_1^2 and H_2^2 are separately Dirac linearizable,[5] for example as

$$H_1 = \boldsymbol{\alpha}\cdot\mathbf{P} - \boldsymbol{\sigma}\cdot\mathbf{L} - 1,$$

$$H_2 = \beta + i\beta\boldsymbol{\alpha}\cdot\mathbf{R}, \tag{2}$$

 0022-2488/84/010150-04$02.50

with standard Dirac matrices β,α,σ, the pieces H_1 and H_2 are *incompatible* in that they cannot, in general, be brought together to give a general overall linear Hamiltonian. The choices for H_1 and H_2 above are not unique but are here selected for simplicity. (A second possibility for H_1 is $\alpha\cdot\mathbf{P}+\alpha\cdot\mathbf{L}-i\gamma_5$, while the roots of $1+R^2$ for H_2 are very numerous; but in all cases a single general Hamiltonian is ruled out.)

To hold to the Hamiltonian framework, and accord to the Hamiltonian its master dynamical role of generator of time (τ) translations, is nevertheless achievable (notwithstanding the incompatibility of H_1 and H_2), provided the spinors to be considered are suitably restricted, and as well the value of the mass parameter $\kappa=(m^2-\frac{1}{4})^{1/2}$.

Clearly, if ψ is a spinor such that

$$H_1\psi=\lambda H_2\psi, \tag{3}$$

then for these spinors an overall linearization of H becomes possible,

$$i\frac{\partial\psi}{\partial\tau}=H\psi=(\alpha_1H_1+\alpha_2H_2)\psi, \tag{4}$$

($\lambda,\alpha_1,\alpha_2$ numerical parameters) since

$$H^2\psi=[\alpha_1^2H_1^2+\alpha_1\alpha_2(H_1H_2+H_2H_1)+\alpha_2^2H_2^2]\psi$$
$$=[(\alpha_1^2+\alpha_1\alpha_2/\lambda)H_1^2+(\alpha_2^2+\lambda\alpha_1\alpha_2)H_2^2]\psi.$$

This requires only that

$$\alpha_1^2+\alpha_1\alpha_2/\lambda=1,\quad \alpha_2^2+\lambda\alpha_1\alpha_2=\kappa^2,$$

or that

$$\alpha_1=(1+\kappa^2/\lambda^2)^{-1/2},\quad \alpha_2=(\kappa^2/\lambda)(1+\kappa^2/\lambda^2)^{-1/2},$$

bringing Eq. (4) to

$$i\frac{\partial\psi}{\partial\tau}=\left(1+\frac{\kappa^2}{\lambda^2}\right)^{1/2}H_1\psi \tag{5}$$
$$=\lambda\left(1+\frac{\kappa^2}{\lambda^2}\right)^{1/2}H_2\psi. \tag{6}$$

If ϕ is some initial spinor, one gets [$\zeta\equiv(1+\kappa^2/\lambda^2)^{1/2}$]

$$\psi=[\exp(-i\zeta H_1\tau)\phi=[\exp(-i\lambda\zeta H_2\tau)]\phi,$$

so that this initial state is constrained to satisfy

$$H_1\phi=\lambda H_2\phi. \tag{7}$$

Stationary states are here ruled out.

As will be shown below, Eq. (7) does not allow arbitrary λ or arbitrary ϕ; rather a discrete spectrum of eigenvalues λ_j and eigenstates ϕ_j is demanded. But then in Eqs. (5) and (6) the operators $(1+\kappa^2/\lambda_j^2)^{1/2}H_1$ or $\lambda_j(1+\kappa^2/\lambda_j^2)^{1/2}H_2$ are not uniquely valued [i.e., are not independent of the index j labeling the eigensolutions of Eq. (7)] unless κ is restricted. Taking uniquely valued spinor wave equations as a basic requirement, two mutually exclusive restrictions on κ stand forth, which may be called cases (A) and (B). These correspond to

$$(1+\kappa^2/\lambda_j^2)^{1/2}=\beta_1 \qquad \text{(A)}$$

or

$$\lambda_j(1+\kappa^2/\lambda_j^2)^{1/2}=\beta_2, \qquad \text{(B)}$$

where β_1, β_2 are arbitrary real numbers independent of the label j. Not both of (A) and (B) can be allowed simultaneously since $\beta_2/\beta_1=\lambda_j$ is ruled out. Then

$$\kappa_j^2=(\beta_1^2-1)\lambda_j^2 \quad \text{(A)}$$

or (8)

$$\kappa_j^2=\beta_2^2-\lambda_j^2, \qquad \text{(B)},$$

prescribe the allowed mass spectra, while the uniquely valued spinor wave equations are

$$i\frac{\partial\psi}{\partial\tau}=\beta_1H_1\psi, \qquad \text{(A)}$$

or

$$i\frac{\partial\psi}{\partial\tau}=\beta_2H_2\psi, \qquad \text{(B)}$$

with β_iH_i remaining Hermitian when H_i are Hermitian (β_1,β_2 may be absorbed into scale changes in τ if desired). It is easily demonstrated that $\mathbf{J}=\mathbf{L}+\frac{1}{2}\sigma$ commutes with both H_1, H_2 of Eq. (2), so that ψ may be an eigenstate of total angular momentum, but it clearly cannot be an eigenstate of energy (either H_1 or H_2).

We may summarize as follows: *Within the Hamiltonian framework in deSitter space, spinors exist which are not eigenstates of the Hamiltonian but rather are eigenstates of a "mass-generating operator" $H_2^{-1}H_1$* [Eq. (7)] *whose eigenvalues prescribe a family of allowed masses (Eq. 8) and whose elements H_1, H_2 are Dirac square roots of well-defined operators within that framework.* In a word, these particular states are unsharp in energy but sharp in mass. To the extent that one may regard the parameters β_1, β_2 as running freely over their real values, the mass spectra are of the nature of bands, with individual bands labeled discretely according to the eigenvalues of the $H_2^{-1}H_1$ operator.

A further perspective on the reduction given above is sketched in the Appendix, where a novel square root process[6] for $H_1^2+\kappa^2H_2^2$ in total is reviewed, and the case where $\lambda=\kappa$ is particularly obtained.

Turning to the eigenvalue problem of λ, Eq. (7), we may use H_1 and H_2 from Eq. (2) as an example. In view of the many possible choices for H_i noted before, this will be understood to be primarily illustrative rather than exhaustive or definitive, demonstrating the principal point that λ has a discrete spectrum. Since the H_1, H_2 of Eq. (2) do not commute, the mass generator $H_2^{-1}H_1$ in $H_2^{-1}H_1\phi=\lambda\phi$ is not Hermitian, so λ cannot be expected to have a completely real spectrum in the present example.

Eq. (7) is readily analyzed upon recognizing certain structural similarities to the classical Dirac–Coulomb problem as set forth particularly by Foldy.[7] First it is convenient to return to the harmonic-oscillator coordinate $\boldsymbol{\rho}$ or ρ,θ,ϕ in polar coordinates ($0\leqslant\rho\leqslant1$) with corresponding momentum $\mathbf{p}=-i\nabla_\rho$. Then employing Foldy's operators

$$\hat{k}=\beta(\sigma\cdot\mathbf{L}+1),$$
$$\alpha_\rho=\alpha\cdot\boldsymbol{\rho}/\rho,$$
$$P_\rho=(1/\rho)(\boldsymbol{\rho}\cdot\mathbf{p}-i),$$

the operators H_1, H_2 are

$$H_1=(1-\rho^2)^{1/2}(\alpha_\rho\,p_\rho+(i/\rho)\alpha_\rho\,\beta\hat{k})$$
$$+\tfrac{1}{2}i\rho\alpha_\rho/(1-\rho^2)^{1/2}-\beta\hat{k},$$
$$H_2=\beta+i\beta\alpha_\rho\,\rho/(1-\rho^2)^{1/2}.$$

The operators $\beta, \hat{k}, L_2, J_z$ are intercommuting and their common eigenvector, which depends only on θ and ϕ, may be designated as ξ, belonging respectively to the eigenvalues $1, k, l(l+1), m_j$. A second angular spin function $\eta \equiv i\alpha_\rho \xi$ is also an eigenvector of $\hat{k}$ and J_z with the same eigenvalues k and m_j as ξ [though it is not an eigenvector of L^2 belonging to $l(l+1)$]. Since η is an eigenvector of β belonging to the eigenvalues -1, it is orthogonal to ξ. Hence when one introduces

$$\phi = (f(\rho)/\rho)\xi + (g(\rho)/\rho)\eta,$$

into $H_1\phi = \lambda H_2\phi$, one obtains terms only in ξ and η, and thence by their orthogonality, the pair of coupled radial equations

$$\frac{df}{d\rho} + \left(-\frac{k}{\rho} - \frac{1}{2}\frac{\rho}{1-\rho^2} - \lambda\frac{\rho}{1-\rho^2}\right)f + \frac{k+\lambda}{(1-\rho^2)^{1/2}}g = 0,$$

$$\frac{dg}{d\rho} + \left(\frac{k}{\rho} - \frac{1}{2}\frac{\rho}{1-\rho^2} + \lambda\frac{\rho}{1-\rho^2}\right)g + \frac{k+\lambda}{(1-\rho^2)^{1/2}}f = 0. \tag{9}$$

Here k is an eigenvalue of $\hat{k}$, namely $k^2 = (j+\frac{1}{2})^2$ with $j = \frac{1}{2}, \frac{3}{2}, \ldots$, that is, $k = \pm 1, \pm 2, \ldots$ or $|k| \equiv s = 1,2,\ldots$.

The normalization of ϕ is defined by

$$\int_0^1 \frac{|f|^2 + |g|^2}{\rho^2}\frac{\rho^2 d\rho}{(1-\rho^2)^{5/2}} = 1, \tag{10}$$

when ξ and η are normalized according to

$$\int \xi^+\xi \sin\theta\, d\theta\, d\phi = 1 = \int \eta^+\eta \sin\theta\, d\theta\, d\phi,$$

where the factor $(1-\rho^2)^{-5/2}$ comes from the invariant line element in ρ,τ variables that prescribe the invariant volume element $(1-\rho^2)^{-5/2}d\rho\, d\tau$ in deSitter space. Consequently f and g must be regular at $\rho = 0$ and vanish sufficiently fast at $\rho = 1$.

One very simple solution to Eqs. (9) stands out at once in the case $k + \lambda = 0$,

$$f = \rho^k(1-\rho^2)^{-1/4-(1/2)\lambda},$$
$$g = \rho^{-k}(1-\rho^2)^{-1/4+(1/2)\lambda}.$$

Not both of these may be retained, but only

$$f = 0 \quad g = \rho^s(1-\rho^2)^{(1/2)s-1/4}$$

or

$$g = 0 \quad f = \rho^s(1-\rho^2)^{(1/2)s-1/4}$$

with eigenvalues

$$\lambda^2(s) = s^2 = 9,16,\ldots$$

for $s = 3,4,\ldots$ in view of Eq. (10).

Proceeding to the general situation, write

$$f = (1-\rho^2)^{1/4}F, \quad g = (1-\rho^2)^{-1/4}G$$

to get rid of roots of $1-\rho^2$,

$$F' - \left(\frac{k}{\rho} + (1+\lambda)\frac{\rho}{1-\rho^2}\right)F + \frac{k+\lambda}{1-\rho^2}G = 0,$$

$$G' + \left(\frac{k}{\rho} + \lambda\frac{\rho}{1-\rho^2}\right)G + (k+\lambda)F = 0,$$

and decouple to obtain a second-order equation in G alone,

$$G'' - \frac{\rho}{1-\rho^2}G' + \left[\frac{-k(k+1)}{\rho^2} - \frac{(k+\lambda)^2 + 2k\lambda + k}{1-\rho^2} + \frac{\lambda - \lambda^2\rho^2}{(1-\rho^2)^2}\right]G = 0.$$

Now extract the characteristic behavior at $\rho = 0$ and $\rho^2 = 1$ through

$$G = \rho^\alpha(1-\rho^2)^\beta S$$

to obtain the indicial roots

$$\alpha = -k,\ k+1,$$
$$\beta = \tfrac{1}{2}\lambda,\ \tfrac{1}{2}(1-\lambda),$$

with S satisfying the differential equation of essentially hypergeometric type

$$S'' + \left(\frac{2\alpha}{\rho} - \frac{(1+4\beta)\rho}{1-\rho^2}\right)S' - \frac{\gamma}{1-\rho^2}S = 0$$
$$\gamma \equiv (k+\lambda)^2 + 2k\lambda + k + \alpha + 2\beta + 4\alpha\beta - \lambda.$$

In the customary way, the series solution $S = \Sigma a_\nu \rho^\nu$ produces the recursion

$$\frac{a_{\nu+2}}{a_\nu} = \frac{(\nu+\alpha+2\beta+q)(\nu+\alpha+2\beta-q)}{(\nu+2)(\nu+1+2\alpha)}, \tag{11}$$

$$q^2 = (\alpha+2\beta)^2 - \gamma = -4k\lambda.$$

The even and odd solutions here are then

$$S_e = {}_2F_1[(\alpha+2\beta+q)/2,(\alpha+2\beta-q)/2;\alpha+\tfrac{1}{2};\rho^2),$$
$$S_o = \rho_3F_2(1,(1+\alpha+2\beta+q)/2,(1+\alpha+2\beta-q)/2;$$
$$\tfrac{3}{2},1+\alpha;\rho^2).$$

The recursion relation Eq. (11) shows that S behaves like $(1-\rho^2)^{1/2-2\beta}$ near $\rho = 1$. This overwhelms the factor $(1-\rho^2)^\beta$ in G when at the outset $\mathrm{Re}(\beta)$ is taken as positive to ensure that G vanishes appropriately at $\rho = 1$. Hence the S series must be broken off in a polynomial,

$$n + \alpha + 2\beta \pm q = 0,$$
$$n = 0,1,2,\ldots .$$

Therefore when $\alpha = -k$ (k negative) $= s$ and $\beta = \lambda/2$ one obtains the λ spectrum

$$\lambda^2 + 2\lambda(n-s) + (n+s)^2 = 0,$$
$$\lambda(s,n) = s - n \pm 2i\sqrt{sn},$$

requiring $s \geqslant n+3$ for satisfactory $\mathrm{Re}(\beta) > 0$ [the root $\beta = (1-\lambda)/2$ of the indicial equation is ruled out].

Similarly, when $\alpha = k+1$ (k positive) and $\beta = (1-\lambda)/2$, the λ spectrum is

$$\lambda(k,n) = n - k + 2 \pm 2i\sqrt{k(n+2)},$$

with $k \geqslant n+4$ for suitable $\mathrm{Re}(\beta)$ (the indicial root $\beta = \lambda/2$ being ruled out here). The case $\alpha = 0$ ($k = -1$) is not allowed.

This concludes the illustration of how the mass generator $H_2^{-1}H_1$ eventuates in a spectrum of eigenvalues $\lambda(s)$,

$\lambda(s,n)$, $\lambda(k,n)$ and corresponding spinors belonging to sharp masses. In the case of the real eigenvalue $\lambda(s)$, the mass spectra according to Eq. (8) are

$$m_j^2 = (\beta_1^2 - 1)(j+\tfrac{1}{2})^2 + \tfrac{1}{4} \qquad \text{(A)},$$

or

$$m_j^2 = \beta_2^2 + \tfrac{1}{4} - (j+\tfrac{1}{2})^2, \qquad \text{(B)}$$
$$j = \tfrac{5}{2},\tfrac{7}{2},\ldots,$$

where (A) describes an infinite real discrete spectrum for $\beta_1^2 > 1$ and a finite real spectrum for $\beta_1^2 - 1$ small and negative; while (B) describes a finite real spectrum for adequately large β_2. Corresponding mass *bands* are defined when β_1, β_2 are allowed to range freely. The complex eigenvalues $\lambda(s,n)$, $\lambda(k,n)$ of course do not admit ready interpretation [though perhaps hinting to a later discrete spectrum of (composite) particle decay times accompanying discrete masses]. Indeed the meaning of *mass* altogether in such totally closed up or 'interior' geometry as that of O(3,2) remains in issue until that geometry is clarified as a locale of an 'exterior' large-scale geometry suited to physical observation.

ACKNOWLEDGMENT

My thanks go to the U.S. Department of Energy for its partial support of this work.

APPENDIX

The fundamental eigenvalue problem $H_1\phi = \lambda H_2\phi$ of the present work also occurs, for $\lambda = \kappa$, upon introducing[6] a novel square root process for

$$P_\tau^2\psi = (H_1^2 + \kappa^2 H_2^2)\psi,$$

($P_\tau = i\partial/\partial\tau$). Namely the linearization

$$I \otimes N_0\, P_\tau\psi = (H_1 \otimes N_1 + \kappa H_2 \otimes N_2)\psi$$

is feasible in that iteration produces

$$(I \otimes N_0)^2\, P_\tau^2\psi = (I \otimes N_0)^2(H_1^2 + \kappa^2 H_2^2)\psi,$$

when $N_0^2 = N_1^2 = N_2^2$, and (to overcome the incompatibility of H_1, H_2) $N_1N_2 = 0 = N_2N_1$. That is, the N_i are suitable *singular* matrices which are nilpotent like $N_i^3 = 0$. The analysis shows[6] that N_i must be at least 4×4 and then of typical structure $N_i = n_i T$ (upon enforcing n_i Hermitian and T unitary)

$$n_1 = \begin{array}{|cccc} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot \\ \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{array} = 0 \oplus \lambda_i,$$

$$n_2 = \begin{array}{|cccc} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & 1 & \cdot \end{array} = 0 \oplus \lambda_6.$$

Here dots stand for zeroes, $\otimes$ for direct product, $\oplus$ for direct sum, and $N_0 = (N_1 + N_2)/\sqrt{2}$ while $T_{13}, T_{22}, T_{31}, T_{41} = 1$ ($T_{ij} = 0$ otherwise) and λ_1, λ_6 are two of the conventional generators of SU(3) [other SU(3) generators and other unitary T are possible, as are higher-dimensional $n_i = 0 \oplus$ SU(N) for all $N \geqslant 3$ but $N < 3$ is ruled out]. In short unitary spin comes forth quite directly in a fusion with Dirac spin, and here is not an ad hoc appendage.

The unitary transform $\Phi = I \otimes T\psi$ brings the linearized wave equation

$$I \otimes n_0\, P_\tau\Phi = (H_1 \otimes n_1 + H_2 \otimes n_2)\Phi, \qquad (12)$$

with $n_0 = (n_1 + n_2)/\sqrt{2}$. Introducing Φ as col$(\Phi_a, \Phi_b, \Phi_c, \Phi_d)$ with indices tied to the n-matrices, Φ_a of course falls aside, leaving Eq. (12) as

$$i\Phi_c' = H_1\Phi_c,$$
$$i(\Phi_b' + \Phi_d') = H_1\Phi_b + \kappa H_2\Phi_d,$$
$$i\Phi_c' = \kappa H_2\Phi_c,$$

where $\Phi' = \partial\Phi/\partial\tau'$ ($\tau' = \tau\sqrt{2}$). It is sufficient to span unitary-spin space, to take $i\Phi_b' = H_1\Phi_b$ and $i\Phi_d' = \kappa H_2\Phi_d$, leaving

$$\Phi_c = [\exp(-iH_1\tau')]\phi = [\exp(-i\kappa H_2\tau')]\phi,$$

and requiring

$$H_1\phi = \kappa H_2\phi, \qquad (13)$$

as in Eq. (7).

Hence in the present spin $\otimes$ unitary-spin scheme the mass parameter κ is directly fixed as eigenvalue of Eq. (13), for example $\kappa = s = |j + \tfrac{1}{2}|$ for $j = \tfrac{5}{2},\tfrac{7}{2},\ldots$ as before. In the latter case,

$$m_j^2 = j(j+1) + \tfrac{1}{2} = \tfrac{37}{4},\tfrac{65}{4},\ldots .$$

This result resembles that of Barut and Böhm[8] for a so-called deSitter "rotator," which, however, stems not from O(3,2) but from O(4,1), and refers not to a particle but to a composite system.

[1] P. A. M. Dirac, Ann. of Math. **36**, 657 (1935); F. Gürsey and T. D. Lee, Proc. Natl. Acad. Sci. **49**, 179 (1963); O. Nachtmann, Comm. Math. Phys. **6**, 1 (1967); G. Börner and H. P. Dürr, Nuovo Cimento A **64**, 669 (1969); M. S. Drew, Ann. Phys. (N.Y.) **103**, 469 (1977).
[2] M. D. Maia, J. Math. Phys. **22**, 538 (1981).
[3] S. Deser and B. Zumino, Phys. Rev. Lett. **38**, 1433 (1977).
[4] A. Salam and J. Strathdee, Phys. Rev. D **18**, 4596 (1978); C. Sivaram and K. P. Sinha, Phys. Rep. **51**, 111 (1979).
[5] E. H. Kerner, Phys. Rev. D **22**, 280 (1980).
[6] E. H. Kerner, Phys. Rev. D **26**, 390 (1982).
[7] L. L. Foldy, in *Quantum Theory III*, edited by D. R. Bates (Academic, New York, 1962), p. 29.
[8] A. O. Barut and A. Böhm, Phys. Rev. **139**, B1107 (1965).

Vortex properties in first- and second-order formulations of abelian gauge theories

John van der Hoek
Department of Pure Mathematics, The University of Adelaide, Adelaide, South Australia, 5000

M. A. Lohe
The Flinders University of South Australia, School of Mathematical Sciences, Bedford Park, South Australia, 5042

(Received 22 October 1982; accepted for publication 7 January 1983)

Properties of noninteracting vortices in a class of models which generalize the Ginzburg–Landau model of superconductivity are described. Previous results of existence and uniqueness for solutions to the first-order equations are extended to cover the case in which the gauge photon and the scalar meson become massless, when long range interactions exist. Several properties of the solutions are also discussed. With some assumptions, and with restrictions on the class of models, all finite-energy solutions of the second-order equations are shown to be solutions of the first-order equations. The second-order equations are formulated in a gauge invariant way, resulting in a second-order elliptic system of two coupled nonlinear equations, which completely determine all gauge invariant quantities.

PACS numbers: 11.15. – q, 74.20.De

I. INTRODUCTION

Finite-energy solutions in field theories are of importance because they serve as good starting approximations for the quantum field theory. For nonabelian gauge theories in three space dimensions these solutions are magnetic monopoles, and detailed properties of these monopoles and their interactions are obtained from a study of the relevant field equations. The simplest of the gauge theories with nontrivial finite energy solutions is the abelian Higgs model in two dimensions, for which the static equations are the Ginzburg–Landau equations of superconductivity. A detailed study of the static solutions (vortices) has been undertaken in Refs. 1–3. Of particular interest is the noninteracting case when the coupling constant λ assumes a critical value ($\lambda = 1$); for this value, static solutions exist which describe vortices located at arbitrary positions in the plane. Evidently, the opposing forces due to the massive gauge photon and the scalar (Higgs) meson cancel exactly.

In Refs. 4 and 5 a model has been described which generalizes the Ginzburg–Landau equations by incorporating into the model an arbitrary nonnegative function $F(|\phi|)$ of the scalar field ϕ. This generalization is of interest because it preserves the noninteracting nature of the vortices; properties of the Ginzburg–Landau equations are revealed to be special cases of similar properties for the general system. Solutions can be found by solving three first order equations, and in Ref. 5 solutions were not shown to exist which describe, as for the Ginzburg–Landau equations, vortices located at arbitrary positions in the plane.

In this paper we extend our previous analysis of the generalized system. First, we strengthen results[5] on the existence and uniqueness to include a class of solutions of particular interest. As mentioned above, in general the class of models we consider share features similar to those of the Ginzburg–Landau theory, which appears as the special case $F(|\phi|)\equiv 1$. An exception arises when $F(|\phi|)$ assumes an asymptotic value $F(1)$, which is zero. The masses of the photon and the scalar meson, which are equal for the noninteracting theory, are given by the value of $F(1)$ so that for $F(1) = 0$ we have massless particles. Instead of the short-range interaction experienced by the massive particles, we now have long-range interactions, with the fields decaying to their asymptotic values according to an inverse power law. In Sec. III we demonstrate the existence and uniqueness of solutions to the first-order equations under very general circumstances, including also the massless case, and dispensing with the assumptions of Ref. 5, excepting, of course, the finite-energy condition. Here we draw on the results of Benilan, Brezis, and Crandall[6] and recent work by Vazquez,[7] which investigates equations of the form

$$-\Delta u + \beta(u) \quad \ni \quad g \quad \text{on } \mathbf{R}^N, \tag{1.1}$$

where $\beta(u)$ is a maximal monotone graph and g is a measure. This equation is precisely of the type which appears in Sec. III. Also discussed in Sec. III are several properties of the solutions, including asymptotic estimates.

Now, we turn attention to the full second-order equations obtained by varying the Lagrangian for the generalized model. We pose the question as to whether all finite-energy solutions of the second-order equations are also solutions of the first-order equations. For the Ginzburg–Landau theory the answer is in the affirmative,[2,3] and we extend this result, using maximum principle type arguments, to the general case provided some assumptions are made on $F(|\phi|)$. One assumption is a growth condition on F, which enables us to conclude that $|\phi|$ is bounded, and another assumption, $F > 0$, is also necessary but excludes the massless case. A convenient feature of the abelian gauge theory under consideration is that the gauge covariant equations are readily expressible in gauge invariant form; we can write a closed second-order system of equations for the two gauge invariant quantities $|\phi|$ and f, where f is the Maxwell field tensor. From the solu-

tions for f and $|\phi|$ the gauge potential A can be constructed in a suitable gauge using Maxwell's equations. The gauge invariant system is derived in Sec. IV and the equivalence of the first- and second-order equations demonstrated in Sec. V. The proofs follow the same strategy as in Refs. 2 and 3 but require modification, particularly with the application of the maximum principle. The difficulty in generalizing the proofs is the appearance in the field equations of a term which lies in $L^1(\mathbb{R}^2)$ [see Eq. (2.6)], and for which *a priori* estimates are difficult to obtain. However, first we discuss in Sec. II some properties of the model.

II. THE MODEL

Define the energy functional[4,5]

$$E = \int [\tfrac{1}{4}(F_{ij})^2 + \tfrac{1}{2}F(|\phi|)|D_i\phi|^2 + \tfrac{1}{2}w^2], \tag{2.1}$$

where the integral is understood to be over $\mathbb{R}^2$, $F(|\phi|)$ is nonnegative, and w is defined for each F according to

$$w(|\phi|) = \int_{|\phi|}^{1} sF(s)\,ds. \tag{2.2}$$

The field tensor F_A is given in terms of the gauge potential $A = A_i(x)\,dx^i$ as follows (for notation see Jaffe and Taubes[3]):

$$F_A = dA = \tfrac{1}{2}F_{ij}\,dx^i \wedge dx^j = \tfrac{1}{2}(\nabla_i A_j - \nabla_j A_i)\,dx^i \wedge dx^j, \tag{2.3}$$

and the covariant derivative by

$$D_A\phi = D_i\phi\,dx^i = (\nabla_i\phi - iA_i\phi)\,dx^i, \tag{2.4}$$

where ϕ is a complex valued function on $\mathbb{R}^2$. The Ginzburg–Landau energy functional is recovered by choosing $F \equiv 1$, in which case the potential $\frac{1}{2}w^2$ reduces to the usual ϕ^4 interaction. Notice that we have set the electric field potential A_0, equal to zero. This follows in fact from the requirement of finite energy, $E < \infty$, provided that $F(1) > 0$ (see also Julia and Zee[8]). The particle masses m can be determined heuristically by identifying the coefficients of the quadratic terms in the fields with m^2, and we find $m^2 = F(1)$, where m is the mass of both the gauge photon and the Higgs meson; these masses are equal provided the coupling constant λ in the interaction $\lambda w^2/2$ is equal to 1, as in Eq. (2.1). For $F(1) = 0$, then, the photon and meson are massless; this is verified by the asymptotic estimates of Sec. III (see Proposition 3.7).

The variational equations which follow from (2.1) are

$$df + |\phi|FJ = 0, \tag{2.5}$$

$$*D_A*(FD_A\phi) + wF\phi - \tfrac{1}{2}F'\hat{\phi}\,|D_A\phi|^2 = 0, \tag{2.6}$$

where $|D_A\phi|^2 = *(D_A\phi \wedge *\overline{D_A\phi})$,

$$f = -*F_A = F_{21}, \tag{2.7}$$

$\hat{\phi} = \phi/|\phi|$ and J is the dual of the Noether current:

$$J = *\mathrm{Im}(\hat{\phi}\,\overline{D_A\phi}). \tag{2.8}$$

Equations (2.5) constitute Maxwell's equations, coupled to a complex scalar field ϕ determined by (2.6). Notice that the generalization of (2.1), by including the arbitrary function $F(|\phi|)$, has not changed the form of Maxwell's equations; by putting $\psi = \phi\sqrt{F}$, Eqs. (2.5) take the usual form

$$*df = \mathrm{Im}(\psi\,\overline{D_A\psi}). \tag{2.9}$$

Observe that when $F(1) = 0$, ψ will attain an asymptotic value of zero, and that in this case there is no symmetry breaking if we regard ψ as the fundamental field. However, (2.6) is of a form different to that when $F \equiv 1$, in particular the term $F'\hat{\phi}\,|D_A\phi|^2$ on the right-hand side is new.

The space of continuous gauge potentials with finite energy separates into disjoint sectors labelled by the vortex number n,[3,9] where

$$n = \frac{1}{2\pi}\int f, \tag{2.10}$$

and is an integer. In each such sector the energy is bounded below,

$$E \geqslant 2\pi w(0)|n|. \tag{2.11}$$

This follows from the decomposition, valid for sufficiently smooth fields, following Bogomol'nyi,[4,10]

$$E = \tfrac{1}{2}\int \{(f \pm w)^2 + F|J \pm d|\phi|\,|^2\} \pm 2\pi w(0)n. \tag{2.12}$$

The lower bound is therefore attained if and only if

$$f = w, \quad J = d|\phi| \quad \text{for } n > 0, \tag{2.13a}$$

or

$$f = -w, \quad J = -d|\phi| \quad \text{for } n < 0. \tag{2.13b}$$

These equations can be reduced to a single equation for $|\phi|$, by eliminating the potential A (see Refs. 2–4):

$$\Delta \log|\phi| + w(|\phi|) = 2\pi\sum_{i=1}^{|n|}\delta(x - a^i), \tag{2.14}$$

where the $2n$ parameters (a^i) are the locations of the n vortices in $\mathbb{R}^2$. The gauge fields are constructed from

$$A = -d\alpha + *d(\log|\phi|), \tag{2.15}$$

where $\alpha(x)$ is a gauge parameter. Therefore, from a solution of (2.14), supplemented by the requirement of finite energy, we obtain a solution of Eqs. (2.5) and (2.6).

Let us also make the following observations. Since solutions of (2.14) satisfy

$$E = 2\pi w(0)|n|, \tag{2.16}$$

we must demand that $w(0) < \infty$. This excludes functions F with behavior that is too singular at $|\phi| = 0$, as is evident from (2.2). This includes $F = |\phi|^{-2}$, i.e., $w = -\log|\phi|$, for which (2.14) is linear. Evidently this corresponds to the free field case for theories of the type in Eq. (2.1), in which the kinetic and potential terms are related by the definition (2.2). This is made manifest by defining a new field $u = -\log|\phi|$, and the fields A and u are then seen to be decoupled in a suitable gauge.

Note also that the Hamiltonian (2.1) retains its form under the transformation

$$|\phi| \to |\phi|^{-1}, \tag{2.17}$$

together with the redefinition $|\phi|^{-4}F(|\phi|^{-1}) \to F(|\phi|)$. This provides a way of defining finite-energy vortices in a model with singular behavior at $|\phi| = 0$. For example, $F = |\phi|^{-4}$ violates $w(0) < \infty$ but under (2.17) the Hamiltonian (2.1) is transformed into the Ginzburg–Landau model, with $F \equiv 1$.

III. EXISTENCE AND UNIQUENESS OF VORTEX SOLUTIONS

We have seen that vortex solutions for the models under consideration can always be constructed from solutions of Eq. (2.14). Let

$$u = -\log|\phi|, \quad \beta(u) = w(e^{-u}). \tag{3.1}$$

From the definition (2.3) for w, the condition $F \geqslant 0$, and assuming local integrability for $sF(s)$, β is continuous monotone nondecreasing on $\mathbb{R}$ and hence maximal monotone. Equation (2.4) becomes

$$-\Delta u + \beta(u) = 2\pi \sum_{i=1}^{|n|} \delta(x - a^i). \tag{3.2}$$

This equation is of the form

$$-\Delta u + \beta(u) \quad \ni \quad g, \tag{3.3}$$

which is studied in Refs. 6 and 7, where β is a maximal monotone graph in $\mathbb{R}$. In Ref. 6, $g \in L^1(\mathbb{R}^2)$, and in Ref. 7 results are extended to the case where $g \in \mathcal{M}(\mathbb{R}^2)$, the space of bounded Radon measures in $\mathbb{R}^2$. This latter result is obtained by approximating $g \in \mathcal{M}(\mathbb{R}^2)$ with a sequence $\{g_n\}$ such that $g_n \in C^\infty(\mathbb{R}^2) \cap L^1(\mathbb{R}^2)$ and using the results of Ref. 6. In order to state the existence results, we define first the Marcinkiewicz space $M^p(\mathbb{R}^2)$ and then the exponential orders of growth of β:

Definition 3.1: Let u be a measurable function on $\mathbb{R}^2$, $1 < p < \infty$ and $1/p' + 1/p = 1$. Then $\|u\|_{M^p} = \min\{c \in [0,\infty] \mid \int_\Omega |u(x)| < c(\text{meas}\,\Omega)^{1/p'}$ for all measurable $\Omega \subset \mathbb{R}^2\}$. $M^p(\mathbb{R}^2)$ is the set of measurable functions u on $\mathbb{R}^2$ satisfying $\|u\|_{M^p} < \infty$.

Definition 3.2: The exponential orders of growth of a maximal monotone graph β at infinity are defined as

$$a^+(\beta) = \begin{cases} \sup\left\{a \mid \int_0^\infty \beta(s)e^{-as}\,ds = \infty\right\} & \text{if } \sup D(\beta) = \infty \\ \infty \quad \text{otherwise,} \end{cases}$$

$$a^-(\beta) = \begin{cases} \sup\left\{a \mid -\int_0^\infty \beta(-s)e^{-as}\,ds = \infty\right\} & \text{if } \inf D(\beta) = -\infty \\ \infty \quad \text{otherwise,} \end{cases}$$

where $D(\beta)$ is the domain of β.

It is assumed for (3.3) that

$$0 \in \beta(0) \cap \text{Int}\,\beta(\mathbb{R}). \tag{3.4}$$

Observe that the condition $0 \in \text{Int}\,\beta(\mathbb{R})$ implies $a^\pm \geqslant 0$. Define also the Sobolev spaces $W^{k,p}(\Omega)$, $W^{k,p}_{\text{loc}}(\Omega)$ in the usual way. We need to consider only $g \in \mathcal{M}(\mathbb{R}^2)$ of the form $g = \Sigma_{i=1}^\infty c_i \delta(x - a^i)$, $a^i \in \mathbb{R}^2$, where the $c_i \in \mathbb{R}$ are the point mass coefficients. We can now state:

Theorem 3.3 (Vazquez[7]): Let β have finite exponential orders and let $g \in \mathcal{M}(\mathbb{R}^2)$. There exists a $u \in W^{1,1}_{\text{loc}}(\mathbb{R}^2)$ with $|\nabla u| \in M^2(\mathbb{R}^2)$ and a $w \in L^2(\mathbb{R}^2)$ such that $w \in \beta(u)$ a. e. and $\Delta u = w - g$ if and only if every point mass coefficient of g, c_i, is such that $c^- \leqslant c_i \leqslant c^+$, where the critical values are defined by $c^\pm = \pm 4\pi/a^\pm$. In addition, the solution is unique of $\beta^{-1}(0) = \{0\}$, or $\int g \neq 0$.

This theorem enables us to generalize the results of Ref. 5; we can now include the case $F(1) = 0$ of massless particles and dispense with other assumptions as well. In order to apply the theorem and its further consequences, we note first from (3.1) that

$$\beta(0) = 0. \tag{3.5}$$

We also demand

$$0 < \beta(\infty), \tag{3.6}$$

and, because of finite energy [see (2.16)],

$$\beta(\infty) < \infty. \tag{3.7}$$

A further natural requirement is

$$\beta^{-1}(0) = \{0\}, \tag{3.8}$$

for this is equivalent to demanding that the potential term $w^2/2$ in the expression (2.1) should have a unique minimum, which will lie at $|\phi| = 1$. This ensures that the symmetry breaking, and the asymptotic value of $|\phi|$, are uniquely defined, and excludes functions F which are identically zero in a neighborhood of $|\phi| = 1$. However, solutions still exist and are unique even if (3.8) is violated, and the asymptotic value of $|\phi|$ is then smallest $|\phi|$ for which $w(|\phi|) = 0$.

Since in our application $g \geqslant 0$, it follows that any solution u satisfies $u \geqslant 0$ (Ref. 7, Proposition 2). Together with (3.5) and (3.6) this fact ensures that (3.4) is satisfied. Furthermore, (3.7) implies that the exponential order a^+ takes the value 0. a^- takes a value which depends on F; but, since $a^- \geqslant 0$, $c^- = -4\pi/a^- \leqslant 0$, and we find the conditions $c^- \leqslant c_i \leqslant c^+$ of the theorem always to be satisfied. We conclude therefore that a solution to Eq. (3.2) exists, and is unique.

Remarks 3.4: (i) The unique solution has finite energy. Given $|\phi|$, we construct the gauge potential according to (2.15) and the vortex energy (2.1) is given by [using $f^2 = w^2, |D_A\phi|^2 = 2(\nabla|\phi|)^2$],

$$E = \int (F(\nabla|\phi|)^2 + w^2). \tag{3.9}$$

In order to demonstrate that $E < \infty$, we apply Lemma A.1 of Ref. 7, which extends Lemma A.13 of Ref. 6. Since $\beta(u) \in L^1(\mathbb{R}^2)$ there is a $k > 0$ such that $\text{meas}[u > k] < \infty$. Provided $\beta \in C^1(\mathbb{R})$, at least on $[0,\infty)$, we can choose $p(u) = \beta(u)/\beta(\infty)$; then $p \in C^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ is nondecreasing, and satisfies $|p| \leqslant 1$. The equation

$$\int p'(u)|\nabla u|^2 + \int p(u)\beta(u) \leqslant 2\pi|n| \tag{3.10}$$

from Lemma A.1, Ref. 7, then shows that $E < \infty$. In addition, $\int \Delta u = 0$ shows that $\int w = 2\pi|n|$, i.e., the solution describes n vortices.

(ii) The regularity of the solution depends on the properties of F. Since $|\phi| \leqslant 1$ ($u \geqslant 0$), the regularity of the solution depends only on that of $F(|\phi|)$ on $[0,1]$:

Proposition 3.5[5]: If the first k derivatives of $F(|\phi|)$ are bounded on the interval $[0,1]$, then $|\phi| \in C^{k+1}(\mathbb{R}^2)$.

(iii) If $|\phi| \in C^2(\mathbb{R}^2)$, an application of the strong maximum principle using $|\phi| \leqslant 1$ shows that $|\phi| < 1$ ($u > 0$).

(iv) Define $\beta_+^{-1}(s) = \sup\{t : \beta(t) \ni s\}$. Then we have:

Proposition 3.6 (Vazquez,[7] Lemma 4): Let $g \in \mathscr{M}(\mathbb{R}^2)$ have support in $B_R(0)$, $R>0$ (choose $R > \max_i\{|a^i|\}$). Then u is locally bounded outside $B_R(0)$, and we have the estimate

$$u(x) \leqslant -2|n|\log(1 - R/|x|) + \beta_+^{-1}(2|n|/R(2|x| - R)). \tag{3.11}$$

Thus, if $\beta^{-1}(0) = \{0\}$, u converges uniformly to 0 at infinity.

It is of interest to improve the estimate (3.11), in particular to demonstrate the different behavior of the massive $[F(1) \neq 0]$ and massless $[F(1) = 0]$ models. The former will have an asymptotic dependence $u \sim \exp(-m|x|)$, where m is the mass, while for the latter u will decay more slowly, $u \sim |x|^{-p}$ for some exponent p, as is shown in the following estimates. Let us note that more precise asymptotic estimates, for $\beta(u)$ of the form $\beta(u) = u|u|^{q-1}$ have been given by Veron.[11]

Proposition 3.7: (i) Suppose F is continuous on $[\delta,1]$; F' exists on $[\delta,1]$ for some $0<\delta<1$, and $F(1) \neq 0$. Then for any $\epsilon > 0$ there exists $M < \infty$, $R(\epsilon) > 0$ such that

$$0 < u(x) < M \exp[-|x|(\sqrt{F(1)} - \epsilon)], \quad |x| > R(\epsilon). \tag{3.12}$$

(ii) Suppose $F^{(n-1)}$, $n \geqslant 1$, is continuous on $[\delta,1]$, and $F^{(n)}$ exists on $[\delta,1]$ for some $\delta > 0$, with $F^{(i-1)}(1) = 0$, $i = 1,\ldots,n$, $F^{(n)}(1) \neq 0$. Then there exist $0 < M_1 \leqslant M_2 < \infty$, $R > 0$ such that

$$M_1|x|^{-2/n} \leqslant u(x) \leqslant M_2|x|^{-2/n}, \quad |x| > R. \tag{3.13}$$

Proof: (i) From Proposition 3.6, for sufficiently small $\delta > 0$ there exists $R(\delta) > 0$ such that $0 < u < \delta$, for $|x| > R$. Using Taylor's theorem for $\beta(u)$ on $[0,\delta]$, there exists $\xi \in [0,\delta]$ with

$$\begin{aligned}\beta(u) &= \beta(0) + u\beta'(\xi) \\ &= uF(e^{-\xi})e^{-2\xi} \\ &\geqslant u(F(1) - \epsilon)\end{aligned}$$

by continuity of F. Hence, for $|x| > R$,

$$\Delta u \geqslant u(F(1) - \epsilon). \tag{3.14}$$

Now, since $u \in C^2(\mathbb{R}^2)$ we can apply Proposition 7.2 of Ref. 3 to obtain the result.

(ii) As in (i), apply Taylor's theorem to $\beta(u)$ for $u \in [0,\delta]$:

$$\beta(u) = [u^{n+1}/(n+1)!]\,\beta^{n+1}(\xi), \quad \xi \in [0,\delta]. \tag{3.15}$$

Hence $C_1 u^{n+1} \leqslant \beta(u) \leqslant C_2 u^{n+1}$, for constants $0 < C_1 \leqslant C_2$. Define, for $|x| \geqslant R$, $v = M|x|^{-2/n}$, satisfying

$$\Delta v = (4M^{-n}/n^2)v^{n+1}. \tag{3.16}$$

Now apply the strong maximum principle to $u - v$, to obtain upper and lower bounds on $u(x)$, $|x| > R$. For example, choosing $4M^{-n}/n^2 \leqslant C_1$,

$$\begin{aligned}\Delta(v-u) &\leqslant C_1(v^{n+1} - u^{n+1}) \\ &= C(x)(v-u),\end{aligned} \tag{3.17}$$

where

$$0 \leqslant C(x) = C_1\left(\frac{v^{n+1} - u^{n+1}}{v-u}\right) \in L^\infty(\mathbb{R}^2).$$

Apply the maximum principle to (3.17) on $\{|x| > R\}$, noting that we can choose M sufficiently large to ensure that $v - u|_{|x|=R} \geqslant 0$, to obtain $v - u \geqslant 0$, for $|x| > R$. Similarly we obtain the lower bound.

For the massless case, it is not difficult to find examples which allow explicit solutions. A simple example is the following, in which the polynomial decay for the massless fields is evident:

Example 3.8:

$$F = 8|1 - |\phi|^2|. \tag{3.18}$$

The unique solution to (2.14), for $n = 1$, is

$$|\phi| = |x|/\sqrt{1 + |x|^2}. \tag{3.19}$$

The gauge potential A (in the Coulomb gauge), the field f, and the vortex mass E are readily calculated using formulas such as (2.15) and (2.16), and we find

$$\begin{aligned}A &= -[|x|^2/(1+|x|^2)]\,d\theta, \\ f &= w = 2/(1+|x|^2), \\ E &= 4\pi.\end{aligned} \tag{3.20}$$

IV. SECOND-ORDER EQUATIONS

Following the existence of solutions which achieve the lower energy bound shown in (2.11), a natural question arises as to whether these solutions exhaust all finite-energy solutions. To answer this, we need to return to the second-order equations (2.5) and (2.6). By using maximum principle type arguments, and by modifying the proofs in Ref. 3, we find that, with some assumptions, no new solutions exist. First we simplify Eqs. (2.5) and (2.6), casting them into a gauge invariant form which requires us to solve only two coupled equations, for f and $|\phi|$. The gauge covariance of Eqs. (2.5) and (2.6) implies that there are only three independent equations, for $|\phi|$ and for the two components of A. The equation for $w(|\phi|)$, which follows directly from (2.6), is

$$\Delta w = \rho w - \gamma F^2|\phi|^2|D_A\phi|^2, \tag{4.1}$$

where

$$\begin{aligned}\rho &= F|\phi|^2, \\ \gamma &= \frac{(F|\phi|^2)'}{2F^2|\phi|^3} = \frac{F'}{2F^2|\phi|} + \frac{1}{F|\phi|^2}.\end{aligned} \tag{4.2}$$

From Eqs. (2.5), which are second order in the potential A, we can derive a second-order equation for f by differentiation. We find [using the definition (2.8) for J]

$$\Delta f = \rho f - \gamma F^2|\phi|^2 i^*(D_A\phi \wedge \overline{D_A\phi}). \tag{4.3}$$

By squaring (2.5) and using

$$|J|^2 = |D_A\phi|^2 - (\nabla|\phi|)^2, \tag{4.4}$$

we find

$$|\nabla f|^2 = F^2|\phi|^2|D_A\phi|^2 - (\nabla w)^2. \tag{4.5}$$

Again, using the definition of J,

$$(J, d|\phi|) = \tfrac{1}{2} i^*(D_A\phi \wedge \overline{D_A\phi}), \tag{4.6}$$

and we obtain the following gauge invariant system, involving only the unknowns f and $|\phi|$:

$$\begin{aligned}&\Delta f - \rho f + 2\gamma\nabla f\cdot\nabla w = 0, \\ &\Delta w - \rho w + \gamma[(\nabla f)^2 + (\nabla w)^2] = 0.\end{aligned} \tag{4.7}$$

The boundary conditions for (4.7) are determined by the fin-

ite-energy requirements, which can be written as follows, again using (4.5):

$$\int f^2 < \infty, \quad \int w^2 < \infty,$$
$$\int \frac{(\nabla f)^2}{F|\phi|^2} < \infty, \quad \int \frac{(\nabla w)^2}{F|\phi|^2} < \infty. \tag{4.8}$$

The system (4.7), (4.8) forms a closed elliptic system for f and $|\phi|$, and our aim is to find all solutions of this system. Evidently, solutions can always be obtained by putting $f = \pm w$, with w determined by (2.14). With the solutions of (4.7),(4.8) we can construct the gauge fields using Maxwell's equations (2.5). In order to see this, put

$$\phi = |\phi| e^{i\alpha}, \tag{4.9}$$

where $\alpha(x)$ is a gauge parameter, necessarily multivalued for nontrivial solutions.[3] Equation (2.5) can be written

$$A = -d\alpha - *df/F|\phi|^2. \tag{4.10}$$

Therefore, given f and $|\phi|$ as determined by (4.7),(4.8), we need only to choose a gauge to be able to write down the solution for A. If we can determine that all solutions satisfy $f = \pm w$, we recover Eqs. (2.15); that is, $f = \pm w$ together with Maxwell's equations imply the remaining first-order equations $J = \pm d|\phi|$, which appear in Eqs. (2.13).

Using (4.10), the equation for f can be cast into a useful divergence form:

$$\nabla(\nabla f/F|\phi|^2) = f - g, \tag{4.11}$$

where $g(x) = [\nabla_1, \nabla_2]\alpha(x)$ is singular, being nonzero only at points where $|\phi| = 0$. This is evident from Eqs. (4.9) and (4.10) since, in order that (A,ϕ) be sufficiently smooth, the zeros of $|\phi|$ should coincide with the points where α is discontinuous. In the next section (Proposition 5.2) we demonstrate, following Ref. 2, that we can always choose a gauge in which A is smooth, provided F is sufficiently smooth and assuming local regularity properties of (A,ϕ). It is worth remarking that Eqs. (4.7) and (4.11) for f and $|\phi|$ can be obtained as the Euler equations of the following functional $\mathscr{A}(f,|\phi|)$:

$$\mathscr{A}(f,|\phi|) = \int \left[\frac{(\nabla f)^2 - (\nabla w)^2}{F|\phi|^2} + f^2 - w^2 - 2fg \right]. \tag{4.12}$$

Next we describe a virial theorem, following Ref. 3. Define the Maxwell stress tensor

$$T_{ij} = \{\nabla_i w \nabla_j w - \nabla_i f \nabla_j f + \tfrac{1}{2}\delta_{ij}[(\nabla f)^2 + (\Delta w)^2]\}/F|\phi|^2 + \tfrac{1}{2}\delta_{ij}(f^2 - w^2). \tag{4.13}$$

It follows from (4.7) that

$$\nabla_j T_{ij} = 0, \tag{4.14}$$

and from (4.8) that

$$\int |T_{ij}| < \infty. \tag{4.15}$$

Proposition 4.1: Let (f,w) be a solution to Eqs. (4.7),(4.8). Then the stress tensor (4.13) satisfies

$$\int T_{ij} = 0. \tag{4.16}$$

Proof: See Jaffee and Taubes,[3] p. 31.

As a consequence, we have the following useful relation:

$$\int f^2 = \int w^2. \tag{4.17}$$

V. EQUIVALENCE OF FIRST- AND SECOND-ORDER EQUATIONS

We now require several assumptions on the behavior of F, and also assume local regularity of (A,ϕ). We show then that $|\phi|$ is bounded, and, following Taubes,[2] show that, with a suitable choice of gauge, (A,ϕ) is smooth. This will imply that f and w are continuous, and from (4.7),(4.8) we then show that $w \geqslant |f|$; combined with (4.17) this implies $f = w$ or $f = -w$ and, as explained above, this is sufficient to demonstrate the equivalence of the first- and second-order equations. The assumptions on F are

(i) $F > 0$, (5.1)

(ii) there exists a constant $K \geqslant 1$ such that for all $s > K$,

$$F(s) + \tfrac{1}{2}sF'(s) \geqslant 0, \tag{5.2}$$

(iii) $F \in C^1[0,\infty)$. (5.3)

The first condition is used to obtain a lower bound on F, although it excludes the massless case. The second condition is used solely to show that $\|\phi\|_\infty \leqslant K$; it means that $F(s)s^2$ is a nondecreasing function of s, for $s > K$, and is satisfied by any positive polynomial F and by any function F which increases for $s > K$. Using (5.3), $|\phi| \leqslant K$ implies that $F(|\phi|)$ is bounded above and below:

$$0 < k_1 \leqslant F(|\phi|) \leqslant k_2, \tag{5.4}$$

for finite constants k_1 and k_2. Similarly, because F' is continuous,

$$|F'(|\phi|)| \leqslant k_3 < \infty. \tag{5.5}$$

The third condition (5.3) also ensures that the solutions f, $w \in C^2(\mathbb{R}^2)$, and so in fact are classical solutions (see Proposition 3.5).

We also assume that the components of A belong to $W^{1,2}_{\text{loc}}(\mathbb{R}^2)$, and that $|\phi| \in W^{2,2}_{\text{loc}}(\mathbb{R}^2)$. This last assumption is stronger than that used by Taubes[2] and has been necessary, in order to ensure that f and w are sufficiently smooth, because of the difficulty posed by the extra L^1 term in the field equations (2.6). This assumption implies that $|\phi|$ is continuous.

Proposition 5.1: With the above assumptions, $|\phi| \leqslant K$.

Proof: Let

$$v = \int_{|\phi|}^{1} ds\, F(s). \tag{5.6}$$

v satisfies the distributional equation

$$\Delta v = |\phi| F w - (F/|\phi| + \tfrac{1}{2}F')|D_A \phi|^2 + (F/|\phi|)(\nabla|\phi|)^2. \tag{5.7}$$

Define $b_R(x) = b(|x|/R)$, where $0 \leqslant b(|x|) \leqslant 1$ is a C_0^∞ monotonically decreasing function with

$$b(|x|) = \begin{cases} 1, & |x| \leqslant 1, \\ 0, & |x| \geqslant 2. \end{cases} \tag{5.8}$$

Define $\eta \in W_0^{1,2}(B_{2R}(0))$ by

$$\eta = b_R \max(0, |\phi| - K). \tag{5.9}$$

Equation (5.7) implies

$$\int_{\Omega_{2R}} [\nabla\eta\cdot\nabla v + F\cdot|\phi|w\eta - (\eta/|\phi|)(F + \tfrac{1}{2}|\phi|F')|D_A\phi|^2 + (\eta F/|\phi|)(\nabla|\phi|)^2] = 0,$$

where $\Omega_{2R} = \{ x \in \mathbb{R}^2 | \, |\phi|(x) > K \} \cap B_{2R}(0)$. Observe that all terms are finite, due to the local regularity assumptions and finite energy. Using definitions (5.6) and (5.9) and collecting terms,

$$\int_{\Omega_{2R}} b_R \{ [(|\phi| - K)/|\phi|](F + \tfrac{1}{2}|\phi|F')|D_A\phi|^2 + (KF/|\phi|)\cdot(\nabla|\phi|)^2 - F|\phi|w\cdot(|\phi| - K)\} = -\int_{\Omega_{2R}} [F\cdot(|\phi| - K)\nabla|\phi|\cdot\nabla b_R]. \tag{5.10}$$

Let

$$G(|\phi|) = \int_{|\phi|}^{1} F(s)(s - K)\, ds. \tag{5.11}$$

For $|\phi| > K \geqslant 1$,

$$|G| \leqslant \int_1^{|\phi|} F(s)(s + K)\, ds \leqslant (K + 1)\int_1^{|\phi|} F(s)s\, ds = (K + 1)|w|. \tag{5.12}$$

The integral of the left-hand side of Eq. (5.10) is nonnegative [using (5.2)], and with the definition (5.11) we obtain

$$\int_{\Omega_R} \{ [(|\phi| - K)/|\phi|](F + \tfrac{1}{2}|\phi|F')|D_A\phi|^2 + (KF/|\phi|)\cdot(\nabla|\phi|)^2 - Fw\cdot|\phi|(|\phi| - K)\} \leqslant \int_{\Omega_{2R}} \nabla b_R\cdot\nabla G \leqslant \left[\int_{\Omega_{2R}} G^2\right]^{1/2} \|\Delta b_R\|_{L^2} \leqslant [(K + 1)^2/R]\|\Delta b\|_{L^2}\|w\|_{L^2}, \tag{5.13}$$

where we have integrated by parts, used Hölder's inequality, the estimate (5.12), and the scaling properties of b_R. Since $\Omega_R \subseteq \Omega_{R'}$ for $R' \geqslant R$ we conclude that Ω_∞ has zero measure and hence $\|\phi\|_\infty \leqslant K$.

Next, with the above assumptions, we prove (following Taubes[2]) that it is always possible to choose a gauge in which the potential A is smooth.

Proposition 5.2 (Taubes[2]): Let (A,ϕ) be a weak solution of Eqs. (2.5) and (2.6). Then there exists a pair $(\tilde{A},\tilde{\phi})$ related to (A,ϕ) by $(\tilde{A},\tilde{\phi}) = A - d\alpha, \phi e^{i\alpha})$, where the components of $\tilde{A} \in C^1(\mathbb{R}^2)$, $\tilde{\phi} \in C^0(\mathbb{R}^2)$ and $\alpha \in W^{2,2}(\Omega)$ for all open sets $\Omega \subset \mathbb{R}^2$ with compact closure.

Proof: We need only outline the proof, which is to be found in Ref. 2. By a weak solution A of Eqs. (2.5) we mean a potential A with locally integrable components, and locally integrable first derivatives, satisfying

$$\int [db \wedge *F_A + b \wedge *\mathrm{Im}(F\phi\ \overline{D_A\phi})] = 0, \tag{5.14}$$

where b has components in $W^{1,2}(\mathbb{R}^2)$ and $|\phi| \in W^{2,2}_{\mathrm{loc}}(\mathbb{R}^2)$. We determine the gauge parameter $\alpha(x)$ which transforms A into the Coulomb gauge, in $B = B_2(0)$; that is, we choose $\alpha \in W^{2,2}(B)$ as the unique solution of

$$\Delta\alpha = *d*A, \quad \alpha|_{\partial B} = 0. \tag{5.15}$$

Then, using $|\phi| \leqslant K$, the standard regularity estimates (Morrey,[12] Chap. 6) and the Sobolev imbedding theorem,[13] we find that $\tilde{A} = A - d\alpha$ is continuous in B. Since we have assumed that $|\phi| \in W^{2,2}(B)$ we can iterate, using $F \in C^1[0,\infty)$, to obtain that A and its first derivatives are continuous in B. This means that $f = -*dA$ is continuous in B. Further iterations, using also Eq. (2.6) for ϕ, are possible if extra smoothness is assumed for F. Since the origin was chosen arbitrarily, we find that f, and by assumption $|\phi|$, are continuous in $\mathbb{R}^2$. By a patching procedure we can also construct α such that $\alpha \in W^{2,2}(\Omega)$ for any bounded set $\Omega \subset \mathbb{R}^2$.

Let us now return to the gauge invariant formulation of the second-order equations (4.7). By adding and subtracting these equations, we obtain

$$\Delta u - \rho u + \gamma(\nabla u)^2 = 0, \tag{5.16}$$

which holds for each of $u = w + f$, $u = w - f$. Using $|\phi| \leqslant K$ we find that $F|\phi|^2$ is bounded above and hence, from (4.8), $\|\nabla F\|_{L^2} < \infty$, $\|\nabla w\|_{L^2} < \infty$. This implies that f, $w \in W^{1,2}(\mathbb{R}^2)$, i.e., $u \in W^{1,2}(\mathbb{R}^2)$. A consequence of this and (5.16) is that $u \geqslant 0$. This is straightforward to prove if F is such that $\gamma \geqslant 0$, by application of the maximum principle,[14] as in Refs. 1 and 2. For more general γ we note:

Lemma 5.3: With the above assumptions on F, $\gamma(|\phi|)$ is bounded below.

Proof: From (4.2), for any $\epsilon > 0$,

$$\gamma \geqslant [(F')^2/16F^4|\phi|^2][16F^3/(F')^2 - \epsilon] - 1/\epsilon.$$

Now,

$$16F^3/(F')^2 \geqslant k > 0,$$

for some positive constant k, since by (5.4) and (5.5) $|F'|$ is bounded above, and $F \geqslant k_1$ for some $k_1 > 0$. Hence, by choosing ϵ sufficiently small,

$$\gamma \geqslant -c, \tag{5.17}$$

for some $c > 0$. ■

Lemma 5.4: The function $(e^v - 1)$ for $v \in W^{1,2}(\mathbb{R}^2)$ is square-integrable on $L^2(\mathbb{R}^2)$.

Proof: See Taubes,[1] Lemma 4.6.

Using Lemma 5.3, we obtain

$$\Delta u - c(\nabla u)^2 - \rho u \leqslant 0. \tag{5.18}$$

Proposition 5.5: For $u \in W^{1,2}(\mathbb{R}^2) \cap C^0(\mathbb{R})$, $c > 0$, $\rho(x) \geqslant 0$, and bounded, (5.18) implies that $u \geqslant 0$.

Proof: Define the test function $v \in W_0^{1,2}(B_R(0))$ by

$$v = \begin{cases} (e^{-cu} - 1)b_R & \text{for } u < 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.19}$$

where b_R is the cutoff function defined above [see Eq. (5.8)]. Since v is compactly supported and $v \geqslant 0$, we can multiply (5.18) by v and integrate by parts:

$$-\int \nabla v\cdot\nabla u - c\int v|\nabla u|^2 - \int \rho uv \leqslant 0. \tag{5.20}$$

Using (5.19) and collecting terms,

$$c\int_{\Omega_R}(\nabla u)^2-\int_{\Omega_R}\rho u(e^{-cu}-1)\leqslant\int_{\Omega_R}(e^{-cu}-1)\nabla u\cdot\nabla b_R, \tag{5.21}$$

where $\Omega_R=\{x\in\mathbb{R}^2|u(x)<0\}\cap B_R(0)$. A bound for the right-hand side of (5.21), using Hölder's inequality, is

$$\left|\int_{\Omega_R}(e^{-cu}-1)\nabla u\cdot\nabla b_R\right|\leqslant(\|\nabla b\|_\infty/R)\|\nabla u\|_{L^2}\|e^{-cu}-1\|_{L^2}. \tag{5.22}$$

Since $u\in W^{1,2}(\mathbb{R}^2)$, we have $\|\nabla u\|_{L^2}<\infty$, $\|e^{-cu}-1\|_{L^2}<\infty$ by Lemma 5.4. Taking lim inf $R\to\infty$, we find that Ω_∞ has zero measure, i.e., $u\geqslant 0$.

Since u can be either $w+f$ or $w-f$, we find $w\geqslant|f|\geqslant 0$. Equation (4.17) implies, using continuity, $f^2=w^2$, or $f(x)=\pm w(x)$. Substituting into Eq. (4.11), we find

$$\Delta\log|\phi|+w=0,\quad|\phi|\neq 0. \tag{5.23}$$

Lemma 5.6: Either $w\equiv 0$ or $w>0$.

Proof: Since we have assumed $F\in C^1[0,\infty)$, $|\phi|\in C^2(\mathbb{R}^2)$ (see Ref. 5, Proposition 3.5); also $w\geqslant|f|$ implies $|\phi|\leqslant 1$. Now apply the strong maximum principle to (5.23) on the set $\{x||\phi|(x)>0\}$ to complete the proof (for details, see Ref. 5, Lemma 5.2).

Finally, using Lemma 5.6 and the continuity properties of f and w as in Ref. 3, we deduce that $f(x)=\pm w(x)$ holds with the same sign everywhere, this sign depending on the sign of n by (2.10):

$$f=w\quad\text{if } n>0,\qquad f=-w\quad\text{if } n<0. \tag{5.24}$$

As explained in Sec. IV, Eqs. (5.24) imply the first-order relations (2.15), which together with an analysis of the zeros of $|\phi|$ (see Refs. 3, Chap. III) imply Eq. (3.2), which was investigated in Sec. III.

ACKNOWLEDGMENTS

We wish to thank Professor H. Brezis for bringing the work of Vazquez[7] to our notice and Professor L. C. Evans for helpful comments concerning Sec. V.

[1] C. H. Taubes, Commun. Math. Phys. **72**, 277 (1980).
[2] C. H. Taubes, Commun. Math. Phys. **75**, 207 (1980).
[3] A. Jaffe and C. Taubes, *Vortices and Monopoles* (Birkhauser, Boston, 1980).
[4] M. A. Lohe, Phys. Rev. D **23**, 2335 (1981).
[5] M. A. Lohe and John van der Hoek, "Existence and uniqueness of generalized vortices," J. Math. Phys. **24**, 148 (1983).
[6] Ph. Benilan, H. Brezis, and M. Crandall, Ann. Scuola Norm. Sup. Pisa II **4**, 523 (1975).
[7] J. L. Vazquez, "On a Semilinear Equation in $\mathbb{R}^2$ Involving Bounded Measures," preprint, Universidad Complutense, Madrid, 1981.
[8] B. Julia and A. Zee, Phys. Rev. D **11**, 2227 (1975).
[9] E. Weinberg, Phys. Rev. D **19**, 3008 (1979).
[10] E. B. Bogomol'nyi, Yad. Fiz. **24**, 861 (1976) [Sov. J. Nucl. Phys. **24**, 449 (1976)].
[11] L. Veron, "Asymptotic behaviour of the solutions of some nonlinear elliptic equations," in *Nonlinear Problems of Analysis in Geometry and Mechanics*," edited by M. Atteia, D. Bancel and I. Gumowski (Pitman, Boston, 1981).
[12] C. Morrey, *Multiple Integrals in the Calculus of Variations* (Springer-Verlag, Berlin, Heidelberg, New York, 1966).
[13] R. Adams, *Sobolev Spaces* (Academic, New York, 1975).
[14] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order* (Springer-Verlag, Berlin, Heidelberg, New York, 1977).

A gravitational Poincaré gauge theory and Higgs mechanism

R. J. McKellar
Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

(Received 14 June 1983; accepted for publication 11 August 1983)

In this paper we shall construct the Lagrangian of a gravitational Poincaré gauge theory using degeneracy in the Euler–Lagrange expressions as a primary restriction. Such a generalization of a Lorentz gauge theory requires the addition of not only a translation gauge connection, but also a Goldstone field. The intractability of the field equations is lessened somewhat by means of a particular choice of gauge which acts like a Higgs mechanism. With one further assumption a complete reduction to the corresponding Lorentz theory can be made, and the Einstein vacuum field equations with cosmological term are recovered.

PACS numbers: 11.30.Cp, 11.15.Kc, 12.25. + e, 04.20.Fy

1. INTRODUCTION

Several authors[1] have sought to show how their gravitational field equations can be characterized as those of a unique Poincaré theory. In most instances the Poincaré transformations involved are actually coordinate transformations with parameters from the Poincaré group and not true internal gauge transformations.

It was shown in an earlier paper[2] how the Einstein vacuum field equations with cosmological term could be derived as a consequence of the Euler–Lagrange equations of a Lorentz gauge theory which is in some sense unique. Since the Lorentz group is a subgroup of the Poincaré group, we could also say we have a Poincaré gauge theory. Nonetheless, the absence of any reference to the translation subgroup in the determined Lagrangian should stop us from using this terminology. The aim of this paper is to construct a true Poincaré gauge theory by generalizing the Lorentz theory.

We shall make use of the formalism developed in two previous papers.[2,3] Thus, a Poincaré gauge transformation is characterized by associating at each point of the space-time manifold M (local coordinates x^i, $i = 1,...,4$) an element $u = u(x^i)$ of the connected component of the identity of the Poincaré group. The coordinates of $u(x^i)$ relative to a canonical chart of the first kind[4] are $u^{\alpha\beta}(x^i) = -u^{\beta\alpha}(x^i)$ and $u^{\alpha}(x^i)$, $\alpha,\beta = 1,...,4$.

To generalize the Lorentz gauge theory to a true Poincaré gauge theory, we shall introduce not only a translation gauge connection $A_i^{\ \alpha}$, but also what turns out to be a Goldstone field[5] Φ^{α}. As was shown in Ref. 2, the inclusion of $A_i^{\ \alpha}$ in the formulation of the variational principle without Φ^{α} is futile since the invariance identities eliminate $A_i^{\ \alpha}$ when the Lagrangian is actually constructed. The insertion of Φ^{α} leads to only one additional term to the Lorentz Lagrangian, viz.,

$$de^{ijkh}\eta_{\alpha\beta}f_i^{\ \alpha}{}_j f_k^{\ \beta}{}_h,$$

where d is an arbitrary constant, ϵ^{ijkh} is the four-dimensional Levi-Civita symbol,

$$\eta_{\alpha\beta} \equiv \mathrm{diag}(-1,-1,-1,1)$$

and $f_i^{\ \alpha}{}_j$ is defined in terms of the Poincaré gauge curvatures[2] $F_i^{\ \alpha\beta}{}_j$ and $F_i^{\ \alpha}{}_j$ as

$$f_i^{\ \alpha}{}_j \equiv F_i^{\ \alpha\beta}{}_j\,\eta_{\beta\gamma}\Phi^{\gamma} + F_i^{\ \alpha}{}_j.$$

A simplification of the resulting field equations is obtained by means of a particular choice of gauge which acts like a Higgs mechanism.[5] In this gauge Φ^{α} vanishes and $A_i^{\ \alpha}$ is no longer regarded as a translation gauge connection but as a set of vectors.

To check the validity of the theory, we find that we can reduce it to the Lorentz theory by imposing

$$\Phi^{\alpha}{}_{\|i} = \kappa h_i^{\ \alpha},$$

where a double bar signifies the double covariant derivative,[2,3,6] κ is an arbitrary constant, and the $h_i^{\ \alpha}$ are the components of the orthonormal tetrad (or vierbein).

2. PRELIMINARIES

With a true gauge theory the gauge potential should be a connection in a principal fiber bundle.[7] In particular, the group acts freely on the fiber, i.e., only the action of the identity leaves each element of the fiber invariant. Thus we violate this condition when the action of the Poincaré group is restricted to being

$$h_i^{\ \beta} = a_{\alpha}^{\beta} h'^{\alpha}_{\ i}, \tag{2.1a}$$

where a_{α}^{β} is a Lorentz matrix and a prime indicates the gauge-transformed quantity.

We need to introduce an additional object in the manner of Pilch[8] whose components Φ^{α} undergo the Poincaré gauge transformation

$$\Phi^{\beta} = a_{\alpha}^{\beta}\,\Phi'^{\alpha} + a^{\beta}, \tag{2.1b}$$

where a^{β} characterizes a translation. A coordinate transformation leaves Φ^{α} invariant. When a canonical chart of the first kind is used, the gauge transformation laws (2.1) can be expressed[2] as

$$h'^{\alpha}_{\ i} = \mathscr{L}_{\beta}^{\alpha}\,h_i^{\beta}$$
and
$$\Phi'^{\alpha} = \mathscr{L}_{\beta}^{\alpha}\,\Phi^{\beta} - \mathscr{L}_{\beta}^{\alpha} l_{\gamma}^{\beta}\,u^{\gamma}, \tag{2.2}$$

where

$$\mathscr{L}_{\beta}^{\alpha} \equiv \exp(-u^{\alpha\gamma}\eta_{\gamma\beta})$$

and

$$l_{\beta}^{\alpha} \equiv \delta_{\beta}^{\alpha} + (1/2!)u^{\alpha\gamma}\eta_{\gamma\beta} + (1/3!)u^{\alpha\gamma}\eta_{\gamma\omega}u^{\omega\nu}\eta_{\nu\beta} + \cdots.$$

 0022-2488/84/010161-06$02.50

In addition to Φ^{α}, we shall also make use of the object with components

$$\Phi^{i} \equiv h^{i}_{\alpha}\, \Phi^{\alpha},$$

which enables us to put the transformation laws (2.2) into the form

$$\begin{bmatrix} h^{i}_{\alpha} \\ \Phi^{i} \end{bmatrix}' = \begin{bmatrix} \delta^{i}_{j} \hat{\mathscr{L}}^{\beta}_{\alpha} & 0 \\ -\delta^{i}_{j}\, l^{\beta}_{\gamma} u^{\gamma} & \delta^{i}_{j} \end{bmatrix} \begin{bmatrix} h^{j}_{\beta} \\ \Phi^{j} \end{bmatrix}, \tag{2.3}$$

where h^{i}_{α} is the inverse of h^{α}_{i} and $\hat{}$ denotes the inverse. The purpose of this is to take advantage of the formalism introduced in a previous paper[3] where we now make the identification

$$\rho^{A} = \begin{bmatrix} h^{i}_{\alpha} \\ \Phi^{i} \end{bmatrix}.$$

Under a coordinate transformation $\bar{x}^{i} = \bar{x}^{i}(x^{j})$ with

$$J^{i}_{j} \equiv \frac{\partial x^{i}}{\partial \bar{x}^{j}}$$

and

$$J \equiv \det J^{i}_{j} > 0,$$

we have

$$\begin{bmatrix} \overline{h^{i}_{\alpha}} \\ \Phi^{i} \end{bmatrix} = \begin{bmatrix} \hat{J}^{i}_{j}\, \delta^{\beta}_{\alpha} & 0 \\ 0 & \hat{J}^{i}_{j} \end{bmatrix} \begin{bmatrix} h^{j}_{\beta} \\ \Phi^{j} \end{bmatrix},$$

where we have used a horizontal bar to denote the corresponding quantity in the new coordinate system.

Since ρ^{A} transforms linearly and homogeneously under both Poincaré and coordinate transformations, it is possible to take its double covariant derivative[2,3,6] and obtain

$$h^{i}_{\alpha \| a} = h^{i}_{\alpha,a} + \{_{j}{}^{i}{}_{a}\} h^{j}_{\alpha} - A^{\beta\gamma}_{a} h^{i}_{\beta}\, \eta_{\gamma\alpha}$$

and

$$\Phi^{i}_{\|a} = \Phi^{i}{}_{,a} + \{_{j}{}^{i}{}_{a}\} \Phi^{j} + A^{\beta}_{a} h^{i}_{\beta},$$

where $\{_{j}{}^{i}{}_{a}\}$ is the Christoffel symbol of the second kind and $A^{\beta\gamma}_{a}$ is the Lorentz gauge connection. The corresponding commutation laws[3] for the second derivatives are then

$$h^{i}_{\alpha \| ab} - h^{i}_{\alpha \| ba} = R_{j}{}^{i}{}_{ab}\, h^{j}_{\alpha} - F_{a}{}^{\beta\gamma}{}_{b} h^{i}_{\beta} \eta_{\gamma\alpha}$$

and

$$\Phi^{i}{}_{\|ab} - \Phi^{i}{}_{\|ba} = R_{j}{}^{i}{}_{ab}\, \Phi^{j} + F_{a}{}^{\beta}{}_{b} h^{i}_{\beta},$$

where $R_{j}{}^{i}{}_{ab}$ is the Riemann curvature tensor. It is also possible to show that

$$\Phi^{\gamma}{}_{\|a} = \Phi^{\gamma}{}_{,a} + A^{\gamma}_{a} \eta_{\beta\omega}\, \Phi^{\omega} + A^{\gamma}_{a}$$

and

$$\Phi^{\gamma}{}_{\|ab} - \Phi^{\gamma}{}_{\|ba} = f_{a}{}^{\gamma}{}_{b} \equiv F_{a}{}^{\gamma\beta}{}_{b}\, \eta_{\beta\omega}\, \Phi^{\omega} + F_{a}{}^{\gamma}{}_{b}. \tag{2.4}$$

Note that the gauge transformation law for $\Phi^{\gamma}{}_{\|a}$ is the same as for h^{γ}_{a}, i.e.,

$$\Phi'^{\gamma}{}_{\|a} = \mathscr{L}^{\gamma}_{\omega} \Phi^{\omega}{}_{\|a},$$

and we also have

$$f'_{a}{}^{\gamma}{}_{b} = \mathscr{L}^{\gamma}_{\omega} f_{a}{}^{\omega}{}_{b}.$$

3. DEGENERACY

In Ref. 2 it was found that

$$\begin{aligned} L = {} & a_1 \epsilon^{ijkh} \epsilon_{\alpha\beta\gamma\omega}\, F_{i}{}^{\alpha\beta}{}_{j}\, F_{k}{}^{\gamma\omega}{}_{h} + a_2 \epsilon^{ijkh} \eta_{\alpha\gamma}\, \eta_{\beta\omega}\, F_{i}{}^{\alpha\beta}{}_{j}\, F_{k}{}^{\gamma\omega}{}_{h} \\ & + b_1 h h^{i}_{\mu} h^{j}_{\nu} \eta^{\mu\gamma} \eta^{\nu\omega} \epsilon_{\alpha\beta\gamma\omega}\, F_{i}{}^{\alpha\beta}{}_{j} \\ & + b_2 h h^{i}_{\alpha} h^{j}_{\beta}\, F_{i}{}^{\alpha\beta}{}_{j} + ch, \end{aligned}$$

where a_1, a_2, b_1, b_2, and c are arbitrary constants, $\epsilon_{\alpha\beta\gamma\omega}$ is a four-dimensional Levi-Civita symbol, and

$$h \equiv \det h^{\alpha}_{i},$$

is the most general Lagrangian of the form

$$L = L(h^{\alpha}_{i}; A^{\alpha\beta}_{i}; A^{\alpha\beta}_{i,j}; A^{\alpha}_{i}; A^{\alpha}_{i,j}),$$

which has the transformation laws

$$\bar{L} = JL$$

and

$$L' = L, \tag{3.1}$$

and is degenerate in the sense that its Euler–Lagrange expressions

$$E^{k}_{\sigma\tau} \equiv \frac{\partial L}{\partial A^{\sigma\tau}_{k}} - \frac{\partial}{\partial x^{h}} \left(\frac{\partial L}{\partial A^{\sigma\tau}_{k,h}} \right)$$

and

$$E^{k}_{\sigma} \equiv \frac{\partial L}{\partial A^{\sigma}_{k}} - \frac{\partial}{\partial x^{h}} \left(\frac{\partial L}{\partial A^{\sigma}_{k,h}} \right)$$

are such that

$$\begin{aligned} & \frac{\partial E^{k}_{\sigma\tau}}{\partial A^{\alpha\beta}_{i,jh}} \equiv 0, \quad \frac{\partial E^{k}_{\sigma\tau}}{\partial A^{\alpha}_{i,jh}} \equiv 0, \\ & \frac{\partial E^{k}_{\sigma}}{\partial A^{\alpha\beta}_{i,jh}} \equiv 0, \quad \text{and} \quad \frac{\partial E^{k}_{\sigma}}{\partial A^{\alpha}_{i,jh}} \equiv 0. \end{aligned} \tag{3.2}$$

We shall now generalize this result to a Lagrangian which includes Φ^{i}, i.e.,

$$L = L(h^{\alpha}_{i}; \Phi^{i}; A^{\alpha\beta}_{i}; A^{\alpha\beta}_{i,j}; A^{\alpha}_{i}; A^{\alpha}_{i,j})$$

and demand the same transformation laws (3.1) and degeneracy (3.2). The construction of the Lagrangian follows closely that of Ref. 2 to which the reader should refer constantly. Also, several lemmas were proved in Ref. 2 which are required here and are listed in the Appendix.

To simplify our calculations, we shall use upper case Greek letters to represent all ten gauge indices, so that, for example, A^{Σ}_{i}, $\Sigma = 1,\ldots,10$, signifies the ordered pair $(A^{\alpha\beta}_{i}, A^{\alpha}_{i})$. The degeneracy condition (3.2) can then be expressed as

$$\frac{\partial E^{k}_{\Sigma}}{\partial A^{\Lambda}_{i,jh}} \equiv 0.$$

As in Ref. 2, this condition, together with the invariance identity corresponding to (4.5) in Ref. 3, implies that $\partial^{2} L / \partial A^{\Lambda}_{i,j}\, \partial A^{\Sigma}_{k,h}$ is totally antisymmetric in its Latin indices. Thus,

$$\frac{\partial^{2} L}{\partial A^{\Lambda}_{i,j}\, \partial A^{\Sigma}_{k,h}} = \epsilon^{ijkh} L_{\Lambda\Sigma}(h^{\mu}_{a}; \Phi^{a}), \tag{3.3}$$

where we have made use of the transformation laws of $L_{\Lambda\Sigma}$ inherited from $\partial^2 L/\partial A^{\Lambda}_{i,j}\,\partial A^{\Sigma}_{k,h}$ and the invariance identity corresponding to (4.6) in Ref. 3. Upon integrating (3.3) twice with respect to $A^{\Lambda}_{i,j}$ while noting the appropriate invariance identities we obtain

$$L = \tfrac{1}{8}\epsilon^{ijkh}L_{\Lambda\Sigma}F_k{}^{\Sigma}{}_h F_i{}^{\Lambda}{}_j + \tfrac{1}{2}L^{ij}_{\Lambda}(h^{\mu}_a;\Phi^a)F_i{}^{\Lambda}{}_j + L_0(h^{\mu}_a;\Phi^a),$$

where L^{ij}_{Λ} and L_0 transform in the same way as $\partial L/\partial A^{\Lambda}_{i,j}$ and L, respectively. When we return to lower case Greek indices, we can express the above as

$$\begin{aligned}L = {}&\epsilon^{ijkh}\mathscr{L}_{a\beta\gamma\omega}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma\omega}{}_h\\ &+\epsilon^{ijkh}\mathscr{L}_{\alpha\beta\gamma}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma}{}_h\\ &+\epsilon^{ijkh}L_{\alpha\beta}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha}{}_j F_k{}^{\beta}{}_h\\ &+\mathscr{L}^{ij}_{\alpha\beta}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j + L^{ij}_{\alpha}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha}{}_j + L_0(h^{\mu}_a;\Phi^a).\end{aligned}$$

It is actually more convenient to express L in terms of $f_i{}^{\alpha}{}_j$ rather than $F_i{}^{\alpha}{}_j$, whereby the Lagrangian becomes

$$\begin{aligned}L = {}&\epsilon^{ijkh}L_{a\beta\gamma\omega}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma\omega}{}_h\\ &+\epsilon^{ijkh}L_{\alpha\beta\gamma}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j f_k{}^{\gamma}{}_h\\ &+\epsilon^{ijkh}L_{\alpha\beta}(h^{\mu}_a;\Phi_a)f_i{}^{\alpha}{}_j f_k{}^{\beta}{}_h\\ &+L^{ij}_{\alpha\beta}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j + L^{ij}_{\alpha}(h^{\mu}_a;\Phi^a)f_i{}^{\alpha}{}_j + L_0(h^{\mu}_a;\Phi^a).\end{aligned} \tag{3.4}$$

All that remains in the construction is to determine the structure of the various concomitants of h^{μ}_a and Φ^a as a consequence of their symmetry properties and transformation laws, viz.:

(i) $L_{\alpha\beta\gamma\omega} = -L_{\beta\alpha\gamma\omega} = -L_{\alpha\beta\omega\gamma}$,
$\overline{L}_{\alpha\beta\gamma\omega} = L_{\alpha\beta\gamma\omega}$,
$L'_{\mu\nu\sigma\tau}\mathscr{L}^{\mu}_{\alpha}\mathscr{L}^{\nu}_{\beta}\mathscr{L}^{\sigma}_{\gamma}\mathscr{L}^{\tau}_{\omega} = L_{\alpha\beta\gamma\omega}$;

(ii) $L_{\alpha\beta\gamma} = -L_{\beta\alpha\gamma}$,
$\overline{L}_{\alpha\beta\gamma} = L_{\alpha\beta\gamma}$,
$L'_{\mu\nu\sigma}\mathscr{L}^{\mu}_{\alpha}\mathscr{L}^{\nu}_{\beta}\mathscr{L}^{\sigma}_{\gamma} = L_{\alpha\beta\gamma}$;

(iii) $\overline{L}_{\alpha\beta} = L_{\alpha\beta}$,
$L'_{\mu\nu}\mathscr{L}^{\mu}_{\alpha}\mathscr{L}^{\nu}_{\beta} = L_{\alpha\beta}$;

(iv) $L^{ij}_{\alpha\beta} = -L^{ji}_{\alpha\beta} = -L^{ij}_{\beta\alpha}$,
$\overline{L}^{ij}_{\alpha\beta}J^a_i J^b_j = JL^{ab}_{\alpha\beta}$,
$L'^{ij}_{\mu\nu}\mathscr{L}^{\mu}_{\alpha}\mathscr{L}^{\nu}_{\beta} = L^{ij}_{\alpha\beta}$;

(v) $L^{ij}_{\alpha} = -L^{ji}_{\alpha}$,
$\overline{L}^{ij}_{\alpha}J^a_i J^b_j = JL^{ab}_{\alpha}$,
$L'^{ij}_{\mu}\mathscr{L}^{\mu}_{\alpha} = L^{ij}_{\alpha}$;

(vi) $\overline{L}_0 = JL_0$,
$L'_0 = L_0$.

We begin by considering the quantity

$$B_0 = B_0(h^{\mu}_a;\Phi^a) = L_0/h,$$

which has the transformation laws

$$\overline{B}_0 = B_0$$

and

$$B'_0 = B_0. \tag{3.5}$$

Expansion of (3.5) gives

$$B'_0(\mathscr{L}^{\mu}_{\beta}h^{\beta}_a;\Phi^a - h^a_{\beta}l^{\beta}_{\phi}u^{\phi}) = B_0(h^{\mu}_a;\Phi^a).$$

By taking the derivative with respect to u^{γ} and evaluating at the identity transformation, we obtain

$$-\frac{\partial B_0}{\partial\Phi^a}h^a_{\gamma} = 0$$

and thus

$$\frac{\partial B_0}{\partial\Phi^a} = 0.$$

Lemma A1 of the Appendix then yields

$$B_0 = c,$$

where c is an arbitrary constant and hence

$$L_0 = ch.$$

In a similar manner the remaining quantities are all independent of Φ^a, and we have:

(i) $L_{\alpha\beta\gamma\omega} = a_1\epsilon_{\alpha\beta\gamma\omega} + \tfrac{1}{2}a_2(\eta_{\alpha\gamma}\eta_{\beta\omega} - \eta_{\alpha\omega}\eta_{\beta\gamma})$,

by Lemma A2 of the Appendix;

(ii) $L_{\alpha\beta\gamma} = 0$,

by Lemma A3 of the Appendix;

(iii) $L_{\alpha\beta} = d\eta_{\alpha\beta}$,

by Lemma A4 of the Appendix;

(iv) $L^{ij}_{\alpha\beta} = hh^i_{\mu}h^j_{\nu}\eta^{\mu\gamma}\eta^{\nu\omega}[b_1\epsilon_{\alpha\beta\gamma\omega} + \tfrac{1}{2}b_2(\eta_{\alpha\gamma}\eta_{\beta\omega} - \eta_{\alpha\omega}\eta_{\beta\gamma})]$,

by applying Lemma A2 of the Appendix to

$$D_{\alpha\beta\gamma\omega} \equiv (1/h)\eta_{\gamma\mu}h^{\mu}_i\eta_{\omega\nu}h^{\nu}_j L^{ij}_{\alpha\beta};$$

(v) $L^{ij}_{\alpha} = 0$,

by applying Lemma A3 of the Appendix to

$$D_{\alpha\beta\gamma} \equiv (1/h)\eta_{\alpha\mu}h^{\mu}_i\eta_{\beta\nu}h^{\nu}_j L^{ij}_{\gamma};$$

where a_1, a_2, b_1, b_2, and d are all arbitrary constants.

We have thus established the following:

Theorem 3.1: If a Lagrangian of the form

$$L = L(h^{\alpha}_i;\Phi^i;A^{\alpha\beta}_i;A^{\alpha\beta}_{i,j};A^{\alpha}_i;A^{\alpha}_{i,j})$$

has the transformation laws

$$\overline{L} = JL$$

and

$$L' = L,$$

and is degenerate in the sense that its Euler–Lagrange expressions satisfy

$$E^k_{\sigma\tau} = E^k_{\sigma\tau}(h^{\alpha}_i;h^{\alpha}_{i,j};\Phi^i;\Phi^i{}_{,j};A^{\alpha\beta}_i;A^{\alpha\beta}_{i,j};A^{\alpha}_i;A^{\alpha}_{i,j})$$

and

$$E^k_{\sigma} = E^k_{\sigma}(h^{\alpha}_i;h^{\alpha}_{i,j};\Phi^i;\Phi^i{}_{,j};A^{\alpha\beta}_i;A^{\alpha\beta}_{i,j};A^{\alpha}_i;A^{\alpha}_{i,j})$$

then L is restricted to being

$$\begin{aligned}L = {}&a_1\epsilon^{ijkh}\epsilon_{\alpha\beta\gamma\omega}F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma\omega}{}_h + a_2\epsilon^{ijkh}\eta_{\alpha\gamma}\eta_{\beta\omega}F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma\omega}{}_h\\ &+b_1hh^i_{\mu}h^j_{\nu}\eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}F_i{}^{\alpha\beta}{}_j + b_2hh^i_{\alpha}h^j_{\beta}F_i{}^{\alpha\beta}{}_j + ch\\ &+d\epsilon^{ijkh}\eta_{\alpha\beta}f_i{}^{\alpha}{}_j f_k{}^{\beta}{}_h,\end{aligned} \tag{3.6}$$

where a_1, a_2, b_1, b_2, c and d are arbitrary constants.

Remark 1: There is only one additional term due to the

presence of Φ^i in the Lagrangian, viz., the coefficient of d.

Remark 2: It was shown in Ref. 2 that the coefficients of a_1 and a_2 are divergences, and thus their Euler–Lagrange expressions are identically zero.

The Euler–Lagrange expressions for the Lagrangian (3.6) take the form

$$\begin{aligned}\mathscr{E}^s_\phi \equiv \frac{\partial L}{\partial h^\phi_s} &= b_1 h\,(h^s_\phi h^i_\mu h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}F_i^{\alpha\beta}{}_j \\ &\quad - 2h^s_\mu\, h^i_\phi h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}\, F_i^{\alpha\beta}{}_j) \\ &\quad + b_2 h\,(h^s_\phi\, h^i_\alpha h^j_\beta\, F_i^{\alpha\beta}{}_j - 2h^s_\alpha\, h^i_\phi h^j_\beta\, F_i^{\alpha\beta}{}_j) \\ &\quad + chh^s_\phi + 2d\,{}^*f^i{}_\alpha{}^j F_i^{\alpha\beta}{}_j \eta_{\beta\phi}\Phi^s, \end{aligned} \tag{3.7}$$

$$E_j \equiv \frac{\partial L}{\partial \Phi^j} = 4d\,{}^*f^a{}_\mu{}^b{}_{\|ba}\, h^\mu_j, \tag{3.8}$$

$$E^s_\alpha = -4d\,{}^*f^s{}_\alpha{}^t{}_{\|t}, \tag{3.9}$$

and

$$\begin{aligned}E^s_{\alpha\beta} &= b_1\,{}^\dagger K^s_{\alpha\beta} + b_2 K^s_{\alpha\beta} - 2d\,(\eta_{\beta\gamma}\,\Phi^\gamma\,{}^*f^s{}_\alpha{}^t)_{\|t} \\ &\quad + 2d\,(\eta_{\alpha\gamma}\,\Phi^\gamma\,{}^*f^s{}_\beta{}^t)_{\|t}, \end{aligned} \tag{3.10}$$

where

$$\begin{aligned}&{}^*f^i{}_\alpha{}^j \equiv \epsilon^{ijkh}\eta_{\alpha\beta}f_k{}^\beta{}_h, \\ &K^s_{\alpha\beta} \equiv -2h\,(h^s_{[\alpha}h^t_{\beta]})_{\|t}, \\ &{}^\dagger K^s_{\alpha\beta} \equiv K^s_{\mu\nu}\,\eta^{\mu\gamma}\,\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}, \end{aligned}$$

and square brackets around indices denotes antisymmetrization. It should be noted that E_j and E^s_α are not independent. In fact, even for a Lagrangian which is not degenerate, one of the conservation laws corresponding to (4.8) in Ref. 3 is

$$E_j = -h^\mu_j E^a_{\mu\|a}.$$

4. A CHOICE OF GAUGE

The lack of independence of the Euler–Lagrange expressions suggests that perhaps a particular gauge transformation could simplify the field equations while reducing the degrees of freedom. Such a transformation is given by

$$u^{\alpha\beta} = 0 \tag{4.1}$$

and

$$u^\alpha = \Phi^\alpha,$$

in which case the transformed field variables (signified by a dot), are

$$\dot{h}^i_\alpha = h^i_\alpha$$

and

$$\dot{\Phi}^i = 0.$$

Thus, Φ^i can be thought of as a Goldstone field.

Even though

$$\dot{\Phi}^\alpha = 0$$

and

$$\dot{F}_i^{\alpha\beta}{}_j = F_i^{\alpha\beta}{}_j,$$

and hence

$$\dot{f}_i{}^\alpha{}_j = \dot{F}_i{}^\alpha{}_j = f_i{}^\alpha{}_j,$$

the double covariant derivative of $\dot{\Phi}^\alpha$ does not vanish; in fact,

$$\dot{\Phi}^\alpha{}_{\|i} = \dot{A}^\alpha_i. \tag{4.2}$$

Thus any reference to Φ^α and its derivatives can be eliminated from both the Lagrangian and the field equations.

We still have the freedom to perform any Lorentz gauge transformation. It is then possible to say that we have obtained a Lorentz gauge theory from a Poincaré gauge theory by means of a Higgs mechanism. The Lagrangian is of the form

$$\dot{L} = \dot{L}\,(\dot{h}^\alpha_i; \dot{A}^{\alpha\beta}_i; \dot{A}^{\alpha\beta}_{i,j}; \dot{A}^\alpha_i; \dot{A}^\alpha_{i,j}),$$

and $\dot{A}^\alpha_i$ is no longer regarded as the translation gauge connection, but as a set of vector fields which transform in the same way as $\dot{h}^\alpha_i$. The Lagrangian (3.6) can then be expressed in this gauge as

$$\begin{aligned}\dot{L} &= a_1\,\epsilon^{ijkh}\epsilon_{\alpha\beta\gamma\omega}\,F_i^{\alpha\beta}{}_j\,F_k^{\gamma\omega}{}_h + a_2\epsilon^{ijkh}\eta_{\alpha\gamma}\eta_{\beta\omega}\,F_i^{\alpha\beta}{}_j\,F_k^{\gamma\omega}{}_h \\ &\quad + b_1 hh^i_\mu h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}\,F_i^{\alpha\beta}{}_j + b_2 hh^i_\alpha h^j_\beta\,F_i^{\alpha\beta}{}_j + ch \\ &\quad + 4d\epsilon^{ijkh}\eta_{\alpha\beta}\dot{A}^\alpha_{i\|j}\dot{A}^\beta_{k\|h}, \end{aligned} \tag{4.3}$$

where we have made use of (2.4) and (4.2) and it is now legitimate to consider the double covariant derivative of $\dot{A}^\alpha_i$.

Corresponding to (3.7)–(3.10), we have the Euler–Lagrange expressions

$$\begin{aligned}\dot{\mathscr{E}}^s_\phi &= b_1 h\,(h^s_\phi h^i_\mu h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}F_i^{\alpha\beta}{}_j \\ &\quad - 2h^s_\mu h^i_\phi h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}F_i^{\alpha\beta}{}_j) \\ &\quad + b_2 h\,(h^s_\phi h^i_\alpha h^j_\beta\, F_i^{\alpha\beta}{}_j - 2h^s_\alpha h^i_\phi h^j_\beta\, F_i^{\alpha\beta}{}_j) + chh^s_\phi, \end{aligned} \tag{4.4}$$

$$\begin{aligned}\dot{E}^s_\alpha &= -8d\epsilon^{stij}\eta_{\alpha\beta}\,\dot{A}^\beta_{i\|jt} \\ &= -4d\epsilon^{stij}\eta_{\alpha\beta}\eta_{\gamma\omega}\,\dot{A}^\gamma_i F_j^{\beta\omega}{}_t, \end{aligned} \tag{4.5}$$

and

$$\dot{E}^s_{\alpha\beta} = b_1\,{}^\dagger K^s_{\alpha\beta} + b_2 K^s_{\alpha\beta} + 4d\eta_{\gamma[\alpha}\eta_{\beta]\omega}\epsilon^{stij}\dot{A}^\omega_{i\|j}\,\dot{A}^\gamma_t. \tag{4.6}$$

Note that we have no Euler–Lagrange expression corresponding to (3.8) due to the elimination of Φ^i.

It should be stressed that we do not have a true Lorentz gauge theory here, but one that has been obtained from a Poincaré gauge theory through symmetry breaking involving a Higgs mechanism. The fields $\dot{A}^\alpha_i$ do not arise in a true Lorentz gauge theory without sources.

5. COMPLETE REDUCTION TO LORENTZ

In the particular gauge (4.1) where only subsequent Lorentz transformations are allowed, the ordered pair $(\dot{A}^{\alpha\beta}_i, \dot{A}^\alpha_i)$ can be regarded as the restriction to the Poincaré subgroup of a generalized affine connection as defined by Kobayashi and Nomizu.[9] Furthermore, if we assume

$$\dot{A}^\alpha_i = h^\alpha_i, \tag{5.1}$$

then we have a Poincaré restriction of their affine connection. In doing so, we have completed a reduction[9,10] of the Poincaré theory to a Lorentz theory by means of soldering[7] in addition to the use of a Higgs mechanism. The fields $\dot{A}^\alpha_i$ have now been eliminated.

Some authors[8,11] regard the assumption (5.1) as essential, while others[12] feel that it is not absolutely necessary to

perform such a reduction in all Poincaré gauge theories. When discussing this point few authors stress the fact that it is possible to identify the translation connection and the vierbein only under this choice of gauge where Φ^{α} vanishes and just Lorentz transformations are then allowed. The two quantities transform differently under a general Poincaré transformation, and it does not make sense to take the double covariant derivative of A_i^{α} except under this choice of gauge when we consider that we have just a Lorentz theory.

The above difficulties are overcome by assuming

$$\Phi^{\alpha}{}_{\|i} = h_i^{\alpha} \tag{5.2}$$

instead,[8,11] which reduces to (5.1) under our particular choice of gauge. This effectively completes the reduction by combining the Higgs mechanism and the soldering into one process. A particular choice of gauge is not required.

To see what effect a complete reduction has on our Poincaré gauge theory, we shall generalize (5.2) to

$$\Phi^{\alpha}{}_{\|i} = \kappa h_i^{\alpha}, \tag{5.3}$$

where κ is a constant. This yields the more useful relation

$$f_i{}^{\alpha}{}_j = \Phi^{\alpha}{}_{\|ij} - \Phi^{\alpha}{}_{\|ji} = \kappa(h^{\alpha}_{i\|j} - h^{\alpha}_{j\|i}). \tag{5.4}$$

There are actually two ways to impose (5.3). *A priori* we can substitute (5.3) and (5.4) into the Lagrangian (3.6) and thereby reduce the number of field variables. *A posteriori* it is possible to adjoin (5.3) to the Euler–Lagrange equations corresponding to (3.7)–(3.10). The results are not always the same.[13]

When (5.4) is substituted into (3.6), the coefficient of d becomes

$$4\kappa^2\epsilon^{ijkh}\eta_{\alpha\beta}\, h^{\alpha}_{i\|j} h^{\beta}_{k\|h},$$

which can be expressed as

$$(4\kappa^2\epsilon^{ijkh}\eta_{\alpha\beta}\, h_i^{\alpha} h^{\beta}_{k\|h})_{\|j} - 4\kappa^2\epsilon^{ijkh}\eta_{\alpha\beta} h_i^{\alpha} h^{\beta}_{k\|hj}.$$

By virtue of the commutation law

$$h^{\beta}_{k\|hj} - h^{\beta}_{k\|jh} = -R_k{}^a{}_{hj} h_a^{\beta} + h_k^{\omega} F_h{}^{\beta\gamma}{}_j\, \eta_{\gamma\omega}$$

and the identities

$$R_k{}^a{}_{hj} + R_h{}^a{}_{jk} + R_j{}^a{}_{kh} = 0$$

and

$$\epsilon^{ijkh} = -hh^i_{\alpha} h^j_{\beta} h^k_{\gamma} h^h_{\omega} \eta^{\alpha\mu}\eta^{\beta\nu}\eta^{\gamma\sigma}\eta^{\omega\tau}\epsilon_{\mu\nu\sigma\tau},$$

the coefficient of d in (3.6) takes the form

$$(4\kappa^2\epsilon^{ijkh}\eta_{\alpha\beta} h_i^{\alpha} h^{\beta}_{k\|h})_{\|j} + 2\kappa^2 hh^i_{\mu} h^j_{\nu}\, \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega} F_i{}^{\alpha\beta}{}_j.$$

Since the first term is a divergence and the second term is proportional to the coefficient of b_1 in (3.6), the effective reduced Lagrangian is

$$L = (b_1 + 2\kappa^2 d)hh^i_{\mu} h^j_{\nu}\, \eta^{\mu\gamma}\eta^{\alpha\omega}\epsilon_{\alpha\beta\gamma\omega} F_i{}^{\alpha\beta}{}_j + b_2 hh^i_{\alpha} h^j_{\beta} F_i{}^{\alpha\beta}{}_j + ch.$$

Therefore, the Euler–Lagrange equations yield[2] the Einstein vacuum field equations with cosmological term, i.e.,

$$b_2 R_{ij} = \tfrac{1}{2} c g_{ij},$$

provided

$$(b_1 + 2\kappa^2 d)^2 + b_2{}^2 \neq 0.$$

In a similar, but more tedious, manner, the *a posteriori* imposition of (5.3) in addition to the Euler–Lagrange equations of (3.6) also yields the Einstein vacuum field equations with cosmological term, subject to the same restriction on the constants.

6. DISCUSSION

We have constructed the Lagrangian of a true Poincaré gauge theory whose Euler–Lagrange equations can be simplified by means of a Higgs mechanism. In this form the translation subgroup is manifested only in the translation connection $\dot{A}_i^{\alpha}$. The usual interpretation of such $\dot{A}_i^{\alpha}$ in a gauge theory using a Higgs mechanism is that they are regarded as a set of vector bosons.[5] Thus the generalization of the theory from Lorentz to Poincaré gives rise to an interaction of a set of vector bosons with the gravitational field. An interesting feature of the Lagrangian (4.3) is that minimal coupling arose without having to impose it.

In this paper complete reduction of the Poincaré theory to the Lorentz theory is regarded merely as a check that the Einstein vacuum field equations can be obtained in some sort of limit. Complete reduction eliminates all aspects of the translation subgroup, and thus we no longer have a Poincaré gauge theory. Therefore, complete reduction should not be required.

ACKNOWLEDGMENT

I would like to thank the Natural Sciences and Engineering Research Council of Canada for its award of an operating grant to conduct this research.

APPENDIX

The following lemmas which are used in the body of the paper were proved in Ref. 2:

Lemma A1: If a quantity $B_0 = B_0(h_i^{\alpha})$ is a scalar under both coordinate and Poincaré gauge transformations, i.e., $\overline{B}_0 = B_0$ and $B'_0 = B_0$, then

$$B_0 = c,$$

where c is an arbitrary constant.

Lemma A2: If a quantity $B_{\alpha\beta\gamma\omega} = B_{\alpha\beta\gamma\omega}(h_i^{\mu})$ has the antisymmetries

$$B_{\alpha\beta\gamma\omega} = -B_{\beta\alpha\gamma\omega} = -B_{\alpha\beta\omega\gamma}$$

and the transformation laws

$$\overline{B}_{\alpha\beta\gamma\omega} = B_{\alpha\beta\gamma\omega}$$

and

$$B'_{\rho\nu\sigma\tau}\, \mathscr{L}^{\rho}_{\mu} \mathscr{L}^{\nu}_{\beta} \mathscr{L}^{\sigma}_{\gamma} \mathscr{L}^{\tau}_{\omega} = B_{\mu\beta\gamma\omega},$$

then

$$B_{\alpha\beta\gamma\omega} = a\epsilon_{\alpha\beta\gamma\omega} + b(\eta_{\alpha\gamma}\, \eta_{\beta\omega} - \eta_{\alpha\omega}\, \eta_{\beta\gamma}),$$

where a and b are arbitrary constants.

Lemma A3: If a quantity $B_{\alpha\beta\gamma} = B_{\alpha\beta\gamma}(h_i^{\mu})$ has the antisymmetry

$$B_{\beta\alpha\gamma} = -B_{\alpha\beta\gamma}$$

and the transformation laws

$$\overline{B}_{\alpha\beta\gamma} = B_{\alpha\beta\gamma}$$

and

$$B'_{\rho\nu\sigma}\mathcal{L}^{\rho}_{\mu}\mathcal{L}^{\nu}_{\beta}\mathcal{L}^{\sigma}_{\gamma} = B_{\mu\beta\gamma},$$

then

$$B_{\alpha\beta\gamma} \equiv 0.$$

Lemma A4: If a quantity $B_{\alpha\beta} = B_{\alpha\beta}(h_i^{\mu})$ has the transformation laws

$$\overline{B}_{\alpha\beta} = B_{\alpha\beta}$$

and

$$B'_{\rho\nu}\mathcal{L}^{\rho}_{\mu}\mathcal{L}^{\nu}_{\beta} = B_{\mu\beta},$$

then

$$B_{\alpha\beta} = b\eta_{\alpha\beta},$$

where b is an arbitrary constant.

[1] Y. M. Cho, Phys. Rev. D **14**, 3335 (1976); F. W. Hehl, P. von der Heyde, G. D. Kerlick, and J. M. Nester, Rev. Mod. Phys. **48**, 393 (1976); T. W. B. Kibble, J. Math. Phys. **2**, 212 (1961).
[2] R. J. McKellar, J. Math. Phys. **22**, 2934 (1981).
[3] R. J. McKellar, J. Math. Phys. **22**, 862 (1981).
[4] F. Brickell and R. S. Clarke, *Differentiable Manifolds* (Van Nostrand-Reinhold, London, 1970).
[5] E. S. Abers and B. W. Lee, Phys. Rep. **9**, 1 (1973).
[6] C. N. Yang, "Gauge Fields," in *Proceedings of the Sixth Hawaii Topical Conference on Particle Physics*, edited by P. N. Dobson, Jr., S. Paksava, V. Z. Peterson, and S. F. Tuan (University of Hawaii Press, Honolulu, 1976).
[7] W. Drechsler and M. E. Mayer, *Fibre Bundle Techniques in Gauge Theories* (Springer-Verlag, New York, 1975).
[8] K. A. Pilch, Lett. Math. Phys. **4**, 49 (1980).
[9] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, (Interscience, New York, 1963), Vol. I.
[10] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds and Physics* (Elsevier, New York, 1982).
[11] R. Giachetti, R. Ricci, and E. Sorace, Lett. Math. Phys. **5**, 85 (1981).
[12] L. K. Norris, R. O. Fulp, and W. R. Davis, Phys. Lett. A **79**, 278 (1980); Y. N. Obukov, Phys. Lett. A **90**, 13 (1982).
[13] See, e.g., J. L. Safko, M. Tsamparlis, and F. Elston, Phys. Lett. A **60**, 1 (1977).

Unified geometrical approach to relativistic particle dynamics

A. P. Balachandran,[a] G. Marmo,[b] N. Mukunda,[c] J. S. Nilsson,[d] A. Simoni,[b] E. C. G. Sudarshan,[e] and F. Zaccaria[b]

Center for Particle Theory and Department of Physics, The University of Texas at Austin, Austin, Texas 78712

(Received 15 June 1982; accepted for publication 10 December 1982)

Models for systems of relativistic particle dynamics are reviewed in terms of a geometrical setting for constraint dynamics. They are derived from the same grand abstract space by means of a common reduction procedure and are put in correspondence with invariant subgroups of the Poincaré group. A new model corresponding to the identity subgroup is also discussed.

PACS numbers: 11.80. – m, 11.30.Cp, 02.20.Rt

I. INTRODUCTION: ON THE DESCRIPTION OF BECOMING

Dynamics is the expression of flow by stringing together sequences of configurations together each labelled by a time evolution parameter according to an explicit rule. The collections of configurations so strung together in a well-ordered sequence constitute trajectories of the system, and each trajectory has certain configurational functionals characterizing them. These would be the constants of motion. In this account the configurations are the conventional coordinate space together with the velocity fibers: whatever constitutes the initial specification to make use of Newton's formulation of the equations of motion.

When such ideas are to be implemented for a relativistic system, we do encounter some new problems. Traditionally, we consider clock time as the time evolution parameter, and a configuration is defined by considering simulataneous specification of coordinates and velocities. In relativistic theory this poses a problem since distant simultaniety is not relativistically invariant. If we insist, nevertheless, on using clock time and a canonical formalism, the no-interaction theorem tells us that the only relativistically invariant descriptions could be for noninteracting systems only. We must therefore be prepared to consider other alternatives.

A satisfactory alternative is to consider a time evolution parameter defined dynamically rather than kinematically. Dynamical evolution is with respect to a temporal parameter that has different significance in different states of motion. The dynamical evolution is self-referring and "the time" is independent of the external reference frames.

It turns out that the temporal parameter so defined, being Lorentz-invariant, must have a generator of dynamical evolution which is also Lorentz-invariant, and is different from any of the ten generators of the Poincaré group. In this 11 parameter generator formalism it has been found possible to construct interacting relativistic systems with invariant world lines.

The natural mechanism for bringing about such a description is to make use of the Dirac constraint formalism starting with a system with excess degrees of freedom and systematically reducing them by imposing constraints. Among those constraints we include one which explicitly depends on a parameter τ, which then gets identified with being the evolution parameter. We have thus the curious situation in which motion is generated by constraints.

In the recent literature there have been a number of such models constructed; they are of three kinds depending upon how the initial configuration and phase spaces are chosen. Each such group made use of a primary set of dynamical variables and a set of constraints. In the first kind of models each individual particle is described by four pairs of canonical variables. A system of $2N$ constraints are then imposed to produce $3N$ pairs of canonical variables and an evolution parameter to describe N particles in motion. In the second kind of model a pair of 4-vectors represent spacetime specification of a uniformly moving "center" of the system and the total 4-momentum of the system, respectively. The constraints then relate these quantities to the particle configurations. In the third kind of model the new collective variables introduced are a Lorentz matrix and its canonical conjugate carrying the burden of the inertial frame. Constraints can then be used to obtain interacting relativistic particles describing world lines.

Each of these kinds of models has its own number of starting variables and judiciously chosen constraints. It would be desirable to have a systematic method of dealing with all three models and to see if there are other possibilities of a similar kind.

The present paper is devoted to this task. We start with grand abstract configuration space $\tilde{\Sigma}$ consisting of the semidirect product of the Lorentz group with the product of N 4-vectors. This configuration space thus has $4N + 10$ dimensions. The phase space has twice this dimension. We then take an invariant subgroup G of the Poincaré group P and take the equivalence classes.

$$\Sigma = \tilde{\Sigma}/G$$

as the configuration space of a model. It turns out that by

[a] Supported by the U.S. Department of Energy under Contract DE-AC02-76ERO 3533. Permanent address: Physics Department, Syracuse University, Syracuse, NY 13210.

[b] Istituto di Fisica Teorica, Università di Napoli and Istituto Nazionale Fisica Nucleare, Sezione di Napoli Mostra d'Oltremare Pad. 19, Napoli, Italy.

[c] Permanent address: Indian Institute of Science, Bangalore 560012, India.

[d] Permanent address: Institute of Theoretical Physics, S-41296, Göteborg, Sweden.

[e] Supported by the U.S. Department of Energy under Contract DE-AS05-76ERO 3992.

 0022-2488/84/010167-10$02.50

choosing G to be P itself, the Lorentz subgroup α, and the translation subgroup T^4, respectively, we get the three kinds of models mentioned above. By choosing the identity subgroup of P we are able to generate another kind of model.

Much of our previous work as well as that of other authors are stated in traditional language of canonical mechanics. For making the ideas accessible to a wider group of people to whom modern differential geometry is a standard tool as well as to expose the essential geometric aspects of the developments, we have carried out our formulation in the language of differential geometry.

The plan of the paper is as follows: Sec. II recapitulates the essential background to establish notation and provide the setting. The world line condition is formulated in its general form in Sec. III. The grand configuration space is introduced in Sec. IV along with the equivalence classes which realize the four kinds of formalisms. In Sec. V we construct the phase spaces and the choice of constraints to build up a suitable family of sections of the fiber bundle for each of the models. Some remarks in Sec. VI conclude the paper.

II. A GEOMETRICAL SETTING FOR CONSTRAINT DYNAMICS

In dealing with constraint dynamics, the situation we are presented with is the following.

On a given $2n$-dimensional manifold[1] $\Gamma = T^*\Sigma$ a set of real functions $K_1,...,K_k$ is given. By choosing a value for each one of them a hypersurface M in Γ is determined. We consider the smooth map

$$\kappa\colon \Gamma \to \mathbb{R}^k,$$
$$\gamma \to (K_1(\gamma),...,K_k(\gamma)),$$

and by fixing a value, say $0 \in \mathbb{R}^k$, we get

$$M = \kappa^{-1}(0) = \{\gamma \in \Gamma\colon K_1(\gamma) = \cdots = K_k(\gamma) = 0\}.$$

We assume M to be a submanifold of Γ, of codimension k. If $0 \in \mathbb{R}^k$ is a regular value for κ, then M is a submanifold.

By means of the symplectic structure ω on Γ we can define Poisson brackets and associate vector fields with functions. The vector field X_f associated with the function f is defined by the relation

$$L_{X_f} g = \{f, g\}$$

for any function g. An equivalent definition is given by

$$i_{X_f}\omega = df$$

if ω is the symplectic form of Γ.

A set of vector fields $X_1,...,X_r$ spans a tangent subspace for each point of Γ on considering span $\{X_1(\gamma),...,X_r(\gamma)\}$. Such spaces will constitute the tangent space of a submanifold if and only if the relations.

$$[X_i, X_j] = c_{ij}^m X_m \tag{2.1}$$

are satisfied, with the c_{ij}^m being functions on Γ. This is the Frobenius theorem.

A vector field X can be evaluated at points of M. If it turns out that $X(m)$ is tangent to M for any $m \in M$, we will say that X is tangent to M.

With the above set of functions we will associate the vector fields X_{K_j} and inquire about the relation (2.1). It is simple to prove that they satisfy the condition of the Frobenius theorem if and only if the following relations hold:

$$d\{K_i, K_j\} = c_{ij}^m\, dK_m.$$

The c_{ij}^m will then be functions of the K_i. We say in this case that the K_i form a function group. Such a situation leads to a foliation on Γ and the relevant analysis has been carried out in Ref. 2, to which we will refer extensively in what follows.

Here we do not require the K_i to form a function group; nevertheless, we shall show how, starting with the vector fields X_{K_i} restricted to M, we can generate a set of vector fields tangent to M and satisfying the condition for the Frobenius theorem.

If

$$i\colon M \to \Gamma$$

is the identification map, we can consider the 2-form $i^*\omega$ on M, which is the pullback of ω by i. In general, $i^*\omega$ is degenerate. If its rank is constant the vector fields on M annihilated by it constitute an involutive distribution $\mathscr{D}$, i.e., they obey the Frobenius theorem. We will prove that they are combinations (with coefficients functions on M) of the X_{K_i} evaluated on M. (Notice that in general the X_{K_i} are not tangent to M.) They will be denoted by Y, and the hypothesis is that they satisfy

$$i_Y(i^*\omega) = 0.$$

This implies that

$$(i_Y\omega)_{|M} = 0$$

and therefore one can write

$$i_Y\omega = c_i\, dK_i \quad \text{(summed on } i\text{)}$$

or

$$Y = c_i\, X_{K_i}$$

with the c_i being functions on M. (Here there is an abuse of notation, as Y is actually a vector field on M, but we do consider it as a vector field on Γ.)

Such an expression for Y implies

$$c_i\{K_i, K_j\} = 0 \quad \text{on } M \text{ for any } j = 1,...,k.$$

When a relation involving Poisson brackets is true only when evaluated on M, it is customary to replace the equality sign $=$ with the sign $\approx$ and it is said to be true in a weak sense. Thus our relations can be written as

$$c_i\{K_i, K_j\} \approx 0 \quad \text{for any } j = 1,...,k. \tag{2.2}$$

It is useful to define the antisymmetric matrix A:

$$A_{ij} = \{K_i, K_j\} \tag{2.3}$$

related to $i^*\omega$ by

$$\operatorname{rank} A(m) = \operatorname{rank}(i^*\omega)(m), \quad m \in M.$$

The set of (c_i) can now be considered as nullvectors of $A_{|M}$ and the number of independent nonvanishing vector fields satisfying (2.2) turns out to be

$$d = \operatorname{codim} M - \operatorname{rank} A_{|M}.$$

If $\operatorname{rank} A_{|M}$ is to be a constant on M, the vector fields Y define an involutive distribution $\mathscr{D}$ on M with the above dimension. This allows us to foliate M and to consider

$$\mathscr{N} = M/\mathscr{D}.$$

In physics it is customary to assume $\mathcal{N}$ to be a manifold having the property that

$$\pi: M \to \mathcal{N}$$

is a submersion. It can be proved that $\mathcal{N}$ inherits a symplectic structure ρ, which allows us to call it the "reduced phase space" or "the frozen phase space."[3]

But so far no dynamics has been defined at all. This is done by introducing a one-parameter family of sections

$$\mathcal{N} \times \mathbb{R} \overset{\sigma}{\to} M.$$

From a global point of view this assumes that a section for $M \overset{\pi}{\to} \mathcal{N}$ does exist. (If the vector fields Y integrate to a Lie group $\mathcal{G}$, such that the leaves of the submersion $\pi: M \to \mathcal{N}$ are diffeomorphic to $\mathcal{G}$, the existence of such a section requires the $\mathcal{G}$-bundle to be trivial.) It is on $\sigma(\mathcal{N} \times \mathbb{R}) \subset M$ that dynamics will be defined, not on M itself. The leaves of π are d-dimensional, and it turns out that $k + d$ is an even number. Therefore,

$$\dim \mathcal{N} = 2n - (k + d)$$

is even, and

$$\dim[\sigma(\mathcal{N} \times \mathbb{R}) \subset M] = 2n - (k + d) + 1, \quad d > 0.$$

Of course, if $d = 0$, then $\mathcal{N} = M$, $\dim \sigma(\mathcal{N} \times \mathbb{R}) = 2n - k$, and our procedure generates a dynamics (the trivial one), i.e., a one-parameter group of transformations on M, which is independent of K_i. But in general this is not the case and the set of K_i has a further role. All possible dynamics that can be defined in such a fashion, corresponding to different choices of σ, have the property that the manifolds of states of motion are all diffeomorphic among themselves.

If $Y_1, Y_2, \ldots, Y_d$ are a basis of vector fields which span $\ker i^*$ each dynamical vector field Δ can be expressed as

$$\Delta = \alpha^i Y_i$$

with α^i functions on M. All this is restricted to the submanifold $\sigma(\mathcal{N} \times \mathbb{R}) \subset M$. This vector field Δ is tangent to the submanifold.

But another way to build up dynamics and the appropriate submanifold is commonly used in dealing with constraint dynamics. Besides the K_i functions, another set of d real functions $X_1, \ldots, X_d$ is chosen to constitute the smooth map

$$\mathbb{X}: \Gamma \times \mathbb{R} \to \mathbb{R}^d,$$
$$(\gamma, \tau) \to X^\tau(\gamma).$$

The requirement on the X is that they are functionally independent and together with the K_i define for each value of the parameter τ a $[2n - (k + d)]$-dimensional surface in Γ on which ω turns out to be nondegenerate. To put it differently, the equations

$$c_m \{ \xi_m, \xi_n \} \approx 0$$
$$(m, n = 1, \ldots, k + d) \quad \text{(summed on } m\text{)}$$

(where ξ_m stands for $K_1, \ldots, K_k; X_1, \ldots, X_d$) do not have nontrivial solutions. Then for each $\tau \in \mathbb{R}$ the surface generated by

$$\mathbb{K} \times \mathbb{X}^\tau: \Gamma \to \mathbb{R}^{d+k}$$

by taking the inverse image of $0 \in \mathbb{R}^{d+k}$ is of dimension $2n - (k + d)$. In this way one recovers what was earlier called $\sigma(\mathcal{N} \times \mathbb{R})$, as will be seen in the next section.

From the previous discussion it is clear that different X_i define different dynamical systems even if all of them have diffeomorphic spaces of trajectories. Their carrier spaces may be different.

In many physical situations, the starting space Γ carries a symplectic action $\overline{\mathcal{R}}$ of some Lie group G, i.e., G acts on Γ via canonical transformations. We ask ourselves what happens to such an action with respect to the constraint surface M. It is obvious that only that part of G which maps M onto itself is relevant as far as dynamics is concerned. If all the infinitesimal generators X^G for $\overline{\mathcal{R}}$ happen to satisfy the relations

$$(i_{X^G} dK_i)_{|M} = 0 \quad (i = 1, \ldots, k)$$

then the action carries over to the manifold M. Furthermore, as the action of G on M preserves $i^*\omega$, it happens that $\mathcal{N}$ also will carry a G-action, $\overline{\mathcal{R}}$, which is symplectic with respect to the symplectic structure ρ. This statement follows from the fact that the vector fields $\tilde{Y}$ defined by

$$i_{\tilde{Y}} \omega = d(c_i K_i)$$

when restricted to M coincide with

$$Y = c_i X_{K^i}.$$

Since $(\overline{\mathcal{R}})^* \omega = \omega$ and M is invariant under $\overline{\mathcal{R}}$, we have also

$$(\overline{\mathcal{R}})^* \mathcal{D} = \mathcal{D}.$$

In fact $(\overline{\mathcal{R}})^*(i_X \omega) = i_{\overline{\mathcal{R}}_* \omega}$, if X is a vector field on Γ.[4]

As we have already said, a dynamics is specified only after we have a section

$$\sigma: \mathcal{N} \times \mathbb{R} \to M$$

and it will be a dynamics on $\sigma(\mathcal{N} \times \mathbb{R})$. The submanifold $\sigma(\mathcal{N} \times \{0\}) \subset \sigma(\mathcal{N} \times \mathbb{R})$ can be thought of as the set of all possible Cauchy data for our dynamics. Furthermore, the projected action of G on $\mathcal{N}$ gives an action of G on $\sigma(\mathcal{N} \times \{0\})$ by setting

$$\mathcal{R}^*(g)\sigma(n, 0) = \sigma(\overline{\mathcal{R}}(g)n, 0), \quad n \in \mathcal{N}, \ g \in G.$$

This can be extended to $\sigma(\mathcal{N} \times \mathbb{R})$ by the relation

$$\mathcal{R}^*(g)\sigma(n, \tau) = \sigma(\overline{\mathcal{R}}(g)n, \tau).$$

It is obvious that $\mathcal{R}^*$ is equivariant with respect to the projection $\pi: M \to \mathcal{N}$ restricted to $\sigma(\mathcal{N} \times \mathbb{R}) \to \mathcal{N}$. It is also clear that it depends on the section $\sigma: \mathcal{N} \times \mathbb{R} \to M$. Moreover, it is canonical with respect to the Poisson brackets on $\sigma(\mathcal{N} \times \mathbb{R})$ defined by the symplectic form $\pi_\tau^* \rho$ the pullback of the symplectic form ρ on $\mathcal{N}$ by the map $\pi_\tau: \sigma(\mathcal{N} \cdot \{\tau\}) \to \mathcal{N}$. This coincides with the usual action generated by Dirac brackets defined on all Γ and restricted to $\sigma(\mathcal{N} \times \{\tau\})$.

But, to connect all this with the evolution of physical objects, it will be necessary to properly define the physical variables, namely positions and momenta in spacetime. In the following sections, maps ϕ_a and ψ_a will be introduced, respectively, for the position and momentum 4-vectors of the ath particle. As the group G involved will be the Poincaré group, it will have the usual action on them. We will denote it by $\mathcal{R}_{\text{reg}}$.

We remark that as both dynamics and states of motion

are given by the choice of a section σ, it is the above action $\mathscr{R}^*$ of the Poincaré group that is the physically relevant one.

In the following sections we are going to apply the above procedure to some specific models.

In some of the models the starting functions K satisfy the relations

$$\{K_i, K_j\} = c_{ij}^m K_m \quad (i,j,m = 1,...,k),$$

i.e.,

$$\{K_i, K_j\} \approx 0.$$

They are then said to form a first class set of constraints. The additional functions χ, meeting the previously stated requirements, are said to form, together with the K, a second class set of constraints. We have

$$\text{rank } A_{|M} = 0, \quad d = k,$$

and the determinant of the matrix

$$B_{m,n} = \{\xi_m, \xi_n\}_{|M}$$

reduces to $(\det|\{K_i, \chi_j\}|)^2$. The Poisson brackets are evaluated on $(\mathbb{K}\times\mathbb{X})^{-1}(0)$.

In other models the structure of the matrix B allows us to carry out the reduction procedure through intermediate steps. For them $A_{|M}$ is singular and has nonzero rank r. A nonsingular submatrix A', of even rank r, is then formed by a subset of the K, which are a second class system of constraints to begin with, so that Dirac brackets can be computed relative to them only. To have the final set of second class constraints, one adds to the remaining K an equal number of χ satisfying the requirement

$$\det B \neq 0.$$

III. WORLD LINE CONDITION

With the space $\mathscr{N}$ we can associate dynamics according to Sec. II. There we have seen that this dynamics is defined on $\sigma(\mathscr{N}\times\mathbb{R}) \subset M$, not on M itself. As already stated, in each model a map $\phi_a : \Gamma \to$ spacetime will be introduced to denote the position 4-vector of particle a. By restricting ϕ_a to $\sigma(\mathscr{N}\times\mathbb{R})$, with each trajectory we associate a world line on spacetime. The physical interpretation of such world lines requires that this association has a definite Poincaré-covariant property. It is this requirement that is usually called the world line condition (WLC). The formal statement of this condition is as follows.

The association

$$n \in \mathscr{N} \mapsto \sigma(n,\mathbb{R})$$

defines a line in $\sigma(\mathscr{N}\times\mathbb{R})$ for each n. On such a set of lines we had defined a Poincaré group action $\mathscr{R}^*$ by setting

$$\mathscr{R}^*(g)\circ\sigma(n,\mathbb{R}) = \sigma(\bar{\mathscr{R}}(g)n,\mathbb{R}), \quad g\in G.$$

We can now state the WLC

$$\phi_a \circ \mathscr{R}^*(g)\circ\sigma(n,\mathbb{R}) = \mathscr{R}_{\text{reg}}(g)\circ\phi_a\circ\sigma(n,\mathbb{R}),$$

where $\mathscr{R}_{\text{reg}}$ is the usual action on the four-dimensional vector space of spacetime positions.

For computations it is convenient to express the WLC in a more explicit way in terms of parametrized lines. Recall the one parameter family of section σ^τ, introduced in Sec II. By varying τ, a line on M is described for each n. Such a line is in turn projected for each a onto $\mathbb{R}^4_a$ by ϕ_a, thus yielding the world line of particle a:

$$c_a^n : \mathbb{R} \to \mathbb{R}^4_a,$$
$$c_a^n(\tau) = \phi_a \circ \sigma^\tau(n).$$

The WLC becomes in this context the requirement that the actions $\mathscr{R}_{\text{reg}}$ defined on each $\mathbb{R}^4$ and $\bar{\mathscr{R}}$ on $\mathscr{N}$ are physically consistent, in the sense that if $n' = \bar{\mathscr{R}}(g)n$, then there is a τ' such that

$$c_a^{n'}(\tau') = \mathscr{R}_{\text{reg}}(g)c_a^n(\tau). \tag{3.1}$$

Here τ' can depend on τ, g, and a. This obviously poses conditions on σ^τ.

To satisfy the WLC, we construct a section of $\pi: M \to \mathscr{N}$ in terms of the real functions χ of the previous section, and choose the χ suitably. We consider the subsets $(\mathbb{X}^\tau)^{-1}(0) \equiv N^\tau \subset \Gamma$. A first requirement is that

$$N^\tau \cap M \neq 0.$$

A second is that $N^\tau_{|M}$ be transversal with respect to the fibers of $\pi: M \to \mathscr{N}$. This condition is satisfied if no vector field exists in $\mathscr{D}$ with a flow tangent to $N^\tau_{|M}$.

While the first demand is met in all cases by requiring that the components of $\mathbb{X}^\tau$ constitute additional constraints not identically vanishing on M, the second one needs some elaboration.

Referring to Sec. II, a vector field lying in $\mathscr{D}$ was seen to be $X_{\psi_{|M}}$, with ψ being such that

$$\psi = c_i K_i \tag{3.2}$$

and

$$\{\psi, K_j\}_{|M} = 0, \quad \forall j = 1,...,k. \tag{3.3}$$

Hence

$$L_{X_\psi} K_j = c_i \{K_i, K_j\} = 0. \tag{3.4}$$

We proceed to determine the functions c_i. Equation (3.4) can be written as

$$(A\mathbf{c})_{|M} = 0, \tag{3.5}$$

where $\mathbf{c} = (c_1,...,c_k)$ and A is the matrix (2.3). We recall that in all the models

$$\text{rank } A = r < k.$$

This allows us to choose r components of $\mathbb{K}$ in terms of which the submatrix A' of nonzero determinant can be built. They will be denoted K'_l $(l = 1,...,r)$ and the remaining ones K''_h $(h = 1,...,d)$ so that

$$\psi = c'_l K'_l + c''_h K''_h.$$

There are ∞^d solutions of (3.5): the c'' can be arbitrarily chosen and the c' are then computed as the unique solution of a linear inhomogeneous system of dimension r. A set of independent solutions is obtained by starting with each K''_h in turn. We denote it by, ψ_h:

$$\psi_h = K''_h - (A')^{-1}_{ll'}\{K'_l, K''_h\}K'_{l'}.$$

The ψ_h constitute a basis for first class constraints.

Returning now to the transversality condition, this can be formulated as the requirement that the equations

$$(b_h\{\psi_h, \chi_{h'}\}) = 0$$

with b_h real functions on Γ, have only the trivial solution $b_h = 0$. This is possible iff

$$\det|\{\psi_h,\chi_{h'}\}_{|M}|\neq 0. \tag{3.6}$$

We note at this point that

$$\{\psi_h,\chi_{h'}\}_{|M} = \{K''_h,\chi_{h'}\}^*_{|M}, \tag{3.7}$$

the bracket on the right-hand side being the Dirac bracket, relative to the K' only.

When

$$\operatorname{rank} A_{|M} = 0,$$

there are no second class constraints (i.e., no K'), and Eq. (3.6) reduces to

$$\det|\{K_j,\chi_{j'}\}|\neq 0, \quad j,j' = 1,...,k. \tag{3.8}$$

In all the schemes considered χ^τ_i are chosen so that all but one, say χ^τ_d, are τ-independent and constitute a Poincaré-invariant set. The χ_i $(i = 1,...,d-1)$ define a line on each fiber and $\overline{\mathscr{R}}_{|M}$ simply permutes these lines among themselves. Thus the WLC is satisfied because in this action on lines $\overline{\mathscr{R}}_{|M}$ and $\mathscr{R}^*$ agree.

Further imposing $\chi^\tau_d = 0$ then puts a parameter τ on each line which is not necessarily preserved under the $\overline{\mathscr{R}}_{|M}$ action. However, this leads us to define a value for τ' in terms of τ, g and other variables such that the WLC in the form (3.1) is satisfied.

IV. THE CHOICE OF THE VARIABLES

In this section we will discuss the variables used in each model to describe systems of N interacting particles.

The physical positions and momenta, in spacetime, will be denoted by 4-vectors q^μ_a and p^μ_a for the ath particle $(a = 1,...,N)$. They transform under the action $\mathscr{R}_{\text{reg}}$ of $\mathscr{P}$ defined by

$$\mathscr{R}_{\text{reg}} = (L,b)q_a, = Lq_a + b$$

and

$$\mathscr{R}_{\text{reg}}(L,B)p_a = Lp_a,$$

where L is a 4×4 Lorentz matrix and b a translation 4-vector. Let $\mathscr{L}$ denote the Lorentz group $\{L\}$ and T^4 the translation group $\{b\}$.

We start with an abstract space $\tilde{\Sigma}$, on which proper actions of $\mathscr{P}$ will be defined. We will then show how the various models equipped with such q_a and p_a emerge.

Let us define

$$\tilde{\Sigma} = \mathscr{P}\times\Sigma_0.$$

$\mathscr{P}$ is the Poincaré group and $\Sigma_0 = \otimes_{a=1,...,N}\mathbb{R}^4_a$. Elements of $\tilde{\Sigma}$ will be denoted $[(\Lambda,a),(x)]$, in which $(\Lambda,a)\in\mathscr{P}$ and (x) stands for $x_1,...,x_N$, x_a being a vector in $\mathbb{R}^4_a$. The following action of $\mathscr{P}$ is defined:

$$\mathscr{R}^{(1)}(L,b)[(\Lambda,a),(x)] = [(\Lambda,a)(L,b)^{-1},(L,b)(x)],$$

where on the right-hand side the right action on $\mathscr{P}$ is given by group multiplication and the left, on (x), is the $\mathscr{R}_{\text{reg}}$ on each $\mathbb{R}^4$, i.e.,

$$(L,b)x_a = \mathscr{R}_{\text{reg}}(L,b)x_a = Lx_a + b.$$

Endowed with such an action, $\tilde{\Sigma}$ has the structure of a fiber bundle associated with the trivial principal bundle $\mathscr{P}$. It is therefore possible to consider equivalence classes with respect to $\mathscr{R}^{(1)}$ and obtain distinct spaces

$$\Sigma = \tilde{\Sigma}/\mathscr{R}^{(1)}(g) \tag{4.1}$$

corresponding to distinct subgroups g of $\mathscr{P}$,

$$\Gamma = \mathscr{T}^*\Sigma,$$

such that the basic (abstract) variables are taken and the analysis of the previous section starts.

Another action of $\mathscr{P}$ on $\tilde{\Sigma}$ commuting with $\mathscr{R}^{(1)}$ can be defined to make $\tilde{\Sigma}$ a trivial principal $\mathscr{P}$-bundle. This is

$$\mathscr{R}^{(2)}(L,b)[(\Lambda,a),(x)] = [(L\Lambda,La+b),(x)].$$

Going to the quotient as in (4.1), it gives rise to an action $\mathscr{R}$ on Σ, which in turn can be lifted to Γ. The symplectic manifold Γ therefore carries a symplectic action $\overline{\mathscr{R}}$ of $\mathscr{P}$.[5]

Maps will be seen to exist from Γ to spacetime for the physical positions, i.e.,

$$q^\mu_a = \phi^\mu_a(\gamma), \quad \gamma\in\Gamma,$$

with the property that

$$\phi_a\circ\overline{\mathscr{R}}(L,b) = \mathscr{R}_{\text{reg}}(L,b)\circ\phi_a$$

and, analogously, for the momenta, i.e.,

$$p^\mu_a = \psi^\mu_a(\gamma), \quad \gamma\in\Gamma,$$

$$\psi_a\circ\overline{\mathscr{R}}(L,b) = \mathscr{R}_{\text{reg}}(L,b)\circ\psi_a.$$

The above physical maps need not be defined on the whole of Γ but rather on the part $\sigma(\mathscr{N}\times\mathbb{R})$, where dynamics operates, i.e., where all the constraints are satisfied. Furthermore, it is there that the generalized mass shell relations

$$p^2_a - m^2_a - v_a = 0 \tag{4.2}$$

will hold.

In what follows we will consider four models. Each of them corresponds to an invariant subgroup of $\mathscr{P}$ with respect to which the quotient (4.1) is taken. Four such subgroups are considered, namely $\mathscr{P}$ itself, the Lorentz group $\mathscr{L}$, the translations T^4, and the identity.

A. The model I[6-9]

The equivalence classes are taken with respect to $\mathscr{P}$, i.e.,

$$\Sigma = \tilde{\Sigma}/\mathscr{R}^{(1)}(\mathscr{P})$$

and each of them can be represented by a set of N 4-vectors (z), so that

$$\Sigma\simeq(\mathbb{R}^4)^{\otimes N}.$$

In fact, the class to which $[(\Lambda,a),(x)]$ belongs contains also $[(L,0),(\Lambda,a)(x)]$ and if

$$(z) = (\Lambda,a)(x)$$

this can be denoted $\{(z)\}$.

The other variables in $\Gamma = T^*\Sigma$ are (η), the canonical conjugates to (z). So a point in Γ is represented by $\{(z);(\eta)\}$. The action $\overline{\mathscr{R}}$ can be seen to be

$$\overline{\mathscr{R}}(L,b)\{(z);(\eta)\} = \{(Lz+b);(L\eta)\}.$$

This allows us to identify these variables with the physical spacetime positions and momenta. The relations (4.2) will enter in the definition of M.

B. The model II[10,11]

Here the subgroup to be taken in (4.1) is the Lorentz group $\mathscr{L}$ and

$$\Sigma = \tilde{\Sigma}/\mathscr{R}^{(1)}(\mathscr{L}).$$

Since

$$\mathscr{R}^{(1)}(L,0)[(\Lambda,a),(x)] = [(\Lambda L^{-1},a),(Lx)],$$

one sees that $[(\Lambda,a),(x)]$ is equivalent to $[(1,a),(\Lambda x)]$. Thus the elements of Σ can be denoted $\{Q,(z)\}$, the Q and z_a $(a = 1,\ldots,N)$ being 4-vectors, $z = \Lambda x$, so that

$$\Sigma \simeq (\mathbb{R}^4)^{\otimes(N+1)}.$$

The additional variables for $\Gamma = T^*\Sigma$ will be R and (η), the canonical conjugates to Q and (z). A point of Γ may be written $\{Q,(z);R,(\eta)\}$. The action of $\bar{\mathscr{R}}(L,b)$ on it gives $\{LQ + b,(Lz);LR,(L\eta)\}$. The physical variables

$$q_a = Q + z_a, \quad p_a = R + \eta_a$$

transform with $\mathscr{R}_{\text{reg}}$ but are not canonically conjugate. The relations (4.2) are satisfied once all the constraints on Γ have been imposed, i.e., when the sections σ have also been introduced.

C. The model III[12]

The equivalence classes are taken with respect to the translation group T^4, i.e.,

$$\Sigma = \tilde{\Sigma}/\mathscr{R}^{(1)}(T^4).$$

Since

$$\mathscr{R}^{(1)}(1,b)[(\Lambda,a),(x)] = [(\Lambda,a - \Lambda b),(x + b)],$$

we have

$$[(\Lambda,a),(x)] = [(\Lambda,0),(x + A^{-1}a)].$$

This allows us to denote a point of Σ by $\{\Lambda,(z)\}$ where

$$z_a = x_a + \Lambda^{-1}a.$$

This gives

$$\Sigma \simeq \mathscr{L} \times (\mathbb{R}^4)^{\otimes N}.$$

The variables for $\Gamma = T^*\Sigma$ include those for Σ and the "momentum" variables $S_{\mu\nu} = -S_{\nu\mu}$ and (η), which are conjugate to Λ^{μ}_{ν} and (z), respectively. The nonvanishing Poisson brackets are

$$\{z_{a\mu},\eta_{b\nu}\} = \delta_{ab}\delta_{\mu\nu},$$

$$\{\Lambda^{\mu}_{\nu},S_{\alpha\beta}\} = g_{\nu\beta}\Lambda^{\mu}_{\alpha} - g_{\nu\alpha}\Lambda^{\mu}_{\beta},$$

$$\{S_{\mu\nu},S_{\alpha\beta}\} = g_{\mu\alpha}S_{\nu\beta} - g_{\nu\alpha}S_{\mu\beta} + g_{\mu\beta}S_{\alpha\nu} - g_{\nu\beta}S_{\alpha\mu}.$$

As far as $\bar{\mathscr{R}}$ is concerned, we see that

$$\mathscr{R}^{(2)}(L,b)[(\Lambda,0),(x + \Lambda^{-1}a)]$$
$$= [(L\Lambda,b),(x + \Lambda^{-1}a)]$$
$$\simeq [(L\Lambda,0),(x + \Lambda^{-1}a + (L\Lambda)^{-1}b)]$$

so that

$$\bar{\mathscr{R}}(L,b)\{\Lambda,(z);S,(\eta)\} = \{L\Lambda,(z + (L\Lambda)^{-1}b);LSL^{-1},(\eta)\}.$$

The position variables in spacetime are defined as

$$q_a = \Lambda z_a,$$

and these transform by means of the action on Γ as under $\mathscr{R}_{\text{reg}}$.

The physical energy–momenta are

$$p_a^{\mu} = \Lambda^{\mu}_{j}\eta_a^{j} + \Lambda^{\mu}_{0}[m_a^2 + V_a(z) + \boldsymbol{\eta}_a\cdot\boldsymbol{\eta}_a]^{1/2}.$$

This allows us to satisfy the relations (4.2). Such p_a transform properly as

$$p_a \to Lp_a$$

since the $v_a(z)$ will be chosen to be functions of the differences $z_b - z_c$.

D. The model IV

The equivalence classes are taken with respect to the identity subgroup so that

$$\Sigma = \tilde{\Sigma} = \Sigma^0 \times \mathscr{P}.$$

The variables of Σ are then Λ, Q, and (z), where $\Lambda \in \mathscr{L}$ and Q and (z) are vectors in $\mathbb{R}^4$. The variables of $T^*\Sigma$ are those of Σ and the "momentum" variables $S_{\mu\nu} = -S_{\nu\mu}$, R, (η). Here R_{μ} is conjugate to Q_{μ} and $\eta_{a\mu}$ is conjugate to $z_{a\mu}$ in the usual sense while $S_{\mu\nu}$ is the four-dimensional "angular momentum" conjugate to Λ^{μ}_{ν}. The Poisson brackets are the same as for model III with the addition of

$$\{Q_{\mu},R_{\nu}\} = \delta_{\nu\mu}.$$

The physical position and momentum variables are given by

$$q_a = \Lambda z_a + Q, \quad p_a = \Lambda\eta.$$

The action of the physical (geometrical) Poincaré group is given by $\mathscr{R}_{\text{reg}}$. Under this action q_a and p_a transform as they should:

$$\mathscr{R}_{\text{reg}}(L,b)q_a = Lq_a + b,$$

$$\mathscr{R}_{\text{reg}}(L,b)p_a = Lp_a.$$

Note that z_a and η_a are invariant under $\mathscr{R}_{\text{reg}}$. The mass shell relations (4.2) will hold as a consequence of the definition of M.

V. REDUCED PHASE SPACES AND SECTIONS

To see how the four models fit within the geometrical setup of Sec. II, we will construct the reduced phase space $\mathscr{N}$ for each of the four models following the procedure outlined before. The additional step will be to consider the choice of the constraints $\mathbb{X}$ to build up a family of sections of the bundle $\pi: M \to \mathscr{N}$.

The dimension of the $\mathscr{N}$'s turns out to be always $6N$; this is another reason to call them phase spaces. Another common feature is that the map $\mathbb{K}$ is taken to be invariant under the Poincaré group, which therefore renders M invariant.

A. The model I

The phase space Γ is of dimension $8N$. The $\mathscr{P}$-invariant submanifold M is constructed by introducing the set of N

real-valued functions on Γ,

$$\mathbb{K} = \{K_a\},$$

$$K_a = p_a^\mu p_{a\mu} - m_a^2 - v_a, \quad a = 1,...,N,$$

having the following properties:

— the zero value is in the image of each of them;
— $(dK_1 \wedge \cdots \wedge dK_N)(m) \neq 0 \quad \forall\, m \in M \equiv \mathbb{K}^{-1}(0)$
(i.e., zero is a regular value for $\mathbb{K}$);
— each of them is $\mathscr{P}$-invariant.

$M \equiv \mathbb{K}^{-1}(0)$ is then a submanifold of Γ. Since $\dim \Gamma = 8N$, we have $\dim M = 7N$.

The v_a satisfy the requirement[6–9]

$$\{K_a, K_b\} = 0, \quad a,b = 1,...,N.$$

Therefore, the matrix A vanishes; and

$$d = \dim \mathscr{D} = N.$$

The vector fields X_a which generate $\mathscr{D}$ are then defined through the relations

$$i_{X_a}\omega = dK_a.$$

The dimension of each leaf is N; hence

$$\dim \mathscr{N} = \dim M/\mathscr{D} = 6N.$$

A point in each leaf, depending on a parameter τ, is obtained by imposing the constraints

$$\chi_a = \left(\sum_{b=1}^{N} p_b\right)\cdot(q_{a+1} - q), \quad a = 1,...,N-1,$$

$$\chi_N = \left(\sum_{b=1}^{N} p_b\right)\cdot q_1 - \tau.$$

As shown in the references quoted, they form, together with the K_a a second class system of constraints; therefore,

$$\det|\{K_a, \chi_b\}|_M| \neq 0, \quad a,b = 1,...,N.$$

Since $A_{|M} = 0$, our transversality condition (3.8) coincides with the above.

B. The model II

The phase space Γ has dimension $8N + 8$. The construction of the $\mathscr{P}$-invariant submanifold M is made by introducing $2N + 5$ functions:

$$K_a^{(1)} = P\cdot z_a,$$
$$K_a^{(2)} = P\cdot \eta_a, \qquad a = 1,...,N,$$

$$K_i^{(3)} = \sum_{a=1}^{N} \eta_{ai}, \quad i = 1,2,3,$$

$$K^{(4)} = \sqrt{P^2} - \sum_{a=1}^{N} (m_a^2 - \eta_a^2 + v_a)^{1/2},$$

$$K^{(5)} = \sum_{a=1}^{N} \eta_{a0}.$$

$$\mathbb{K} \equiv (K_a^{(1)}, K_a^{(2)}, K_i^{(3)}, K^{(4)}, K^{(5)}),$$

$$M = K^{-1}(0).$$

The "potentials" v_a are taken to be $\mathscr{P}$-invariant functions of $z_b - z_c$ and η_b. Only $2N + 4$ of them are functionally independent as, for instance, $K^{(5)}$ is a combination of the $K_i^{(3)}$ due to the $K_a^{(2)}$ vanishing; however,

$$(dK_1^{(1)} \wedge \cdots \wedge dK_N^{(1)} \wedge dK_1^{(2)} \wedge \cdots \wedge dK_N^{(2)} \wedge dK_1^{(3)} \wedge \cdots \wedge dK^{(4)} \wedge dK^{(5)})(m) \neq 0$$

for all $m \in M$. We have

$$\operatorname{codim} M = 2N + 4.$$

Again the zero value is regular and M is $\mathscr{P}$-invariant since $\bar{\mathscr{R}}$ either leaves the components of $\mathbb{K}$ invariant or permutes them among themselves.

The $(2N + 5)$-dimensional antisymmetric matrix A, the elements of which are the Poisson brackets of components of $\mathbb{K}$, has the form

$A =$

	$K_a^{(1)}$	$K_a^{(2)}$	$K_i^{(3)}$	$K^{(4)}$	$K^{(5)}$
$K_a^{(1)}$	0	$c \cdots 0$ $0 \cdots c$	$P_1\ P_2\ P_3$ $\cdot\ \cdot\ \cdot$ $\cdot\ \cdot\ \cdot$	x_1 $\cdot$ x_N	P_0 $\cdot$ $\cdot$
$K_a^{(2)}$	$-c \cdots 0$ $0 \cdots -c$	0	$0\ 0\ 0$ $\cdot\ \cdot\ \cdot$ $\cdot\ \cdot\ \cdot$	x_{N+1} $\cdot$ x_{2N}	0 $\cdot$ $\cdot$
$K_i^{(3)}$	$-P_1 \cdots$ $-P_2 \cdots$ $-P_3 \cdots$	$0 \cdots$ $0 \cdots$ $0 \cdots$	$\cdot\ \cdot\ \cdot$ $\cdot\ \cdot$ $\cdot$	0 $\cdot$ $\cdot$	$\cdot$ $\cdot$ $\cdot$
$K^{(4)}$	$-x_1 \cdots -x_N$	$-x_{N+1} \cdots -x_{2N}$	$0 \cdots$	$\cdot\cdot$	$\cdot$
$K^{(5)}$	$-P_0 \cdots$	$0 \cdots$	$\cdots$	$\cdots$	$\cdot\cdot$

where

$$c = P^\mu P_\mu,$$

$$x_a = \{P\cdot z_a, K^{(4)}\},$$

$$x_{a+N} = \{P\cdot \eta_a, K^{(4)}\}.$$

To compute the latter matrix elements and then to evaluate them on M, use is made of the $\mathscr{P}$ invariance of the v_a. This means that both x_a and x_{a+N} are combinations of terms each of which has $P\cdot z_2$ or $P\cdot\eta_a$ as factors; then they vanish on M.

To compute the rank of A on M, we note that its minor A', formed by the first $2N$ rows and $2N$ columns, is nonsingular. We then act on the remaining rows and columns, adding to the elements of a line those of other parallel lines multiplied by suitable constants, to transform A on M into a new matrix $\bar{A}$ having the same rank:

$$\bar{A} = \left[\begin{array}{ccccc:ccc} & & & & & P_1 & 0 & \cdots \\ & & & & & 0 & & \cdot \\ & & & A' & & \cdot & & \cdot \\ & & & & & \cdot & & \cdot \\ \hdashline -P_1 & 0 & \cdots & & & & & \\ 0 & & & & & & & \\ \cdot & & & & & & 0 & \\ \cdot & & & & & & & \\ \cdots & & & & & & & \end{array}\right]$$

the rank of which cannot be $2N+1$ it being antisymmetric. Since $\det A' \neq 0$, we conclude that

$$\operatorname{rank} A = 2N;$$

$$\dim \mathscr{D} = \operatorname{codim} M - \operatorname{rank} A = 4.$$

Since

$$\dim M = 8(N+1) - (2N+4) = 6N+4,$$

we have

$$\dim \mathscr{N} = \dim M/\mathscr{D} = 6N.$$

The set of constraints K' leading to the nonsingular matrix A' is made up of the $K^{(1)}_a$ and $K^{(2)}_a$ $(a = 1,\ldots,N)$.

The four remaining functionally independent K form the K'' set. To these are added four constraints χ to form a second class set. (The explicit form for χ is discussed in Ref. 11.) The transversality condition (3.7) involves the Dirac bracket $\{K'',\chi\}^*$ relative to the K' constraints only. This is satisfied as a previous analysis of this model shows.[11]

C. The model III

The phase space Γ has dimension $8N+12$. Here $2N+5$ functions are introduced to construct the $\mathscr{P}$-invariant submanifold M They are

$$K^{(1)}_\alpha = z^0_\alpha - z^0_{\alpha+1},$$
$$K^{(2)}_\alpha = \eta^0_\alpha - \eta^0_{\alpha+1}, \qquad \alpha = 1,\ldots,N-1,$$

$$K^{(3)}_i = \sum_{a=1}^{N} \eta^i_a, \quad i = 1,\ldots,3,$$

$$K^{(4)}_i = \tfrac{1}{2}\epsilon_{ikl}S^{kl} + \sum_{a=1}^{N} (\mathbf{z}_a \wedge \boldsymbol{\eta}_a)_i,$$

$$K^{(5)} = \eta^0_1 + v(\Delta z) + \sum_{a=1}^{N} (m_a^2 + \boldsymbol{\eta}^2 + v_a)^{1/2}.$$

They are either invariant or transformed into each other under $\mathscr{P}$.

Furthermore the wedge product of their differentials is a $(2N+5)$-form which does not vanish on $M \equiv \mathbb{K}^{-1}(0)$. This is therefore a $\mathscr{P}$-invariant submanifold of dimension $6N+7$.

The antisymmetric matrix A of dimension $(2N+5)$ has the following form:

$$A = \left[\begin{array}{c|ccccc} & K^{(1)} & K^{(2)} & K^{(3)} & K^{(4)} & K^{(5)} \\ \hline K^{(1)} & 0 & A^{(1)} & 0 & 0 & \begin{matrix}1\\0\\\vdots\end{matrix} \\ \hline K^{(2)} & -A^{(1)} & 0 & 0 & 0 & \vdots \\ \hline K^{(3)} & 0 & 0 & 0 & A^{(2)} & \vdots \\ \hline K^{(4)} & 0 & 0 & -A^{(2)} & A^{(3)} & \vdots \\ \hline K^{(5)} & -1\ 0\ \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right]$$

Here

$$A^{(1)} = \begin{vmatrix} 2 & 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 2 & 1 & 0 & \cdot & \cdot & \cdot & \cdot \\ 0 & 1 & 2 & 1 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 & 1 & 2 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 1 & 2 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 & 2 \end{vmatrix},$$

$$A^{(2)} = \begin{vmatrix} 0 & K^{(3)}_3 & -K^{(3)}_2 \\ -K^{(3)}_3 & 0 & K^{(3)}_1 \\ K^{(3)}_2 & -K^{(3)}_1 & 0 \end{vmatrix},$$

and

$$A^{(3)} = \begin{vmatrix} 0 & K^{(4)}_3 & -K^{(4)}_2 \\ -K^{(4)}_3 & 0 & K^{(4)}_1 \\ K^{(4)}_2 & -K^{(4)}_1 & 0 \end{vmatrix}.$$

Evaluated on M,

$$\operatorname{rank} A \leqslant 2(N-1) + 1.$$

A further reduction to

$$\operatorname{rank} A \leqslant 2(N-1)$$

is obtained since an antisymmetric matrix of odd order has vanishing determinant. Direct computation shows $\det A^{(1)} \neq 0$. Therefore,

$$\operatorname{rank} A = 2(N-1).$$

In this case $\mathscr{D}$ has dimension $(2N+5) - 2(N-1) = 7$ and again $\dim \mathscr{N} = 6N$.

The set of K' is formed by the $2(N-1)$ $K^{(1)}_\alpha$ and $K^{(2)}_\alpha$. Imposing only these constraints means restricting the analysis to a subset M' of Γ having dimension $6N+14$. If $i_{M'}$ is the

identification map

$$i_{M'}: M' \to \Gamma,$$

then the original symplectic form ω[17],

$$\omega = \sum_{a=1}^{N} \sum_{\mu=0}^{3} dz_a^\mu \wedge d\eta_{a\mu} + \omega',$$

when pulled back to M' gives

$$i_{M'}^* \omega = \sum_{a=1}^{N} \sum_{i=1}^{3} dz_a^i \wedge d\eta_a^i - N\, dz_{10} \wedge d\eta_{10} + \omega',$$

where ω' pertains to the variables Λ_μ^ν and $S_{\mu\nu}$. This 2-form on M' is seen to be nondegenerate as a consequence of the K' being second class. Introducing new variables to replace z_{10} and η_{10},

$$Q = \sqrt{N} z_{10}, \quad R = \sqrt{N} \eta_{10},$$

we can write

$$i_{M'}^* \omega = \sum_{a=1}^{N} \sum_{i=1}^{3} dz_a^i \wedge d\eta_a^i - dQ \wedge dR + \omega',$$

This is actually the starting symplectic form for the model described in Ref. 12 since the relation between a symplectic form

$$\omega = \tfrac{1}{2} \omega_{\mu\nu}(\xi)\, d\xi^\mu \wedge d\xi^\nu$$

and its associated Poisson brackets

$$\{f, g\} = \omega^{\mu\nu}(\xi) \frac{\partial f}{\partial \xi^\mu} \frac{\partial g}{\partial \xi^\nu}$$

is given by

$$\omega^{\mu\nu} \omega_{\nu\lambda} = \delta_\lambda^\mu.$$

To form the section $\sigma(\mathcal{N} \times \mathbb{R})$ we need to make specific choice of χ as described in Ref. 12.

D. The model IV

The dimension of $T^*\Sigma$ is $8N + 20$ so that a second class system of $2N + 20$ constraints is required to obtain $\dim \mathcal{N} = 6N$. We may choose them to be the following:

$$K_\mu^{(1)} = R_\mu - \sum_{a=1}^{N} p_{a\mu}, \quad \mu,\nu = 1,...,4,$$

$$K_{\mu\nu}^{(2)} = (Q \wedge R)_{\mu\nu} + S_{\mu\nu} - \sum_{a=1}^{N} (q_a \wedge p_a)_{\mu\nu},$$

$$K_a^{(3)} = \eta_a^2 - m_a^2 - v_a, \quad a = 1,...,N,$$

$$\chi_\mu^{(1)} = \sum_{a=1}^{N} \epsilon_a z_{a\mu}, \quad \left(\sum_{a=1}^{N} \epsilon_a = 1,\ \epsilon_a > 0 \right),$$

$$\chi_\alpha^{(2)} = z_{1\alpha} - z_{2\alpha}, \quad \alpha \leqslant 2,$$

$$\chi_\alpha^{(3)} = z_{1\alpha} - z_{3\alpha}, \quad \alpha \leqslant 1,$$

$$\chi^{(4)} = z_{10} - z_{40},$$

$$\chi_\alpha^{(5)} = R \cdot (q_\alpha - q_N), \quad \alpha = 1,...,N-1,$$

$$\chi^{(5)} = R \cdot q_N - \tau.$$

Here we choose v_a in $K_a^{(3)}$ to be functions only of the internal variables z_a and η_a. We choose them to be also invariant under the "Poincaré" group with generators $\Sigma\, \eta_a$, $\Sigma\, (z_a \wedge \eta_a)$ and adjust their functional dependence so that the $K_a^{(3)}$ form a first class set. (This is always possible.[9]) With such a choice $K^{(1)}$, $K^{(2)}$, and $K^{(3)}$ together form a first class set of $(N + 10)$ constraints.

The remaining constraints χ turn this first class set into a second class set. Of these, $\chi^{(1)}$ to $\chi^{(4)}$ are generalizations of those in model III. The functions ϵ_a are functions only of the internal variables (z) and (η) and are thus invariant under the physical Poincaré group. In the free particle limit $v_a \to 0$, they become the "renormalized energies" so that the usual free particle trajectories are recovered as in Ref. 12. The conditions $\chi^{(2)}$ to $\chi^{(4)}$ are designed to fix a Lorentz frame, and thus they are conjugate to $K^{(2)}$. For $N \leqslant 3$ they are clearly inadequate: They must then be replaced by some other "frame fixing" condition. Conditions $\chi^{(5)}$ are the familiar constraints conjugate to $K^{(3)}$.

Since $K^{(1)}$ to $K^{(3)}$ form a first class set $\mathbb{K}$ and the $(N + 10) \times (N + 10)$ matrix of their Poisson brackets with the constraints $\mathbb{X}$ is by construction nondegenerate, it is clear that the $(2N + 20) \times (2N + 20)$ matrix of Poisson brackets is nondegenerate. That is, the constraints $\mathbb{K}$ and $\mathbb{X}$ form a second class set. To be precise, there are degeneracies in these matrices whenever $\chi^{(2)}$ to $\chi^{(4)}$ fail to fix a frame, for instance, when z_1, z_2, and z_3 are parallel. Such situations have to be handled as in Ref. 12.

Thus $M = \mathbb{K}^{-1}(0)$ has dimension $7N + 10$ and the distribution $\mathcal{D}$ has dimension $N + 10$ and is formed by the vector fields X_K. The transversality condition for the σ defined in terms of χ reduces to (3.8) and is satisfied as K and χ form a second class set.

We note the following. The constraints $K^{(1)}$ and $K^{(2)}$ ensure that in the reduced phase space the generators of the physical Poincaré group have the desirable expressions $\Sigma\, p_a$ and $\Sigma\, q_a \wedge p_a$. Also, by virtue of the constraints $\chi^{(1)}$, Q becomes the weighted average $\Sigma\, \epsilon_a q_a$ as in other models.[10,12]

VI. DYNAMICS AS A GATHERING OF MANY INTO A SYSTEM

In the present paper we have started with a grand configuration in which we have a private world to each particle with a 4-vector all to itself and a Lorentz matrix describing the inertial frame. At this stage we had no particles and no motion, no interaction, and no dynamics: We need to generate some *togetherness* and some *self-referral* mechanism to introduce evolution. *Interaction comes from togetherness.*

To form a *system*, this "preparticle" collection has to give up part of its free-wheeling style and subject themselves to some constraints. It is from such constraints that the dynamical system specification and even the notion of dynamical evolution and the evolution parameter emerge.

In this paper we show many alternate patterns to the same goal and how the intermediate stage formulations appear drastically different. We also see in the course of time that not all constraints are on the same footing. Some are gauge constraints which change only the language of description; but some are essential constraints. *Changing the latter means changing the physical system.*

It is fairly straightforward to make choice of the constraints so that the world line condition is satisfied thus fulfilling one of the elementary requirements on relativistic interacting systems. But it was essential to go beyond the

ten-parameter descriptions to the generalized *11-parameter form of Dirac's relativistic dynamics.*

In all this discussion the question of separability for systems with more than two particles has not been answered. We have addressed ourselves to this question elsewhere.[13]

In conclusion, we wish to stress the unifying power of geometry allowing us to view different models for relativistic interacting particles from a common perspective. The emphasis on the role of geometry in description of nature goes back to Plato, and this point of view has been enriched over the centuries by many illustrious scholars.[14] We hope that our work is in keeping with this tradition.

[1]We refer to R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Addison-Wesley, Reading, Mass., 1978) for the theory and notation of the calculus on manifolds.

[2]G. Marmo, E. J. Saletan, and A. Simoni, Nuovo Cimento B **50**, 1 (1979).

[3]Peter G. Bergman and Arthur Komar, "The Hamiltonian in Relativistic Systems of Interacting Particles," Syracuse University, 1980; F. Rohrlich, Phys. Rev. D **25**, 2576 (1982).

[4]Ref. 1, p. 116.

[5]Ref. 1, p. 180.

[6]Ph. Droz-Vincent, Lett. Nuovo Cimento **1**, 839 (1969); **1**, 206 (1973); Phys. Scripta **2**, 129 (1970); Rep. Math. Phys. **8**, 79 (1975); Ann. Inst. H. Poincaré **27**, 407 (1977); Phys. Rev. D **19**, 702 (1979).

[7]I. T. Todorov, "Dynamics of Relativistic Point Particles as a Problem with Constraints," Commun. of the JINR, EZ-10125, Dubna, 1976.

[8]A. Komar, Phys. Rev. D **18**, 1881, 1887, 3017 (1978); **19**, 2908 (1979).

[9]E. C. G. Sudarshan, N. Mukunda, and J. N. Goldberg, Phys. Rev. D **23**, 2218 (1981).

[10]F. Rohrlich, Ann. Phys. **117**, 292 (1979); Physica A **96**, 290 (1979); M. J. King and F. Rohrlich, Ann. Phys. **130**, 350 (1980).

[11]N. Mukunda and E. C. G. Sudarshan, Phys. Rev. D **23**, 2210 (1981).

[12]A. P. Balachandran, G. Marmo, N. Mukunda, J. S. Nilsson, A. Simoni, E. C. G. Sudarshan, and F. Zaccaria, Nuovo Cimento A **67**, 121 (1982).

[13]A. P. Balachandran, D. Dominici, G. Marmo, N. Mukunda, J. S. Nilsson, J. Samuel, E. C. G. Sudarshan, and F. Zaccaria, Phys. Rev. D **26**, 3492 (1982).

[14]G. Galilei, Il Saggiatore, **VI**, 232 (1623); S. Lie, "Zur Theorie der Transformationsgruppen," Christ. Forh. Aar. 1888 Nr. 13, Ges. Abh. BNdv, XXIII (especially pp. 554–7); S. Lie and F. Engle, *Theorie der Transformationsgruppen*, (Teubner, Liepzig, 1888–1893), Vols. I–III (in particular Vol. II); E. Cartan, *Leçons sur les invariants integraux* (Hermann, Paris, 1922); C. Caratheodory, *Calculus of Variations and Partial Differential Equation of the First Order (Part I)* (Holden Day, San Francisco, 1965); A. Lichenerowicz, J. Differential Geom. **12**, 253 (1977); R. Herman, *Interdisciplinary Mathematics* (Math. Sci. Press, Brookline, Mass., 1975, 1977), Vol. 14, Chap. 14, and Vol. 15, p. 52.

A smooth transonic flow in the plane

P. D. Smith[a)]
Institute for Advanced Study, Princeton, New Jersey and Johns Hopkins University, Baltimore, Maryland 21218

(Received 18 May 1982; accepted for publication 7 January 1983)

The implicit function theorem is used to study a symmetric exterior problem for the gas dynamics equation—an equation of mixed type. The existence of families of smooth C^1 solutions is demonstrated. These solutions are families of smooth transonic flows in the plane and are of applied interest. Some of these results have appeared in the literature with an incorrect derivation using the Hodograph mapping. This mapping is not invertible in the transonic case. The methods of this paper do not use the Hodograph mapping and extend to general (e.g., plasma) flows.

PACS numbers: 47.40.Hg, 02.30. + g

INTRODUCTION

Recently L. M. and R. J. Sibner have constructed a family of smooth transonic flows on a symmetric torus.[1] Smooth transonic flows are interesting because of the transonic flow controversy (see Bers[2]), but a physicist might object that flows constrained to a torus are not physical. In this paper the method of Ref. 1 is extended to construct families of smooth transonic flows in an exterior plane domain, showing that the above objection is unfounded.

The extension of their method is necessary because of technical difficulties: In a limiting case of our plane flow, certain derivatives, which are always finite in toroidal flow, become infinite. Also our flow domain is not compact. Together, these two facts require the modification of certain Arzela–Ascoli arguments in Ref. 1. The new arguments use Dini's theorem on the convergence of monotone function sequences instead of the Arzela–Ascoli theorem.

In the toroidal flows shock solutions may also occur. In plane flows, when the polytropic constant $\gamma = 3$, we show that shocks do not occur. Our proof uses the Prandle–Rankine–Hugonant relations for shocks in a polytropic gas. The author conjectures that shocks do not exist for any value of $\gamma > 1$.

Our construction of smooth transonic flows is interesting because it never uses the Hodograph mapping, a mapping which may not be invertible in transonic flow. See Bers,[2] for a discussion of the inapplicability of the Hodograph method in transonic flow, and Courant[3] for the Hodograph approach.

1. DESCRIPTION OF THE PROBLEM

We seek an irrotational, stationary polytropic flow in the exterior of the unit circle considered as a domain in the Euclidean plane. This flow is assumed to have a constant angular speed, i.e., to be independent of the polar angle. We show, directly from the defining differential equation, that there are three flows of this type: purely rotational vortex flow, purely radial source flow with constant mass flow through the circle, and spiral flow with constant mass flow through the circle. The most interesting flow is the spiral flow because this case includes a family of smooth transonic flows.

Our results follow from a complete analysis of the mass flow–circulation problem below:

The mass flow–circulation problem

Consider the exterior of the unit circle as a domain in the Euclidean plane: Show that, in this domain, there exists an irrotational, stationary, polytropic flow that is independent of the polar angle, that has prescribed circulation about the circle, and that has prescribed radial mass transport through the circle.

Remark: The data for the mass flow–circulation problem must lie in certain ranges determined later. The reader will find a complete statement of the results in Sec. 4.

2. THE DIFFERENTIAL EQUATION

We now describe the model of polytropic flow used in this discussion. This model was developed by Sibner and Sibner[4,5] to describe stationary irrotational polytropic flow on a Riemannian manifold.

In this model a flow is described by its velocity field given as a differential 1-form ω that satisfies the equations below:

$$d\omega = 0, \tag{2.1a}$$

$$\delta\rho(Q(\omega))\omega = 0, \tag{2.1b}$$

where $Q(\omega) = g^{ij}\omega_i\omega_j$ is the square speed and $\rho = (1 - \frac{1}{2}(\gamma - 1)Q(\omega))^{1/(\gamma-1)}\gamma > 1$ is the polytropic density function (see Bers[2]). We require that ρ be nonnegative which forces $0 \leqslant Q(\omega) \leqslant 2/(\gamma - 1)$.

Remark: Physically, Eq. (2.1a) is the irrotationality of flow, and Eq. (2.1b) is the conservation of mass. These equations are a mixed quasilinear system. When $0 \leqslant Q(\omega) < 2/(\gamma + 1)$, this system is elliptic and the flow is subsonic; hence, $2/(\gamma + 1)$ is the *square sonic speed*; when $Q(\omega) = 2/(\gamma + 1)$, the system is parabolic; when $Q(\omega) > 2/(\gamma + 1)$, the system is hyperbolic, and the flow is said to be supersonic. This system is the prolongation of the gas dynamics equation to the cotangent bundle of a Riemannian manifold. See Ref. 4 for details.

a) Supported in part by NSF Grant MCS 77-18723 A04.

 0022-2488/84/010177-08$02.50

For flows exterior to the unit circle in the Euclidean plane, it is convenient to use polar coordinates R and θ. With these coordinates $g_{11} = 1$, $g_{22} = R^2$, and $g_{12} = g_{21} = 0$.

Equation (2.1a) reduces to

$$\text{(A)}\quad \alpha_\theta = B_R, \quad \text{where } \omega = \alpha\, dR + \beta\, d\theta. \tag{2.2}$$

Equation (2.1b) reduces to

$$\text{(B)}\quad \frac{\partial}{\partial R}[R\rho\alpha] + \frac{\partial}{\partial\theta}\left[\frac{1}{R}\rho\beta\right] = 0.$$

In the next section, we show that solutions of (A) and (B), which are independent of θ, satisfy a nonlinear algebraic equation. Compare Ref. 1.

3. THE MASS-FLOW RELATION

From now on we consider only flows in the exterior of the unit circle, which are independent of the polar angle. These flows have constant angular velocity (hence β is constant) and are radial, rotational, or spiral flows.

We show that such flows satisfy a conservation law given by a nonlinear algebraic equation—the mass flow relation.

Consider Eqs. (A) and (B). In combination they tell us that any solution $\omega = \alpha\, dR + \beta\, d\theta$, which is independent of θ, must satisfy

$$B_R = \alpha_\theta \frac{\partial}{\partial R}(R\rho\alpha) + \frac{\partial}{\partial\theta}\left[\frac{1}{R}\rho\beta\right] = 0. \tag{3.1}$$

Since β is constant, this implies that $\alpha = \alpha(R)$ is a function only of R. Recall that, in this geometry, $g_{11} = 1$, $g_{22} = R^2$, $g_{12} = g_{21} = 0$. Since $Q(\omega) = g^{ij}\omega_i\omega_j = \alpha^2 + \beta^2/R^2$, we see that Q is independent of θ. Moreover, because $\rho^2 = (1 - \frac{1}{2}(\gamma - 1)Q)2/(\gamma - 1)$, we also see that the density ρ is independent of θ.

Thus the partial differential equation above becomes the nonlinear algebraic equation

$$R^2\rho^2\alpha^2 = K, \quad \text{for some nonnegative const } K, \tag{3.2}$$

and, since $Q = \alpha^2 + \beta^2/R^2$ with β constant, we obtain the mass flow relation (MF) $R^2\rho^2(Q - \beta^2/R^2) = K$ with β constant and K a nonnegative constant.

The mass flow relation has physical as well as mathematical importance. Physically, it says that the mass flow through the circle is zero in rotational vortex flow and constant in both radial and spiral flow. Mathematically the mass flow relation is (for fixed values of K and β) a relation for $Q(\omega)$ as a function of the radius R and this relation determines α from $Q = \alpha^2 + \beta^2/R^2$.

In any case, a flow satisfying the mass flow relation is presented by two parameters: β (which determines the circulation $C = 2\pi\beta$) and K (which determines the radial mass flow).

4. A DESCRIPTION OF PLANE FLOWS EXTERIOR TO THE UNIT CIRCLE—A LIST OF THE RESULTS

This section describes all the solutions to the mass flow–circulation problem. These solutions are symmetric, stationary flows, with fixed circulation $2\pi\beta$: Purely radial (source) flow, purely rotational (vortex) flow and spiral flow. Both vortex and spiral flow include solution families that are everywhere smooth and transonic.

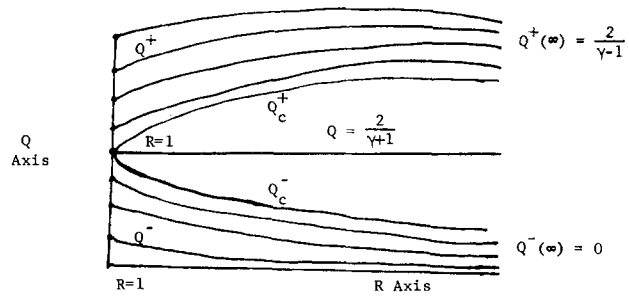

FIG. 1. Radial flow.

Purely radial (source) flow

Here β is zero. The speed is given by $Q = \alpha^2(R)$ and the mass flow relation reduces to $K = R^2\rho(\alpha^2)\alpha^2$.

The constant K parametrizes the solutions. If K is too large, there is no flow at all. Any smaller value K corresponds to two flows. The first flow is everywhere supersonic with a limiting square speed of $2/(\gamma - 1)$ at infinity. We denote this flow by Q^+. The second flow is everywhere subsonic with a limiting speed of zero at infinity. We denote this flow by Q^-. See Fig. 1.

Since $\rho^2(R) = (1 - \frac{1}{2}(\gamma - 1)Q)2(\gamma - 1)$, we see that if $Q = 2/(\gamma - 1)$, the density vanishes. Thus $Q = 2/(\gamma - 1)$ is the square vacuum (or cavitation) speed. Since $Q^+ = 2/(\gamma - 1)$ at infinity and $K = R^2\rho(\alpha^2)\alpha^2$, both Q^+ and Q^- have no mass flow at infinity.

When the initial speed is sonic, i.e., $Q(1) = 2/(\gamma + 1)$, corresponding to the largest value of K for which there is flow, the two flows bifurcate from $R = 1$. See Fig. 1. We call such flow critical flow and $K = K_c$ critical mass flow. Both critical solutions Q_c^+ and Q_c^- are everywhere smooth (C^1), except when $R = 1$, where $d[Q^+]/dR$ is infinite.

Remark: Shock flow—that is, flow that starts on the Q^+ curve and finishes on the Q^- curve labeled by the same K (see Figs. 2 and 3)—might occur. However, we show in Sec. II that, when $\gamma = 3$, shocks never occur.

Purely rotational (vortex) flow

This is a trivial case. This flow is a vortex with constant angular speed parametrized by β. The streamlines are circles, concentric and exterior to the unit circle. There is no radial mass transport ($K = 0$).

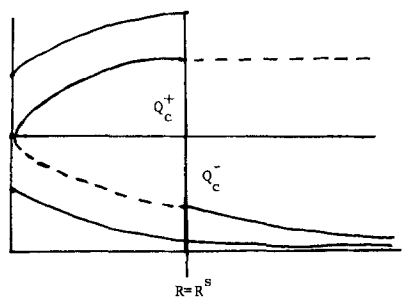

FIG. 2. Shock in critical flow.

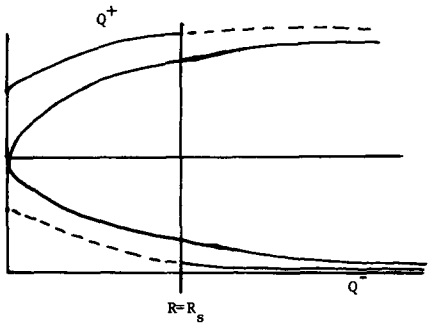

FIG. 3. Shock in noncritical flow.

Since in this case $Q = \beta^2/R^2$, the flow is everywhere subsonic when $\beta > \sqrt{2/(\gamma+1)}$ and otherwise smooth transonic. The limiting speed is zero.

Spiral flow

This is the most interesting flow. It combines the bifurcation feature of radial flow with the smooth transonic flow feature of spiral flow.

In this case the flows are parametrized by K and β. Physically this says the flows are parametrized by mass flow (K) and circulation C ($C = 2\pi\beta$).

Just as in the case of purely radial (source) flow, in spiral flow we have two solutions corresponding to each value of K, up to a critical value K_c of K, and then no solutions if $K > K_c$.

However, in spiral flow the critical mass flow constant K_c depends on β. It is a consequence of this dependence that spiral flows have families of everywhere smooth transonic flows.

To see this, we must think in terms of bifurcation points. In purely radial flow, bifurcation occurred at $R = 1$ when $K = K_c$ and $Q_c^{\pm}(1) = 2/(\gamma+1)$, the sonic speed. But in spiral flow bifurcation occurs at $R = 1$ when $K = K_c$, and K depends on β. Because of the mass flow relation K_c determines $Q_c^{\pm}(1)$ [just let $R = 1$ in (MF)] and this means that bifurcation occurs when $R = 1$ and

$$Q(1) = \hat{Q}(1) = [2/(\gamma+1)](1+\beta^2). \tag{4.1}$$

The new bifurcation point is called the critical speed. The critical speed is generally larger than the sonic speed (if $\beta > 0$) and replaces the sonic speed as a bifurcation point in spiral flow. See Fig. 4.

Since k determines $Q(1)$ from the mass flow relation, we could also parametrize the $Q^{\pm}$ flows by β and $Q^{\pm}(1)$.

In spiral flow shocks might occur (although the author conjectures that they do not). When $\gamma = 3$, we show in Sec. 11 that shocks never occur.

However, we do have families of smooth transonic flows. To see this, consider the Q_- flows, with $\beta^2 < 2/(\gamma-1)$, and with $2/(\gamma+1) < Q^-(1) \leqslant \hat{Q}(1)$. These flows are a family of everywhere smooth transonic spiral flows with zero limiting speed at infinity.

5. TWO EQUIVALENT PROBLEMS

We reformulate the mass flow–circulation problem as an initial value problem.

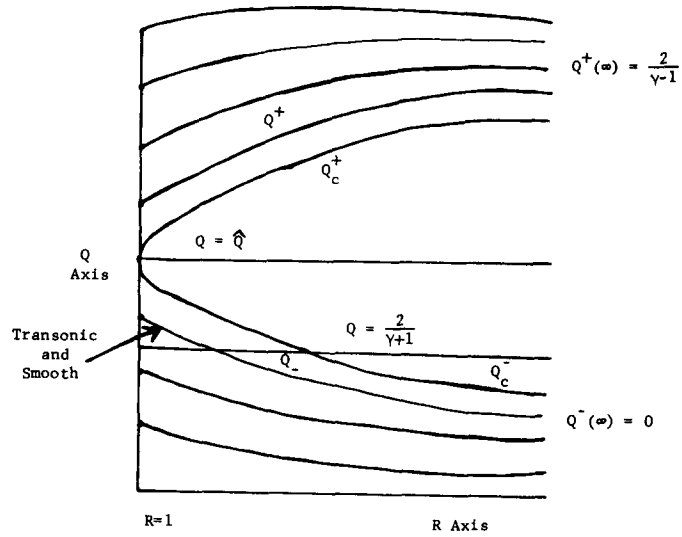

FIG. 4. Spiral flow.

Consider the following two problems:

Problem I: Mass flow–circulation problem: Find C^1 functions $Q^{\pm}(R)$, satisfying $R^2\rho^2(Q - \beta^2/R^2) = K$, where $R \geqslant 1$, $K > 0$, $0 < \beta^2 < 2/(\gamma-1)$, K, and β are prescribed constants, and where $\rho^2 = (1 - \frac{1}{2}(\gamma-1)Q)2/(\gamma-1)$.

Probelm II: Initial value problem: Find C^1 solutions $Q^{\pm}(R)$ with prescribed β, $0 < \beta^2 < 2(\gamma-1)$, with prescribed initial data $Q^+(1) > \hat{Q}(1)$ or $Q^-(1) < \hat{Q}(1)$ satisfying

$$R^2\rho^2(Q - \beta^2/R^2) = K = \rho^2[Q^{\pm}(1) - \beta^2]. \tag{5.1}$$

These problems are equivalent. More precisely, we have:

Proposition 5.1: Solutions of I satisfy II and vice-versa provided that $K = \rho^2(Q^{\pm}(1))(Q^{\pm}(1) - \beta^2)$ in I.

Which follows from:

Proposition 5.2: Solutions of either I or II satisfy the algebraic mass flow relation

$$(MF) \quad R^2\rho^2(Q^{\pm})(Q^{\pm} - \beta^2/R^2) = K,$$

and C^1 solutions to the relation (MF) satisfy problems I and II for noncritical K and noncritical initial value $Q^{\pm}(1)$ such that $\rho^2(Q^{\pm}(1))(Q^{\pm}(1) - \beta^2) = K$.

Proof: The relation $\rho^2(Q^{\pm}(1))(Q^{\pm}(1) - \beta^2) = K$ is simply the mass flow relation (MF) when $R = 1$. The two problems are equivalent since Problem II is a Problem I with this relation used to replace K by $Q^{\pm}(1)$.

6. LOCAL EXISTENCE THEORY

We seek a local solution $Q = Q(R)$ of the initial value problem:

$$R^2(Q - \beta^2/R^2)\rho^2 = K \quad \text{on } [R_1, R_2] \subset [1, \infty),$$

K a positive constant, $Q(R_1) = Q_1$, and $\beta^2/R_1^2 < Q_1 < 2/(\gamma-1)$.

Definition: $\hat{Q}(R) = [2/(\gamma+1)](1 + \beta^2/R^2)$ is called the critical curve. A solution that lies above $\hat{Q}$ is called supercritical and is denoted by Q^+. A solution that lies below $\hat{Q}$ is called subcritical and is denoted by Q^-.

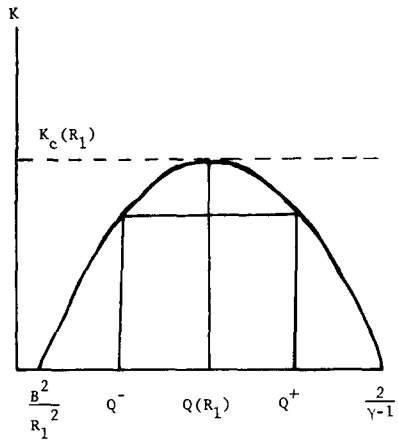

FIG. 5. Dependence of K on Q.

Theorem 6.1: There exists a critical value of K, $K = K_c$, such that for given R_1, Q_1, and K, where $0 \leqslant K < K_c(R_1)$, there is an interval $[R_1, R_2]$ in which there exist two local C^1 solutions Q^+ and Q^- of the initial value problem above.

Moreover, Q^+ and Q^- satisfy:

(i) $K = R^2(Q^\pm - \beta^2/R^2)[1 - \frac{1}{2}(\gamma-1)Q^\pm]2/(\gamma-1)$;

(ii) $\beta^2/R^2 < Q^- < \hat{Q} < Q^+ < 2/(\gamma-1)$;

(iii) (a) Q^+ (resp. Q^-) is a monotone increasing (decreasing) function of K, (b) K increases monotonically as a function of increasing Q^-, (c) K decreases monotonically as a function of increasing Q^+; (6.1)

(iv) $Q^- \nearrow \hat{Q}$ and $Q^+ \searrow \hat{Q}$ as $K \to K_c$, see Fig. 5;

(v) $K_c(r_1) \leqslant K(r_2)$ if $r_1 \leqslant r_2$;

(vi) if $Q^\pm(R)$ satisfying $Q^\pm(R_1) = Q_1$ can be continued to $R = R_2$ then $Q(R_2)$ is an increasing function of $Q(R_1)$.

Proof: These results are proved for toroidal flow in Ref. 1, p. 372. Our theorem follows identically with the substitution $R = f$. The proofs in Ref. 4 follow by application of the implicit function theorem to the function

$$F(R,Q) = R^2(Q - \beta^2/R^2)\rho^2 - K;$$

thus,

$$E_Q(R_1,Q) = R_1^2[1 - \tfrac{1}{2}(\gamma-1)Q_1](3-\gamma)/(\gamma-1) \quad (6.2)$$

and

$$\frac{dQ}{dR} = \frac{-F_R}{F_Q} = \frac{-8}{\gamma+1}\,\frac{1}{R}\left\{\frac{Q[1 - \frac{1}{2}(\gamma-1)Q]}{\hat{Q} - Q}\right\}.$$

The conclusions follow from arithmetic consideration of these expressions. Note that dQ/dR is infinite if $Q = \hat{Q}$. This is the analytic meaning of $\hat{Q}$.

Remark: (vi) tells us that the local solutions are monotone increasing as functions of their initial values. In other words, if $Q_a^+(1) \geqslant Q_b^+(1)$, then $Q_a^+(r) \geqslant Q_b^+(r)$ for any $r \geqslant 1$. When we have found that global solutions of the mass flow–circulation problem (vi) will also tell us that these solutions $Q^\pm$ are also monotone increasing functions of their initial data.

The formula for dQ/dR above tells us that each solution Q^+ is monotone increasing and that each solution Q^- is monotone decreasing as functions of the radius R. The same facts hold for global solutions of the mass flow–circulation problem.

The considerations of this remark are an important tool in the proof of convergence to the critical solutions of Sec. 9.

7. SOME LOCAL LEMMAS

We state some necessary technical lemmas that are almost identical to the corresponding technical lemmas of Ref. 1. Indeed the proofs in Ref. 1 apply to our case with just a change in notation and domain. We need these technical lemmas in the global existence theory.

Recall Problems I and II of Sec. 5. Let $R \geqslant 1$ as usual.

Proposition 7.1: Let $Q^\pm(R)$ be a solution of the initial value problem II on some interval $[R_{\min}, R_{\max}]$ for which $\varphi(R) = [Q(R) - Q^\pm(R)]^2$ is positive. Then $\varphi(R)$ has a unique minimum at $R = R_{\min}$.

The proof of this proposition follows from this lemma:

Lemma 7.1: (a) The function $(\hat{Q} - Q^\pm)^2$ satisfies the differential equation

$$\frac{d}{dR}[\hat{Q} - Q^\pm]^2 = \frac{8}{R} g(R, Q^\pm(R)), \quad (7.1)$$

where

$$g(R,Q^\pm(R)) = [1/(\gamma+1)]\{Q^\pm(R)[1 - \tfrac{1}{2}(\gamma-1)Q^\pm(R)] + (\beta^2/R^2)[Q^\pm(R) - \hat{Q}(R)]\}. \quad (7.2)$$

(b) If $[\hat{Q}(R) - Q(R)] > 0$ on an interval $[R_{\min}, R_{\max}]$, then $g(R,Q(R)) > C(K) > 0$, where

$$C(K) = \frac{2}{(\gamma+1)^2}\left(1 + \frac{\beta^2}{R_{\max}^2}\right) \times \left[\min \frac{K}{R_{\max}^2}, \left(\frac{\gamma-1}{2}\,\frac{K}{R_{\max}^2}\right)^{2/(\gamma-1)}\right] \quad (7.3)$$

is a monotone increasing function of the mass flow constant K.

Proof: With the substitution $R = f$ and by replacing the torus with the interval $[R_{\min}, R_{\max}]$ as domain, the proof is identical to that of Lemma 5.2 of Ref. 4.

Proof of Proposition 7.1: In Ref. 1 the second derivative test was used, but here it does not apply since the minimum now occurs on the left end point.

We see this because the derivative of the function in question is positive and this function is continuous on a compact set. Here, the minimum occurs on the left end point.

Corollary 7.1: If Q^- is a solution of Problem II in some interval $[R_{\min}, R_{\max}]$ in which $Q^+(R) - Q^-(R) > 0$, then

$$Q^+(R) - \hat{Q}(R) > Q^+(R_{\min}) - \hat{Q}(R_{\min}) > 0. \quad (7.4)$$

We also have:

Lemma 7.2: Let $N = [R_{\min}, R_{\max}]$ be an interval contained in $[1, \infty)$. There exists a unique C^1 subcritical (resp. critical) solution Q^- (resp. Q^+) of Problem II (equivalently of Problem I) with noncritical data on N.

Proof: This follows just like Theorem 5.1 in Ref. 1 with the substitution of N for the torus and R for f.

Remark: The above supercritical solutions Q^+ are monotone increasing, and the above subcritical solutions Q^- are monotone decreasing because of the differential equation above: The sign of the derivative of $Q^\pm$ depends on whether $\hat{Q} - Q^\pm$ is positive or negative.

8. GLOBAL EXISTENCE THEORY FOR NONCRITICAL FLOWS

We establish the global existence and uniqueness of noncritical flow solutions of the mass flow–circulation problem. The method of proof is somewhat different from the method of Ref. 1 because of complications due to the noncompact nature of our domain.

We now state and prove the main theorem of this section.

Theorem 8.1: Let $0<\beta^2<2/(\gamma-1)$ and $0\leqslant K\leqslant K_c$, where K_c is the critical mass flow constant, corresponding to β. There exists a unique C supercritical (resp. subcritical) flow Q^+ (resp. Q^-) with mass flow constant K and circulation $2\pi\beta$ about the unit circle.

At no loss of generality, we carry out the proof in the supercritical Q^+ case.

Proof: First, we show uniqueness.

Suppose that we have two solutions. We show that the set S on which they are equal is nonempty and open and closed in the relative topology that S inherits as a subset of $[1,\infty)$. Then, because $[1,\infty)$ is connected, S is $[1,\infty)$.

We start by showing that S is nonempty. Any solution of the mass flow–circulation problem (Problem I) is also a solution of the initial value problem (Problem II) with initial data $Q(1)$ determined by K. Since both solutions have the same mass flow constant K, they have the same initial value $Q(1)$. Thus they agree at $R=1$, and S is nonempty.

We now show that S is relatively closed and open. Any solution of the mass flow–circulation problem must satisfy the mass flow condition (MF) at every point of $[1,\infty)$. This condition is a continuous algebraic functional relation on Q. Thus S is closed. The implicit function theorem shows that S is open.

Thus S is $[1,\infty)$, and the two solutions are equal—uniqueness is proved.

Remark: We used the noncritical nature of K, when we envoked the implicit function theorem. Because K is noncritical dF/dQ is nonzero and noninfinite. See Sec. 6 for the formula giving DF/dQ.

We now show existence by construction. Let $N=[1,R]$ and $M=[1,\tilde{R}]$ be subintervals of $[1,\infty)$ such that $N\subset M$. The local existence theorem (Theorem 6.1) gives us C^1 solutions ${Q^+}_N$ and ${Q^+}_M$ of Problems I and II on N and M. We show that ${Q^+}_M$ is the unique continuous extension of Q^+ to M that is a solution of Problems I and II on M.

Let S be the set of points where ${Q^+}_N$ and ${Q^+}_M$ agree. Clearly $S\subset N$. We show that S is nonempty, and also closed and open in the relative topology that N inherits as a subset of M. The proof is very similar to the uniqueness proof above.

S is nonempty because ${Q^+}_N$ and ${Q^+}_M$ have the same initial value at $R=1$ as solutions of Problem II. S is closed because ${Q^+}_N$ and ${Q^+}_M$ both satisfy the algebraic mass flow relation (MF), which is a continuous algebraic functional relation on Q. Finally, S is open by the implicit function theorem since K is noncritical and $0<dF/dQ<\infty$.

Thus $S=N$ and ${Q^+}_M$ is the unique continuous extension of ${Q^+}_N$ as a solution of Problem I.

Now let $\tilde{R}\to\infty$. The local solution ${Q^+}_M$ tends to a global solution Q^+ of Problem I. QED

The argument for Q^- is identical.

Corollary 8.1: Q^+ is monotone increasing as a function of R with limiting speed $2/(\gamma-1)$ at infinity. Q^- is monotone decreasing as a function of R with limiting speed zero at infinity.

Proof: The monotonicity follows from the sign of $dQ^\pm/dr$ (see Sec. 6), now that we know that Q^+ and Q^- exist. We see the limiting speed behavior by looking at the algebraic mass flow relation (MF). Both Q^+ and Q^- must satisfy (MF). Let $R\to\infty$. Because K/R^2 then goes to zero, either $\rho(Q_\infty)$ or Q_∞ must vanish.

Since Q^+ is monotone increasing, this forces the limiting square speed to be the square cavatation speed $2/(\gamma-1)$, similarly, because Q^- is monotone decreasing its limiting speed at infinity must be zero. QED

We also have two more corollaries.

Corollary 8.2: The global C^1 solutions $Q^\pm$ satisfy the differential equation

$$\frac{d}{dR}[\hat{Q}-Q^\pm]^2=\frac{8}{R}g(R,Q(R)), \tag{8.1}$$

where g is the same as it was in Lemma 7.1.

Proof: $Q^\pm$ exist. The conclusion of this corollary is a local condition on $Q^\pm$ which was proved in Lemma 7.1. QED

Corollary 8.3: The global C^1 solutions $Q^\pm$ satisfy the integral equation:

$$Q^\pm=\hat{Q}\pm\left\{[\hat{Q}(1)-Q(1)]^2+\int_1^R(8/t)g(t,Q(t))\,dt\right\}^{1/2}. \tag{8.2}$$

Proof: Integrate the differential equation of the previous corollary. QED

9. CONVERGENCE TO THE CRITICAL SOLUTIONS

We show that as K approaches its critical value K_c, the solutions Q^+ and Q^- approach limiting functions Q_c^+ and Q_c^- that solve the mass flow–circulation problem when $K=K_c$.

Slightly abusing terminology, we call Q_c^+ and Q_c^- subcritical. The two critical solutions Q_c^+ and Q_c^- satisfy the critical mass flow relation:

$$\text{(MFC)}\quad K_c=R^2(Q_c^\pm-\beta^2/R^2)\rho^2(Q_c^\pm). \tag{9.1}$$

Q_c^+ is monotone increasing as a function of R with a limiting square speed $2(\gamma-1)$ at infinity, Q_c^- is monotone decreasing as a function of R with a limiting square speed of zero at infinity, and Q_c^+ and Q_c^- bifurcate at $R=1$ with vertical slope. See Fig. 4.

Although the critical solutions $Q_c^\pm$ satisfy the algebraic relation (MFC), we cannot prove local existence using the implicit function theorem alone because the required derivative is infinite at $R=1$ (compare Sec. 6). We prove local and global existence and uniqueness using convergence arguments.

These convergence arguments construct $Q_c^\pm$ as the limit of noncritical solutions $Q^\pm$ as K approaches its critical

value $K = K_c$. The convergence arguments are more delicate than one might at first suspect because our domain is *not* compact, and thus we do the convergence arguments in careful detail.

We now state and prove the theorem.

Theorem 9.1: There exist unique solutions Q_c^+ and Q_c^- to the mass flow–circulation problem with critical mass flow $K = K_c$.

These solutions are C^1 when $R > 1$ and satisfy:

(a) Q_c^+ is monotone increasing as a function of R with limiting square speed $2/(\gamma - 1)$ at infinity.

(b) Q_c^- is monotone decreasing as a function of R with zero limiting speed at infinity.

(c) $Q_c^+ > \hat{Q}$ when $R > 1$ (we say then that Q^+ is supercritical) and $Q_c^- < \hat{Q}$ when $R > 1$ (we say then that Q^- is subcritical).

(d) Q_c^+ and Q_c^- bifurcate from $R = 1$ with infinite slope at $R = 1$.

(e) Let Q_K^+ denote solutions of the mass flow–circulation problem with noncritical mass flow K. Then as K approaches the critical value $K = K_c$, $Q_K^\pm$ approach $Q_c^\pm$ pointwise. In fact, $Q_K^+ \searrow Q_c^+$ and $Q_K^- \nearrow Q_c^-$ uniformly on any compact subinterval of $[1, \infty)$.

(f) Q_c^+ and Q_c^- also solve the initial value problem (Problem II) with *critical* initial value $Q_c(1) = \hat{Q}_c(1)$.

(g) Let $Q_{Q_1}^\pm$ denote solutions of the initial value problem with noncritical initial value $Q(1) = Q_1$. Then, as the initial values Q_1 approach the critical initial value $Q_c(1) = \hat{Q}(1)$, $Q_{Q_1}^\pm$ approach $Q_c^\pm$. In fact, $Q_{Q_1}^+ \searrow Q_c^+$ and $Q_{Q_1}^- \nearrow Q_c^-$ and the convergence is uniform on any compact subinterval of $[1, \infty)$.

(h) The critical solutions satisfy the critical mass flow relation

$$K_c = R^2(Q_c^\pm - \beta^2/R^2)\rho^2(Q_c^\pm)). \tag{9.2}$$

(i) They satisfy the integral equation

$$Q_c^\pm(R) = \hat{Q}(R) \pm \left\{ [\hat{Q}_c(1) - Q_c^\pm(1)]^2 + \int_1^R g(t, Q^{\pm c}(t))\, dt \right\}^{1/2}. \tag{9.3}$$

(j) They have derivatives for $R > 1$ given by

$$\frac{dQ_c^\pm}{dR} = \frac{d\hat{Q}}{dR} - \frac{1}{2}\frac{8}{R} g(RQ_c^\pm(R)) \times \left\{ [\hat{Q}_c(1) - Q_c^\pm(1)]^2 + \int_1^R (8/t) g(t, Q^\pm(t))\, dt \right\}^{-1/2}. \tag{9.4}$$

Proof: We prove the theorem for Q^+. The proof for Q^- is identical.

Proof of Uniqueness: Let $Q_{c,1}^+$ and $Q_{c,2}^+$ be C^1 (when $R > 1$) global solutions of the mass flow–circulation problem with critical mass flow constant $K = K_c$. Then, $Q_{c,1}^+$ and $Q_{c,2}^+$ satisfy the critical mass flow relation

$$\text{(MFC)} \quad [Q_c^+(R) - \beta^2/R]\rho^2(Q_c^+)R^2 = K_c.$$

Let S be the set of points where $Q_{c,1}^+ = Q_{c,2}^+$. Then because these solutions have the same initial value, S is nonempty (i.e., $K = 1 \in S$). Because (MFC) is algebraic and thus continuous as a function of Q_c^+, the set S is closed. When $R > 1$, the initial value theorem implies the local solvability of the relation (MFC) and thus S is relatively open in $[1, \infty)$. Therefore, S is all of $[1, \infty)$ and $Q_{c,1}^+ = Q_{c,2}^+$. Now, we prove the existence of a global critical solution that is C^1 when $R > 1$.

We now construct the supercritical solution Q_c^+.

Consider a sequence $(Q_{K_n}^+)$ of solutions to Problem I corresponding to a sequence of noncritical mass flow values K_n (for β fixed!) converging to the critical mass flow value K_c. Since K determines $Q(1)$ from the mass flow relation at $R = 1$, the remark at the end of Sec. 6 (and Theorem 6.1) imply that at each point x of $[1, \infty)$ the sequence $(Q_{K_n}^+)$ is monotone decreasing and bounded from below (by zero); hence it has a limit point $Q_c^+(x)$. Thus the function sequence $(Q_{K_n}^+)$ converges pointwise to a limit function Q_c^+. Moreover, a standard interlacing sequence argument shows that the limit function Q_c^+ is independent of the choice of the sequence (K_n).

The limit function Q_c^+ is continuous, as can be easily proved by a standard "three epsilon" argument. More is true: Because the sequence $(Q_{K_n}^+)$ is monotone, Dini's theorem (Ref. 6, p. 248, Ex. 9.9) assures us that the convergence is *uniform* on any compact subset of $[1, \infty)$.

Each member $(Q_{K_n}^+)$ of the sequence satisfies the mass flow relation

$$K_n = R^2(Q_{K_n}^+ - \beta^2/R^2)\rho^2(Q_K^+). \tag{9.5}$$

Because ρ is continuous as a funciton of Q, the sequential continuity of this relation implies that Q_c^+ satisfies the critical mass flow relation

$$\text{(MFC)} \quad K_c = R^2(Q_c^+ - \beta^2/R^2)\rho^2(Q_c^+), \tag{9.6}$$

which proves (h).

We now show that Q_c^+ is C^1 when $R > 1$. We do this by computing dQ_c^+/dR. Along the way, we establish (i) and (j).

By Corollary 8.3 each element $(Q_{K_n}^+)$ of the sequence globally satisfies the integral equation

$$Q_{K_n}^+(R) = \hat{Q}(R) + \left\{ [\hat{Q}(1) - Q_{K_n}^+(1)]^3 + \int_1^R (8/t) g(t, Q_{K_n}^+)\, dt \right\}^{1/2}, \tag{9.7}$$

where g is given by

$$g(R, Q(R)) = [1/(\gamma + 1)]\{Q(R)[1 - \tfrac{1}{2}(\gamma - 1)Q(R)] + (\beta^2/R^2)[Q(R) - \hat{Q}(R)]\}. \tag{9.8}$$

Because $g(R,Q)$ is continuous as a function of Q, and because the convergence of $Q_{K_n}^+$ to Q_c^+ is uniform on compact sets by Dini's theorem, we have that Q_c^+ satisfies the integral equation

$$Q_c^+ = \hat{Q}(R) + \left\{ [\hat{Q}(1) - Q_c^+(1)] + \int_1^R (8/t) g(t, Q_c^+)\, dt \right\}^{1/2}. \tag{9.9}$$

Now by the fundamental theorem of calculus we can differentiate this relation at any interior point of $[1, \infty)$ to obtain

$$\frac{dQ_c^+}{dR} = \frac{d\hat{Q}}{dR} - \frac{\frac{1}{2}[(8/R)g(R,Q_c^+(R))]}{\left\{[\hat{Q}(1) - Q_c^+(1)]^2 + \int_1^R (8/t)g(t,Q_c^+)\,dt\right\}^{1/2}}. \tag{9.10}$$

Thus Q_c^+ is C^1 if $R > 1$. These last two equations have many consequences.

From the equation for Q_c^+ we see that $\hat{Q}(1) = Q_c^+$ [and similarly $\hat{Q}(1) = Q_c^-$]; thus $Q_c^+(1) = Q_c^-(1) = Q(1)$, showing that Q_c^+ and Q_c^- bifurcate from $R = 1$. Also from this, we see that Q_c^+ solves initial value problem II as well as mass flow–circulation problem I.

From the above expression for dQ_c^+/dR we see that Q_c^+ is monotone increasing as a function of R (similarly Q_c^- is monotone decreasing as a function of R) and also that dQ_c^+/dR is infinite when $R = 1$.

From the integral equation above for Q_c^+ it follows by algebra that:

$$\frac{d}{dR}(\hat{Q} - Q_c^+)^2 = \frac{8}{R}g(R,Q_c^+), \tag{9.11}$$

where

$$g(R,Q_c^+(R)) = [1/(\gamma-1)] \times \{Q_c^+(R)[1 - \tfrac{1}{2}(\gamma-1)Q_c^+(R)] + (\beta^2/R^2)[Q_c^+(R) - \hat{Q}(R)]\}, \tag{9.12}$$

from which, repeating the proof of Corollary 8.1, it follows that Q_c^+ is supercritical. Similarly, Q_c^- is subcritical.

Finally the critical mass flow relation implies that at infinity $Q^+ = 2/(\gamma-1)$ (the square cavitation speed) and $Q^- = 0$. QED

10. THE CASE OF $\gamma = 3$

When $\gamma = 3$, the mass flow relation becomes quadratic and $Q^\pm$ satisfy a quadratic equation. Compare Ref. 1.

In this case we have

$$Q^{\pm 2} - 2\hat{Q}(R)Q^\pm + (\beta^2 + K)/R^2 = 0, \tag{10.1}$$

and $Q^\pm$ satisfy

$$Q^\pm = -\hat{Q}(R) \pm [\hat{Q}^2(R) - (\beta^2 + K)/R^2]^{1/2}, \tag{10.2}$$

where $\hat{Q}(R) = \frac{1}{2}(1 + \beta^2/R^2)$. The discriminate vanishes when $K = K_c$ and $R = 1$, which shows that $K_c = \frac{1}{4}(1+\beta^2)^2 - \beta^2$.

11. SHOCK SOLUTIONS

Consider Figs. 2 and 3. Each figure shows a flow that starts out on the Q^+ curve and drops to the Q^- curve at $R = R_s$. Such a flow is a possible solution of the mass flow–circulation problem because Q^+ and Q^- have the same mass flow constant K.

These solutions ae called *shocks*. In Ref. 1 such solutions also occurred. There because of the periodic nature of the flow, it was shown that shocks were possible only for critical mass flow and indeed must occur there.

In our problem the flow is not periodic, so shocks might also occur if K is not critical. See Fig. 3. However, they might not occur at all. In fact, when $\gamma = 3$, we give a proof that shocks do not exist.

The proof is based on the Prandle–Rankine–Hugonant condition below for shocks in a polytropic gas. See Ref. 3.

Prandle–Rankine–Hugonant condition

Let a shock occur at $R = R_s$. Let V_{n1} be the velocity normal to the shock ahead of the shock and let V_{n2} be the velocity normal to the schock behind the shock. Then:

$$V_{n1}V_{n2} = 2/(\gamma+1) \tag{11.1}$$

We now show:

Theorem 11.1: Let $\gamma = 3$. There are no shock solutions to the mass flow–circulation problem.

Proof: We first show that the shock is oblique and normal to R at R_s.

Since $Q^+(R_s)$ and $Q^-(R_s)$ share the same value of β, the θ-velocity is invariant across the shock. So, the θ direction is tangent to the shock. See Fig. 6.

We now show that $R_s < 1$, which proves shocks are impossible since our flows are exterior to the unit circle.

Recall that when $\gamma = 3$, Q^+ and Q^- satisfy the quadratic equation

$$Q^{\pm 2} - (1 + \beta^2/R^2)Q^\pm + (\beta^2 + K)/R^2 = 0. \tag{11.2}$$

Let

$$Q^+(R_s) = Q_s^+, \quad \alpha^+(R_s) = \alpha_s^+, \tag{11.3}$$

$$Q^-(R_s) = Q_s^-, \quad \alpha^-(R_s) = \alpha_s^-. \tag{11.4}$$

Then,

$$Q_s^{\pm 2} - (1 + \beta^2/R_s^2)Q_s^\pm + (\beta^2 + K)/R_s^2 = 0. \tag{11.5}$$

Thus

$$Q_s^+ + Q_s^- = 1 + \beta^2/R_s^2, \tag{11.6}$$

$$Q_s^+ Q_s^- = (\beta^2 + K)/R_s^2. \tag{11.7}$$

In terms of α_s^+ and α_s^+ we have

$$\alpha_s^{+2} + \alpha_s^{-2} = [Q_s^+ - \beta^2/R_s^2] + (Q_s^- - \beta^2/R_s^2) \tag{11.8}$$

or

$$\alpha_s^{+2} + \alpha_s^{-2} = 1 - \beta^2/R_s^2 \tag{11.9}$$

and also

$$(\alpha^{+2} + \beta^2/R_s^2)(\alpha^{-2} + \beta^2/R_s^2) = (\beta^2/R_s^2) = (\beta^2 + K)/R_s^2 \tag{11.10}$$

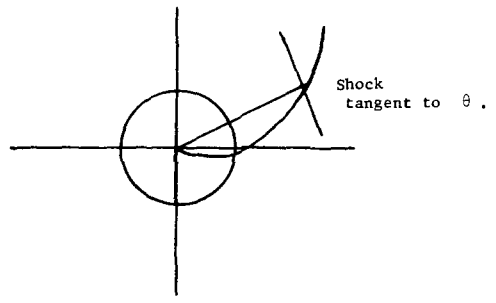

FIG. 6. Oblique shock.

or, equivalently,

$$\alpha_s^{+2}\alpha_s^{-2} + (\alpha_s^{+} + \alpha_s^{-2})\beta^2/R_s^2 + (\beta^2/R_s^2)^2 = (\beta^2 + K)/R_s^2. \quad (11.11)$$

Since $\alpha^{+2} + \alpha^{-2} = 1 - \beta^2/R_s^2$, we have

$$\alpha_s^{+2}\alpha_s^{-2} + (1 - \beta^2/R_s^2)\beta^2 R_s^2 + (\beta^2/R_s^2)^2 = (\beta^2 + K)/R_s^2. \quad (11.12)$$

Consider $\alpha_s^{+2}\alpha_s^{-2}$. Because the shock is normal to R at $R = R_s$, $\alpha_s^{+} = V_{n1}$, and $\alpha_s^{-} = V_{n2}$. Thus by the Prandle–Randkine–Hugonant condition above, $\alpha_s^{+2}\alpha_s^{-2} = 2/(\gamma + 1) = \frac{1}{2}$.

Thus we have

$$\tfrac{1}{2} + (1 - \beta^2/R_s^2)\beta^2/R_s^2 + (\beta^2/R_s^2)^2 = (\beta^2 + K)/R_s^2. \quad (11.13)$$

At present let $K = K_c$. When $\gamma = 3$, we showed in the previous section that

$$\beta^2 + K_c = \tfrac{1}{4}(1 + \beta^2). \quad (11.14)$$

The last two equations give the equation below for R_s.

$$\tfrac{1}{2}R_s^4 + (\tfrac{3}{4}\beta^2 - \tfrac{1}{4})R_s^2 + (\beta^2 - \beta^4) = 0, \quad (11.15)$$

which has roots

$$R_s^2 = (\tfrac{1}{4} - \tfrac{3}{4}\beta^2) \pm \sqrt{(\tfrac{1}{4} - \tfrac{3}{4}\beta^2)^2 - 2(\beta^2 - \beta^4)}. \quad (11.16)$$

Thus,

$$R_s^2 < 1 \quad (\text{when } \beta = 0,\ R_s = 1/\sqrt{2}). \quad (11.17)$$

So, if $K = K_c$, shocks do not occur. Now if $K \neq K_c$, then $K < K_c$ and $K + \beta^2 < \beta^2 + K_c < \frac{1}{4}(1 + \beta^2)$. Replacing the quadratic above by a quadratic inequality proves that again $R_s < 1$. QED

We have thus proved that if $\gamma = 3$, shocks do not occur.

When $\gamma \neq 3$, the author conjectures that shocks also do not occur. A proof proabably would involve Prandle's condition and careful estimates based on the integral equation for $Q^{\pm}$.

12. CONCLUSION

We have demonstrated a family of smooth transonic flows in the plane. The method also can be used to analyze other transonic plane flows (e.g., Ringleb flow[3]), previously incorrectly treated by the Hodograph method. In addition, the author has also treated certain three-dimensional flows by this method (e.g., pipe flow), and here too, smooth transonic families occur.

[1] L. M. Sibner and R. J. Sibner, "Transonic Flow on an Axially Symmetric Torus," J. Math. Anal. (to be published).

[2] L. Bers, *Mathematical Aspects of Subsonic and Transonic Gas Dynamics* (Wiley, New York, 1958).

[3] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves* (Interscience, New York, 1948).

[4] L. M. Sibner, and R. J. Sibner, "Non-Linear Hodge–de-Rham Theorem," Acta Math. **125**, 57–73 (1970).

[5] L. M. Sibner, and R. J. Sibner, "Non-Linear Hodge Theory: Applications," Adv. Math. **31**, 1–15 (1979).

[6] T. Apostol, *Mathematical Analysis* (Addison-Wesley, Reading, Mass., 1974), 2nd ed.

Symmetry of the complete second-order conductivity tensor in a Vlasov plasma

Jonas Larsson
Department of Plasma Physics, Umeå University, S-901 87 Umeå, Sweden

(Received 22 June 1983; accepted for publication 2 September 1983)

This paper has two purposes. The first is to consider the origin of a recently derived symmetry property including the pole contributions of the second-order conductivity. The second is to show how certain general formulas for the conductivities easily lead to much more convenient expressions than those used in the above-mentioned derivation of the symmetry.

PACS numbers: 52.25.Fi, 02.30. + g

The second-order conductivity tensor for a plasma described by the Vlasov–Maxwell equations can be expressed in terms of an integral involving poles due to resonant wave-particle interaction. The nonresonant particles determine the principal part of the integral while the resonant particles give pole contributions. Neglecting the pole contributions, we obtain the very well-known symmetry leading to the Manley–Rowe relations. Recently a symmetry relation was found[1] involving also the pole contributions. The derivation was a straightforward but lengthy calculation resulting in a very extensive formula for the second-order conductivity tensor in an unmagnetized relativistic plasma.

In this paper we observe that previously derived[2–4] general formulas for the conductivities directly lead to symmetries involving both the principal parts and the pole contributions. The symmetries are valid for a relativistic plasma also in the magnetized case. It will be shown below that the symmetry in Ref. 1 is included.

It is convenient to consider the quantities V, related to the second-order conductivity as[4]

$$\mathbf{V}(0,1,2)\equiv V(k_0,\mathbf{E}_0,k_1,\mathbf{E}_1,k_2,\mathbf{E}_2) = (2i/\omega_0)\mathbf{E}_0\cdot\boldsymbol{\sigma}^{(2)}_{k_1,k_2}(\mathbf{E}_1,\mathbf{E}_2), \tag{1a}$$

where $k_j = (\omega_j,\mathbf{k}_j)$, $j = 0,1,2$ and

$$\omega_0 + \omega_1 + \omega_2 = 0, \quad \mathbf{k}_0 + \mathbf{k}_1 + \mathbf{k}_2 = 0 \tag{1b}$$

and $\mathbf{E}_j$ are arbitrary vectors used as arguments in (1a). Now V may be writen as[4]

$$V(0,1,2) = \int f_0(\mathbf{v})A(k_0,\mathbf{E}_0,k_1,\mathbf{E}_1,k_2,\mathbf{E}_2,\mathbf{v})d^3v \equiv \int f_0(\mathbf{v})A(0,1,2\mathbf{v})d^3v, \tag{2}$$

where A is symmetric in the indices (0,1,2), i.e.,

$$A(0,1,2,\mathbf{v}) = A(\alpha,\beta,\gamma,\mathbf{v}) \quad \text{for } \{\alpha,\beta,\gamma\} = \{0,1,2\}. \tag{3}$$

There are, however, poles in the integrand of (2) due to denominators $(\omega_j - \mathbf{k}_j\cdot\mathbf{v})$ for an unmagnetized plasma and $(\omega_j - k_{jz}v_z - n\omega_c)$ for a magnetized plasma. These poles must be treated properly. Let us introduce operators P and R_j, where P stands for the principal part and R_j stands for pole contribution of the denominators containing ω_j mentioned above treated according to the prescription $\omega_j + i\eta$, $\eta \to 0+$. Then we have[4]

$$V(0,1,2) = PV - R_0V + R_1V + R_2V. \tag{4}$$

In (4) we have for brevity not indicated the arguments on the right-hand side since the symmetry (3) implies

$$PV(0,1,2) = PV(\alpha,\beta,\gamma), \quad R_jV(0,1,2) = R_jV(\alpha,\beta,\gamma) \quad \text{for} \quad \{\alpha,\beta,\gamma\} = \{0,1,2\}. \tag{5}$$

It is clear from (4) that $V(0,1,2)$ does not have the corresponding symmetry. A calculation of $V(0,1,2)$ naturally means a calculation of each term in (4). Then we have determined not only $V(0,1,2)$ but also all $V(\alpha,\beta,\gamma)$, where $\{\alpha,\beta,\gamma\} = \{0,1,2\}$. More substantial symmetries may be obtained in situations where some of the pole contributions may be neglected.

Considering resonant wave interaction between two high-frequency waves k_0 and k_1 with the low-frequency wave k_2, we may sometimes take $R_0V = R_1V = 0$. Then $V(0,1,2) = V(1,0,2) = PV + R_2V$, while $V(2,0,1) = PV - R_2V$. The coupled mode equations, in which we now may omit the linear damping of wave 0 and 1, are then simplified. Different particular forms of these equations are considered in Ref. 5.

Let us now compare with the symmetry result (26) in Ref. 1. We may write it in the form

$$V(0,1,2) = (P + R_1 + R_2)S(1,0,2) + (P + R_1 - R_0)S(1,2,0), \tag{6}$$

where S is related to the tensor S_{ijl} in Ref. 1 as

$$S(0,1,2)\equiv S(k_0,\mathbf{E}_0,k_1,\mathbf{E}_1,k_2,\mathbf{E}_2) = -iq(2\pi)^4(\omega_0\omega_1\omega_2)^{-1}S_{ijl}(-k_0,k_1,k_2)E_{0i}E_{1j}E_{2l}, \tag{7}$$

together with (1b). We also have

$$V(0,1,2) = S(0,1,2) + S(0,2,1). \tag{8}$$

It follows directly from (1) in Ref. 1 that

$$R_2S(1,0,2) = R_2V(1,0,2), \quad R_0S(1,2,0) = R_0V(1,2,0), \tag{9}$$

and we may thus rewrite (6) as

$$V(0,1,2) = PV(1,0,2) + R_1V(1,0,2) + R_2V(1,0,2) - R_0V(1,2,0). \tag{10}$$

But (10) follows directly from (4) and the derivation is thus completed.

 0022-2488/84/010185-02$02.50

We finally give a formula for the second-order conductivity tensor in an unmagnetized relativistic plasma. It is a particular case of the general formula[3] rewritten in more familiar notations and is clearly much more convenient to use than the formula which one obtains by straightforward calculations.[1] The result is

$$V(0,1,2) = -\frac{i}{m_0^2}\int F_0(\mathbf{v}) \times \frac{1}{(\omega_0 - \mathbf{k}_0\cdot\mathbf{v} - i\eta)(\omega_1 - \mathbf{k}_1\cdot\mathbf{v} + i\eta)(\omega_2 - \mathbf{k}_2\cdot\mathbf{v} + i\eta)} \times \left(\frac{\mathbf{k}_0\cdot\mathbf{F}_0 - (q\omega_0/c^2)\mathbf{v}\cdot\mathbf{E}_0}{\omega_0 - \mathbf{k}_0\cdot\mathbf{v} - i\eta}(\mathbf{F}_1\cdot\mathbf{F}_2 - q^2c^{-2}\mathbf{v}\cdot\mathbf{E}_1\mathbf{v}\cdot\mathbf{E}_2) + \text{even permutations of } (0,1,2) \right)(1 - \mathbf{v}^2/c^2)d^3v, \tag{11}$$

where

$$F_j = q\left(\mathbf{E}_j + \mathbf{v}\times\frac{\mathbf{k}_j\times\mathbf{E}_j}{\omega_j}\right) \quad \text{and } \eta \to 0+. \tag{12}$$

The property (3) is manifest in (11). The tensor S_{ijl} is explicitly obtained from (7) and (11) by substituting $\mathbf{E}_j = \hat{\mathbf{x}}_j$, where $(\hat{x}_0,\hat{x}_1,\hat{x}_2)$ are our orthonormal unit vectors.

The expression (11) is a good example of the usefulness of the general current response formulas[2,3] and it may simplify future application of them if we consider the notational change needed to obtain (11). From (2.11) and (2.13)–(2.15) in Ref. 3 we obtain in the notation of that paper

$$iu\cdot\kappa_\alpha\,\delta\tilde{x}_\alpha = \delta\tilde{u}_\alpha, \tag{13}$$

$$iu\cdot\kappa_\alpha\delta\tilde{u}_a = i(q/m_0c^2)\,\kappa_\alpha\Lambda\,\tilde{\phi}_\alpha\cdot u, \tag{14}$$

$$\tilde{\phi}_0\cdot\Lambda^{(2)}_{\kappa_1,\kappa_2}:\tilde{\phi}_1\,\tilde{\phi}_2 = \tfrac{1}{2}ic^3m_0\int f_0(u)[\kappa_0\cdot\delta\tilde{x}(0)\delta\tilde{u}(1)\cdot\delta\tilde{u}(2) + \text{even permutations of } (0,1,2)]\,du, \tag{15}$$

where $\kappa_\alpha = (\omega_\alpha/c)e_0 + \mathbf{k}_\alpha$ and $u = u^0(e_0 + \mathbf{v}/c)$ so that

$$u\cdot\kappa_\alpha = -(u^0/c)(\omega_\alpha - \mathbf{k}_\alpha\cdot\mathbf{v}), \tag{16}$$

where $u^0 = (1 - \mathbf{v}^2/c^2)^{-1/2}$.

In (15) the 4-vectors $\tilde{\phi}_\alpha$ are arbitrary. If we take $\tilde{\phi}_\alpha$ related to $\mathbf{E}_\alpha$ as the 4-potential is related to the electric field in Fourier space we obtain

$$\tilde{\phi}_0\cdot\Lambda^{(2)}_{\kappa_1,\kappa_2}:\tilde{\phi}_1\,\tilde{\phi}_2 = -(ic/\omega_0)\mathbf{E}_0\cdot\sigma^{(2)}_{\kappa_1,\kappa_2}[\mathbf{E}_1,\mathbf{E}_2] \tag{17}$$

and

$$\kappa_\alpha\Lambda\,\tilde{\phi}_\alpha\cdot u = -iu^0 \times\left(c^{-1}\mathbf{v}\cdot\mathbf{E}_\alpha\,e_0 + \mathbf{E}_\alpha + \mathbf{v}\times\frac{\mathbf{k}_\alpha\times\mathbf{E}_\alpha}{\omega_\alpha}\right). \tag{18}$$

Finally we need the relation between the distribution functions $f_0(u)$ and $F_0(\mathbf{v})$. The 4-current is

$$qc\int_S f_0(u)u\,du = \mathrm{q}\int(ce_0 + \mathbf{v})F_0(\mathbf{v})d^3v. \tag{19}$$

Taking the e_0-part of (19) we get the correspondence

$$f_0(u)u^0\,du = F_0(\mathbf{v})d^3v. \tag{20}$$

Or in more exact words, when we make the variable change $u \to \mathbf{v}$ defined by the relation $cu = u^0(ce_0 + \mathbf{v})$, then (20) is valid. From (1a), (13)–(18) and (20) we now obtain (11).

ACKNOWLEDGMENT

I thank the referee for checking all the equations above by rederiving them and for pointing out a number of typographical errors.

[1] H. E., Brandt, J. Math. Phys. **24**, 1332 (1983).
[2] J. Larsson, J. Math. Phys. **20**, 1321 (1979).
[3] J. Larsson, J. Math. Phys. **20**, 1331 (1979).
[4] J. Larsson, J. Plasma Phys. **21**, 519 (1979).
[5] J. Weiland and H. Wilhelmsson, *Coherent Non-Linear Interaction of Waves in Plasmas* (Pergamon, New York, 1977).

ERRATA

Erratum: Some remarks on the classical vacuum structure of gauge field theories [J. Math. Phys. 22, 179 (1981)]

M. Asorey
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, Spain

(Received 12 October 1983; accepted for publication 28 October 1983)

PACS numbers: 11.10.Np, 02.40.Vh, 99.10. + g

(1) Page 182 left column: Delete

$$\mathcal{A}_1^N U(1) = \{(\exp i\lambda_1,\ldots,\exp i\lambda_N)\in U(1)^N: \lambda_i\in[0, 2\pi), \lambda_1\leqslant\cdots\leqslant\lambda_{N-1}, (1/2\pi)\sum_{i=1}^{N}\lambda_i\in\mathbb{N}\},$$

and replace it by

$$\mathcal{A}_1^N U(1) = \{(\exp i\lambda_1,\ldots,\exp i\lambda_N)\in U(1)^N: \lambda_i\in[0, 2\pi), \lambda_1\leqslant\cdots\leqslant\lambda_N, (1/2\pi)\sum_{i=1}^{N}\lambda_i\in\mathbb{N}\}.$$

(2) Page 182 left column: Delete

$$\mathcal{V}^{SU(2)}_{\mathbb{R}^3\times S^1}\approx U(1)\approx\mathcal{V}^{SO(3)}_{\mathbb{R}^3\times S^1}\approx SO(2)\approx\mathcal{V}^{U(1)}_{\mathbb{R}^3\times S^1}\approx U(1),$$

and replace it by

$$\mathcal{V}^{SU(2)}_{\mathbb{R}^3\times S^1}\approx[0,\pi];\ \mathcal{V}^{SO(3)}_{\mathbb{R}^3\times S^1}\approx SO(2)\approx\mathcal{V}^{U(1)}_{\mathbb{R}^3\times S^1}\approx U(1).$$

Erratum: Splines and the projection collocation method for solving integral equations in scattering theory [J. Math. Phys. 24, 177 (1983)]

M. Brannigan
Department of Statistics and Computer Science, University of Georgia, Athens, Georgia 30602

D. Eyre
National Research Institute for Mathematical Sciences of the CSIR, P. O. Box 395, Pretoria 0001, Republic of South Africa

(Received 6 October 1983; accepted for publication 19 October 1983)

PACS numbers: 24.10. − i, 02.30.Rz, 25.10. + s, 02.60.Nm, 99.10. + g

1. The line after Eq. (2.1) should read "... space of continuous functions ... ."

2. Since the integral operator $\mathcal{K}$, containing the principal value integral, is not bounded on a space of continuous functions then our proof of convergence is not valid. However, convergence for this method is shown in our subsequent paper [J. Math. Phys. **24**, 1548 (1983)].

We are indebted to Ian H. Sloan for calling our attention to these points.

Erratum: Splines and the Galerkin method for solving the integral equations of scattering theory [J. Math. Phys. 24, 1548 (1983)]

M. Brannigan
Department of Statistics and Computer Sciences, University of Georgia, Athens, Georgia 30602

D. Eyre
National Research Institute for Mathematical Sciences of the CSIR, P. O. Box 395, Pretoria 0001, Republic of South Africa

(Received 6 October 1983; accepted for publication 19 October 1983)

PACS numbers: 03.80. + r, 02.30.Rz, 05.30.Jp, 99.10. + g

1. On page 1553 the scattering energy should read $(k/k_B)^2 = 0.64$.

2. Table II shows the square of the L 2-norm.

Decomposition of representations into basis representations for the classical groups[a)]

E. D'Hoker
Center for Theoretical Physics, Laboratory for Nuclear Science and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 24 February 1983; accepted for publication 24 June 1983)

We prove decomposition formulae for an arbitrary representation in terms of basis representations for the classical compact Lie groups. Using these decomposition formulae, simple rules are obtained for the product of two arbitrary representations and for the restriction of a representation to a classical subgroup.

PACS numbers: 02.20.Qs

I. INTRODUCTION

E. Cartan[1] has classified all simple compact Lie groups into four infinite sequences of classical groups SU($n+1$), SO($2n+1$), Sp(n), and SO($2n$) of rank n and in addition five exceptional groups, E_6, E_7, E_8, F_4, and G_2. Weyl[2] has shown that every finite-dimensional irreducible representation of a classical group is in one-to-one correspondence with a complex-valued function on the group, called the group character—or simply character—of the representation. If $\lambda(g)$ is a representation, then the character χ_λ is the trace of $\lambda(g)$:

$$\chi_\lambda(g) = \operatorname{tr} \lambda(g). \tag{1.1}$$

It has the following properties:

$$\chi_\lambda(hgh^{-1}) = \chi_\lambda(g), \tag{1.2}$$

$$\chi_{\lambda\oplus\mu}(g) = \chi_\lambda(g) + \chi_\mu(g), \tag{1.3a}$$

$$\chi_{\lambda\otimes\mu}(g) = \chi_\lambda(g)\chi_\mu(g). \tag{1.3b}$$

The set of all characters forms a basis for the regular class functions on the group. Weyl[2] has also shown that the character functions of a classical group of rank n are classified by n nonnegative integers. In addition, for the orthogonal groups, there are the so-called double-valued or spinor representations, which are specified by n half-odd integers. Since the character functions are invariant under conjugation—property (1.2)—they may be completely reconstructed from their value on a Cartan subgroup. Weyl's *first* formula gives the character in terms of n angles $\phi_1, \phi_2, \ldots, \phi_n$, which parametrize the Cartan subgroup in the standard fashion.[2,3] We record Weyl's first formulae[2] here for later reference. We shall henceforth suppress the argument of the character function.

(A) SU(n):

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1},\ldots,\epsilon^{l_n}|}{|\epsilon^{l_1^0},\ldots,\epsilon^{l_n^0}|}. \tag{1.4a}$$

Here we use the definition $|\epsilon^{l_1},\ldots,\epsilon^{l_n}| = \det(E)$, with $E_{ij} = \epsilon_i^{l_j}$ where $\epsilon_i = e^{i\phi_i}$, $l_i^0 = n - i$, $l_i = f_i + l_i^0$ and the integers f_i obey $f_1 \geqslant f_2 \geqslant f_3 \geqslant \cdots \geqslant f_n$ and $(f_1+1, f_2+1, \cdots, f_{n+1}) \equiv (f_1, f_2, \cdots, f_n)$.

(B) SO($2n+1$):

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1}-\epsilon^{-l_1},\ldots,\epsilon^{l_n}-\epsilon^{-l_n}|}{|\epsilon^{l_1^0}-\epsilon^{-l_1^0},\ldots,\epsilon^{l_n^0}-\epsilon^{-l_n^0}|}. \tag{1.4b}$$

Here the l_i^0 are half-integers given by $l_i^0 = n - i + \frac{1}{2}$ and $l_i = f_i + l_i^0$ with f_i either all integers or all half-odd-integers and $f_1 \geqslant f_2 \geqslant \cdots \geqslant f_n \geqslant 0$.

(C) Sp(n):

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1}-\epsilon^{-l_1},\ldots,\epsilon^{l_n}-\epsilon^{-l_n}|}{|\epsilon^{l_1^0}-\epsilon^{-l_1^0},\ldots,\epsilon^{l_n^0}-\epsilon^{-l_n^0}|}. \tag{1.4c}$$

Here the l_i^0 are integers and are given by $l_i^0 = n - i + 1$ and $l_i = f_i + l_i$ with f_i integer and $f_1 \geqslant f_2 \geqslant \cdots \geqslant f_n \geqslant 0$.

(D) SO($2n$):

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1}+\epsilon^{-l_1},\ldots,\epsilon^{l_n}+\epsilon^{-l_n}|}{|\epsilon^{l_1^0}+\epsilon^{-l_1^0},\ldots,\epsilon^{l_n^0}+\epsilon^{-l_n^0}|}. \tag{1.4d}$$

Here the l_i^0 are integers defined by $l_i^0 = n - i$ and $l_i = f_i + l_i^0$ with f_i either all integers or all half-integers and $f_1 \geqslant f_2 \geqslant \cdots \geqslant |f_n| > 0$. When $f_n = 0$, the right-hand side of (1.4d) is divided by a factor of 2.

The ordered set $(f_1, f_2, \ldots, f_n)$ is called the signature[1] and is also equal to the highest weight vector. To every irreducible representation corresponds one and only one signature $(f_1, f_2, \ldots, f_n)$ such that $f_1 \geqslant f_2 \geqslant \ldots \geqslant f_n$, and this signature is called dominant.

Using algebraic manipulations, one can rewrite expressions (1.4a–d) in terms of a set of characters of generating representations instead of in terms of the exponential functions ϵ_i. Weyl's *second* formula[2,3] gives the characters in terms of the so-called symmetric representations. For SU(n), Weyl's second formula reads

$$\chi_{(f_1,f_2,\ldots,f_n)} = \det\Sigma, \tag{1.5a}$$

$$\Sigma_{ij} = \chi_{d^{i-j+f_i}}. \tag{1.5b}$$

d^k has signature $(k, 0, 0, \ldots, 0)$ when $k \geqslant 0$, and is defined to vanish when $k < 0$.

Similar formulae exist for the other three series of classical groups.[2] Weyl's second formula is quite useful: It provides a practical algorithm for the decomposition of the ten-

[a)] This work is supported in part through funds provided by the U. S. Department of Energy (DOE) under contract DE-AC02-76ERO3069.

 0022-2488/84/010001-12$02.50

sor product of two representations into irreducible representations.[3] Indeed, from (1.5) it is clear that the tensor product of two arbitrary representations λ and μ can be expressed in terms of products of λ with *symmetric* representations. The latter products may be evaluated using a set of rules deduced from Weyl's first formula.[2,3]

In the present paper, we shall show that, for the classical groups the character of an arbitrary representation may also be expressed as a determinant only involving the so-called *basis* representations in an elementary way. For a group of rank n, there are precisely n basis representations whose signatures are listed below.[3] (Henceforth, we identify a representation with its dominant signature.)

(A) $SU(n+1)$: $d_p = (\underbrace{1,1,...,1}_{p},0,...,0)$ and $p = 1,...,n$; (1.6a)

(B) $SO(2n+1)$: d_p, $p = 1,...,n-1$ and the spinor representation $s = (\frac{1}{2},...,\frac{1}{2})$; (1.6b)

(C) $Sp(n)$: d_p, $p = 1,...,n$; (1.6c)

(D) $SO(2n)$: d_p, $p = 1,n-2$, and the two spinor representations $s^+ = (\frac{1}{2},...,\frac{1}{2},\frac{1}{2})$ (1.6d)
$s^- = (\frac{1}{2},...,\frac{1}{2},-\frac{1}{2})$.

Our formulae give all characters in terms of only a finite number of generators[4]: $\{\chi_{d_p}\}_{p=1,n}$ for $SU(n+1)$ and $Sp(n)$, $\{\chi_{d_p},\chi_s\}_{p=1,n-1}$ for $SO(2n+1)$ and $\{\chi_{d_p},\chi_{s^+},\chi_{s^-}\}_{p=1,n-2}$ for $SO(2n)$. The proof of these relations, henceforth called *decomposition formulae*, is the main objective of the present paper, and is given in Secs. II, III, IV, and V, respectively for $SU(n)$, $SO(2n+1)$, $Sp(n)$, and $SO(2n)$. For each of these groups, we shall first determine rules for the product of a basis representation with an arbitrary representation and then prove the decomposition formulae, essentially by explicit calculation of the determinant.

The case of $SU(n)$ is simplest, and will be developed in much detail; the case of $SO(2n+1)$ requires several important modifications, which we shall fully describe. For $Sp(n)$, only the final results will be given, and, for $SO(2n)$, special attention will be devoted to subtleties like double characters. Finally, in the last section we shall discuss three applications. First, we show that our decomposition formulae provide rules for the tensor multiplication of two arbitrary representations of any of the classical groups, just as Weyl's second formula did.[2] These rules are only slightly more complicated for the groups $Sp(n)$ or $SO(n)$ than for the group $SU(n)$, and may present an interesting alternative to the rather involved rules discussed in standard references.[5] Second, we prove a relation between the dimensions of the representations of $Sp(n)$ and these of the spinor representations of $SO(2n+1)$. Finally, we show that our decomposition formulae yield a simple algorithm for the calculation of the restriction of a representation to a classical subgroup of the original group. Thus branching rules for nonmaximal subgroups can be obtained. Let us also remark that the simple rules for products and branching of representations could be easily implemented in a computer program.

The extension of our formulae to the case of exceptional groups is presently under investigation.

II. THE SPECIAL UNITARY GROUPS SU(n)

Multiplication of a basis representation with an arbitrary representation

The weight diagram[3] for the basis representations is deduced from Weyl's first formula (1.4a)

$$\chi_{d_p} = \sum_{i_1<i_2<\cdots<i_p} \epsilon_{i_1}\cdots\epsilon_{i_p}. \tag{2.1}$$

The character of the tensor product of d_p with a representation $\lambda = (f_1, f_2, ..., f_n)$ is found using (1.3):

$$\chi_{d_p\otimes\lambda} = \sum_{i_1<i_2<\cdots<i_p} \epsilon_{i_1}\epsilon_{i_2}\cdots\epsilon_{i_p} \frac{|\epsilon^{l_1},...,\epsilon^{l_n}|}{|\epsilon^{l_1^0},...,\epsilon^{l_n^0}|}. \tag{2.2}$$

The integers l_i are defined in terms of the f_i by $l_i = f_i + l_i^0$.

Note that χ_λ and χ_{d_p} are invariant under the action of the Weyl group,[2,3] which permutes the angles ϕ_i. Using this invariance for $\chi_{d_p\otimes\lambda}$, we find

$$\chi_{d_p\otimes\lambda} = \sum_{i_1<i_2<\cdots<i_p} \frac{|\epsilon^{l_1},...,\epsilon^{l_{i_1}},...,\epsilon^{l_{i_p}},...,\epsilon^{l_n}|}{|\epsilon^{l_1^0},...,\epsilon^{l_n^0}|} \tag{2.3}$$

so that

$$\chi_{d_p\otimes\lambda} = \sum_{i_1<i_2<\cdots<i_p} \chi_{(f_1,...,f_{i_1}+1,...,f_{i_p}+1,...,f_n)}. \tag{2.4a}$$

In (2.4a), a character corresponding to a signature which is not dominant must be omitted. Expression (2.4a), together with the one-to-one correspondence between dominant characters and irreducible representations, implies the following formula for the representations:

$$d_p\otimes\lambda = \sum_{i_1<i_2<\cdots<i_p} (f_1,...,f_{i_1}+1,...,f_{i_p}+1,...,f_n). \tag{2.4b}$$

Here again, nondominant signatures are deleted.

The decomposition formula for the symmetric representations

Before attacking the full problem, we shall first prove a decomposition formula for the symmetric representation d^k of $SU(n)$ [defined in (1.5)].

Theorem 1. Let M^k be the following determinant[6]

$$M^k = \begin{vmatrix} d_1 & 1 & 0 & & 0 \\ d_2 & d_1 & 1 & & 0 \\ d_3 & d_2 & d_1 & & 0 \\ \vdots & \vdots & & \ddots & \\ d_k & d_{k-1} & & & d_1 \end{vmatrix}\otimes. \tag{2.5}$$

Then we have $M^k = d^k$.

In formula (2.5) it is understood that $d_k = 0$ if $k > n$ or $k < 0$.

Proof: Upon multiplication by the determinant

$$1 = \begin{vmatrix} 1 & 0 & 0 & & 0 \\ -d^1 & 1 & 0 & & 0 \\ d^2 & -d^1 & 1 & & 0 \\ & & & \ddots & \vdots \\ (-1)^{k-1}d^{k-1} & \cdots & & & 1 \end{vmatrix}\otimes, \tag{2.6}$$

making use of the well-known[3] duality relation

$$\sum_{p=0}^{n-1}(-1)^p d_p \otimes d^{k-p} = \begin{cases} 1 & \text{if } k=0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.7}$$

it is clear that $M^k = d^k$, as announced.

We now give also a different proof, the method of which will generalize to the case of an arbitrary representation of SU(n) as well as to the other classical groups. The expansion of the determinant in (2.5) along the first column yields a sum of products of a basis representation d_j with a minor Δ_j. The crucial remark is that this minor Δ_j is of the same form as the original determinant: $\Delta_j = M^{k-j}$. Thus we have

$$M^k = \sum_{\alpha=1}^{k} d_\alpha \otimes M^{k-\alpha}(-1)^{\alpha-1}. \tag{2.8}$$

We can prove (2.5) by induction. Suppose that $M^k = d^k$ for all $k \leqslant p-1$ and clearly $M^1 = d^1$; then we wish to prove that $M^p = d^p$. The induction hypothesis together with (2.8) yields

$$M^k = \sum_{\alpha=1}^{k} d_\alpha \otimes d^{k-\alpha}(-1)^{\alpha-1}. \tag{2.9}$$

Using (2.4b), we see that

$$d_\alpha \otimes d^{k-\alpha} = B_\alpha + B_{\alpha+1}, \tag{2.10}$$

where

$$B_\alpha = (\underbrace{k-\alpha+1,1,\cdots,1}_{\alpha},0,\ldots,0) \tag{2.11}$$

and $B_{k+1} = 0$ since it corresponds to a nondominant signature. Hence we have

$$M^k = \sum_{\alpha}^{k}(-1)^{\alpha-1}(B_\alpha + B_{\alpha+1}) \tag{2.12}$$

so that $M^k = B_1 = (k, 0, \ldots, 0) = d^k$ as announced. Upon replacing M^j by d^j in (2.8) we get precisely (2.7). This finishes the proof of Theorem 1.

Combination of (2.5) and (1.5) shows that every representation can be written as a function of basis representations d_p alone. We shall now prove a much more convenient formula for the decomposition in terms of basis representations.

The general decomposition formula

Let λ be a representation with dominant signature $(f_1, f_2, \ldots, f_n)$; the nonnegative projection numbers r_i are obtained from the projection of the highest weight vector onto the roots[3]:

$$r_i = f_i - f_{i+1}, \quad i = 1,\ldots,n-1. \tag{2.13}$$

We shall now prove the following general decomposition formula for the unitary groups[6,7]:

Theorem 2: Let

$$\mu = \begin{vmatrix} d_1 & 1 & \cdots & 0 & \cdots & & & & \\ d_2 & d_1 & & \vdots & & & & & \\ \vdots & & \ddots & & & & & & \\ d_{r_1} & \cdots & & d_1 & 1 & 0 & & & \\ d_{r_1+2} & \cdots & & d_3 & d_2 & d_1 & \cdots & & \\ & & & & & & d_2 & & \\ & & & & & & & \ddots & \\ & & & & & & & & d_{n-1} \\ & & & & & & & \ddots & d_{n-1}\ d_{n-2} \\ & & & & & & 0 & \cdots & 1\ \ d_{n-1} \end{vmatrix} \begin{matrix} \otimes \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} r_1 \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} r_2 \\ \vdots \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} r_{n-1} \end{matrix} \tag{2.14}$$

or

$$\mu = \otimes det(\mathscr{D}) \quad \text{with } \mathscr{D}_{ij} = d_{i-j+k} \tag{2.15}$$

and let k be defined by $r_1 + r_2 + \cdots + r_{k-1} < i \leqslant r_1 + r_2 + \cdots + r_k$. Then we have $\mu = \lambda$.

Observe that in formula (2.14) we have r_i times the representation d_i on the diagonal. The off-diagonal elements of the determinant are obtained by incrementing (resp. decrementing) the index i by one unit when moving to the left (resp. to the right).

Proof: In analogy with Theorem 1, the expansion of the determinant (2.14) along the first column yields a sum of products of a basis representation and a minor, which is of the same form as the original determinant μ. We proceed with a proof by induction on the first coordinate of the signature f_1. Suppose $\mu = \lambda$ for all $f_1 \leqslant p-1$; then we want to prove that $\mu = \lambda$ for *all* representations such that $f_1 = p$. Clearly, we have $\mu = \lambda$ for $f_1 = 1$. As a consequence of the induction hypothesis, we see that every minor corresponds to one and only one irreducible representation. Indeed, the minor associated with d_1 has signature $(f_1 - 1, f_2, \ldots, f_n)$, the minor associated with d_2 has signature $(f_1 - 2, f_2, \ldots, f_n)$; this pattern continues until one encounters the element d_{r_1} in the first column which has minor $(f_2, f_2, f_3, \ldots, f_n)$. If $r_2 \neq 0$, then at least one d_2 is present on the diagonal, and the next element in the first column is d_{r_1+2} with minor $(f_2 - 1, f_2 - 1, f_3, \ldots, f_n)$. Upon increasing the index of the element in the first column by 1, the second entry in the signature of the minor decreases by 1. It is remarkable that each minor in the expansion of determinant (2.14) is again an irreducible representation with a signature such that its first entry is always smaller than f_1. Thus we must prove that the expansion of the determinant, for a representation with dominant signature $(f_1, f_2, \ldots, f_n)$, with all minors replaced by their actual value precisely yields $\mu = \lambda$.

Using the signature notation, the above described expansion becomes

$$\mu = \oplus \sum_{\alpha=1}^{r_1} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}$$
$$\oplus \sum_{\alpha=1}^{r_2} d_{r_1+1+\alpha} \otimes (f_2-1, f_2-\alpha, f_3, \ldots, f_n)$$
$$\times(-1)^{\alpha+r_1-1} \oplus \cdots$$
$$\oplus \sum_{\alpha=1}^{r_{n-1}} d_{r_1+\cdots+r_{n-2}+\alpha+n-2}$$
$$\otimes (f_2-1, f_3-1, \ldots, f_{n-1}-1, f_n-\alpha)$$
$$\times(-1)^{r_1+\cdots+r_{n-2}+\alpha-2}. \tag{2.16}$$

All signatures appearing in (2.16) are dominant by construction.

Formula (2.16) may, however, be simplified through the use of nondominant signatures—henceforth called signatures as opposed to dominant signatures. We shall generalize (1.4a) to signatures $(f_1, f_2, \ldots, f_n)$, which need not be dominant, by defining their character as

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1}, \epsilon^{l_2}, \ldots, \epsilon^{l_n}|}{|\epsilon^{l_1^0}, \epsilon^{l_2^0}, \ldots, \epsilon^{l_n^0}|} \tag{2.17}$$

even when f is not dominant. Of course, every signature is either related to a dominant signature by permutation of columns in (2.16) or must vanish.[8] Thus we have, e.g., $\chi_{(2,4,1)} = -\chi_{(3,3,1)}$ but $\chi_{(2,3,1)} = 0$. Using the definition of (non dominant) signatures, (2.16) may be rewritten

$$\mu = \oplus \sum_{\alpha=1}^{r_1} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}$$
$$\oplus \sum_{\alpha=1}^{r_2} d_{r_1+1+\alpha} \otimes (f_2-\alpha-1, f_2, \ldots, f_n)(-1)^{r_2+\alpha}$$
$$\oplus \sum_{\alpha=1}^{r_3} d_{r_1+r_2+2+\alpha} \otimes (f_3-\alpha-2, f_2, f_3, \ldots, f_n)$$
$$\times(-1)^{r_1+r_2+\alpha+1}$$
$$\oplus \cdots. \tag{2.18}$$

A shift in the summation variable produces

$$\mu = \oplus \sum_{\alpha=1}^{r_1} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}$$
$$\oplus \sum_{\alpha=r_1+2}^{r_1+r_2+1} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}$$
$$\oplus \sum_{\alpha=r_1+r_2+3}^{r_1+r_2+r_3+2} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1} \oplus \cdots$$
$$\oplus \sum_{\alpha=r_1+r_2+\cdots+r_{n-2}+n-1}^{r_1+r_2+\cdots+r_{n-1}+n-2} d_\alpha \otimes (f_1-\alpha, f_2, f_3, \ldots, f_n)$$
$$\times(-1)^{\alpha-1}. \tag{2.19}$$

The signature vanishes at values of α which are missing from the summation, through the use of (2.17). Hence we have, e. g.,

$$(f_1-r_1-1, f_2, f_3, \ldots, f_n) = (f_2-1, f_2, f_3, \ldots, f_n) = 0. \tag{2.20}$$

Using the above property, we obtain our final formula:

$$\mu = \otimes \sum_{\alpha=1}^{f_1+n-2} d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}. \tag{2.21}$$

We shall now prove that $\mu = \lambda$, by explicit calculation of μ.

First we need the generalization (2.4) to nondominant signatures:

$$d_p \otimes (f_1, f_2, \ldots, f_n)$$
$$= \oplus \sum_{i_1<i_2<\cdots<i_p} (f_1, \ldots, f_{i_1}+1, \ldots, f_{i_p}+1, \ldots, f_n). \tag{2.22}$$

In (2.21), all terms may be kept since the nondominant signatures automatically vanish when $(f_1, f_2, \ldots, f_n)$ is dominant. By permutation of columns in (2.2) and (2.17), it is also clear that formula (2.22) holds even when f is not dominant.

The explicit calculation of all terms in the tensor products in (2.21) is unnecessary, and we shall introduce the following convenient shorthand

$$(f_1, \{f_2, \ldots, f_n\}_+^p)$$
$$= \oplus \sum_{1<i_1<i_2\cdots<i_p} (f_1, \ldots, f_{i_1}+1, \ldots, f_{i_p}+1, \ldots, f_n) \tag{2.23}$$

with

$$(f_1, \{f_2, \ldots, f_n\}_+^0) = (f_1, f_2, \ldots, f_n) \tag{2.24a}$$
$$(f_1, \{f_2, \ldots, f_n\}_+^p) = 0 \quad \text{when} \quad p<0 \text{ or } p>n-1. \tag{2.24b}$$

Then the tensor products in (2.21) can be computed using (2.23) and (2.24):

$$d_\alpha \otimes (f_1-\alpha, f_2, \ldots, f_n) = (f_1-\alpha+1, \{f_2, \ldots, f_n\}_+^{\alpha-1}) + (f_1-\alpha, \{f_2, \ldots, f_n\}_+^{\alpha}) \tag{2.25}$$

so that

$$\mu = \oplus \sum_{\alpha=1}^{f_1+n-2} (-1)^{\alpha-1}(f_1-\alpha+1, \{f_2, \ldots, f_n\}_+^{\alpha-1})$$
$$\oplus \sum_{\alpha=1}^{f_1+n-2} (-1)^{\alpha-1}(f_1-\alpha, \{f_2, \ldots, f_n\}_+^{\alpha}). \tag{2.26}$$

A shift in the summation variable of the second sum yields

$$\mu = (f_1, f_2, f_3, \ldots, f_n)$$
$$\oplus (-1)^{f_1+n-1}(-n+2, \{f_2, \ldots, f_n\}_+^{f_1+n-2}). \tag{2.27}$$

Since $f_1 \geqslant 1$, the second bracket vanishes with the use of (2.24b) and since $(f_1, f_2, \ldots, f_n)$ is dominant, we have proven that $\mu = \lambda$.

Example: The representation λ with signature (4,2,1,0,0) of SU(5) is decomposed as follows:

$$(4,2,1,0,0) = \begin{vmatrix} d_1 & 1 & 0 & 0 \\ d_2 & d_1 & 1 & 0 \\ d_4 & d_3 & d_2 & d_1 \\ 0 & 1 & d_4 & d_3 \end{vmatrix}.$$

One can, e. g., check the dimensions by taking the character of both sides of (2.27) and computing the determinant at the identity element. With the use of tables of dimensions,[9] we find

$$700 = \begin{vmatrix} 5 & 1 & 0 & 0 \\ 10 & 5 & 1 & 0 \\ 5 & 10 & 10 & 5 \\ 0 & 1 & 5 & 10 \end{vmatrix}$$

III. THE ORTHOGONAL GROUPS SO(2n + 1)

The decomposition formulae for the spinor representation are different and will be treated separately from the formulae for single valued representations. Furthermore, it will appear natural to express the decomposition formulae in terms of the basis representations of (1.6b) *plus* the representation $d_n = (1,1,...,1)$. Later we shall prove that the latter representation is simply expressed in terms of the former ones.

Multiplication of a basis representation with an arbitrary representation

The weight diagram of the representations d_p is deduced from Weyl's first formula (1.4b):

$$\chi_{d_p} = \frac{|\epsilon^{l_1} - \epsilon^{-l_1},...,\epsilon^{l_n} - \epsilon^{-l_n}|}{|\epsilon^{l_1^0} - \epsilon^{-l_1^0},...,\epsilon^{l_n^0} - \epsilon^{-l_n^0}|}, \tag{3.1}$$

where $l_i^0 = n - i + \frac{1}{2}$, $l_i = l_i^0$ if $i > p$ and $l_i = l_i^0 + 1$ if $i \leqslant p$. We shall obtain a more convenient expression for this weight diagram by introducing the function

$$\mathscr{A}_p = |\epsilon^n + \epsilon^{-n},...,\epsilon^{n-p+1} + \epsilon^{-n+p-1},\epsilon^{n-p-1} + \epsilon^{-n+p+1},...,\epsilon + \epsilon^{-1},1|. \tag{3.2}$$

After division of numerator and denominator in (3.1) by the common factor $\Pi_{i=1}^n(\epsilon_i^{1/2} - \epsilon_i^{-1/2})$, χ_{d_p} can be rewritten as follows.

$$\chi_{d_p} = (\mathscr{A}_p + \mathscr{A}_{p-1})/\mathscr{A}_0. \tag{3.3}$$

The function $\mathscr{A}_p$ can be easily evaluated using the binomial coefficients C_n^k:

$$\mathscr{A}_p = \sum_{\alpha=0}^{[p/2]} C_{n-p+2\alpha}^{\alpha} R(p-2\alpha)\mathscr{A}_0, \tag{3.4}$$

with

$$R(q) = \sum_{i_1<i_2<\cdots<i_q} \sum_{\sigma_{i_j}=\pm 1} \epsilon_{i_1}^{\sigma_{i_1}}\epsilon_{i_2}^{\sigma_{i_2}}\cdots\epsilon_{i_q}^{\sigma_{i_q}}. \tag{3.5}$$

The function $R(q)$ may be thought of as the character of the representation d_q of a unitary group with ϵ_i replaced with $\epsilon_i + \epsilon_i^{-1}$. Combination of formulae (3.3)–(3.5) gives us the weight diagram of d_p:

$$\chi_{d_p} = \sum_{\alpha=0}^{[p/2]} C_{n-p+2\alpha}^{\alpha} R(p-2\alpha) + \sum_{\alpha=0}^{[(p-1)/2]} C_{n-p+1+2\alpha}^{\alpha} R(p-1-2\alpha). \tag{3.6}$$

For the spinor representation, the weight diagram is computed directly from (1.4b)

$$\chi_s = \sum_{\sigma_i=\pm 1} \epsilon_1^{\sigma_1/2}\epsilon_2^{\sigma_2/2}\cdots\epsilon_n^{\sigma_n/2}. \tag{3.7}$$

The functions $R(q)$, χ_{d_p} and χ_s are invariant under the action of the Weyl group which permutes the ϕ_i and changes their sign. Again we generalize the dominant signatures in (1.4b) to (nondominant) or generalized signatures, defined through the same formula (1.4b) but where the signature $(f_1, ..., f_n)$ need not be dominant. Every generalized signature is again either related to a dominant signature or must vanish. An important special case is

$$\chi_{(f_1,f_2,...,f_{n-1},-1)} = -\chi_{(f_1,f_2,...,f_{n-1},0)} \tag{3.8}$$

for $(f_1, f_2, ..., f_{n-1}, 0)$ dominant.

With the help of the weight diagram computed previously, we can evaluate the tensor product of d_p and s with a representation $\lambda = (f_1, f_2, ..., f_n)$. First we need the product of $R(q)$ with χ_λ, obtained using the invariance under the action of the Weyl group:

$$R(q)\chi_\lambda = \sum_{i_1<i_2<\cdots<i_q} \sum_{\sigma_{i_j}=\pm 1} \chi_{(f_1,...,f_{i_1}+\sigma_{i_1},...,f_{i_q}+\sigma_{i_q},...,f_n)}. \tag{3.9}$$

Formula (3.9) also holds for $(f_1, f_2, ..., f_n)$ *not* dominant. The product of d_p and λ is gotten by combining (3.6) and (3.9). The tensor product of the spinor representation with λ is deduced from (3.7) in an analogous fashion.

$$\chi_s\chi_\lambda = \sum_{\sigma_i=\pm 1} \chi_{(f_1+\sigma_1/2, f_2+\sigma_2/2,..., f_n+\sigma_n/2)}. \tag{3.10}$$

As an example of these multiplication rules, we compute the following product for SO(9):

$$\begin{aligned}(1,1,0,0)\otimes(3,1,0,0) &= (4,\{1,0,0\}^1) + (3,\{1,0,0\}^2) + (2,\{1,0,0\}^1) + \cdots \\ &= (4,2,0,0) + (4,1,1,0) + (4,0,0,0) + (3,2,1,0) \\ &\quad + (3,1,1,1) + (2,2,0,0) + 2(3,1,0,0) \\ &\quad + (2,1,1,0) + (2,0,0,0).\end{aligned}$$

The decomposition formula for nonspinor representations

It is not hard to generalize formula (2.14) to the case of the group SO(2n + 1). The explicit form of the weight diagram in (3.6) suggests that we should take the linear combinations

$$D_k = d_k \ominus d_{k-1} \oplus d_{k-2} \oplus \cdots \oplus (-1)^k d_0 \tag{3.11}$$

as elementary building blocks in a determinantal expression like (2.14). Examination of a few simple special cases shows that this is basically correct, provided one modifies (2.14) in a way which we shall now specify. Let λ be a representation with dominant signature $(f_1, ..., f_n)$ and define the integers

$$r_i = f_i - f_{i+1}, \quad i = 1,2,...,n-1, \quad r_n = f_n, \tag{3.12}$$

as well as the sequence of direct sums and differences of basis representations

$$\begin{aligned} &0 \leqslant k \leqslant n, && D_k = d_k \ominus d_{k-1} \oplus d_{k-2} + \cdots \oplus (-1)^k d_0, \\ &n \leqslant k \leqslant 2n, && D_k = D_{2n-k}, \\ &k > 2n \text{ or } k < 0, && D_k = 0. \end{aligned} \tag{3.13}$$

Theorem 3: Let

$$\mu = \begin{vmatrix} D_1 & D_0 & & \cdots & & & 0 & \Big\} \otimes \\ D_2 & D_1 & & & & & 0 & \Big\} r_1 \\ \vdots & & \ddots & & & & \vdots & \\ D_{r_1} & \cdots & D_1 & & & & 0 & \\ D_{r_1+2} & \cdots & D_3 & D_2 & & & & \Big\} r_2 \\ & & & & \ddots & & & \\ & & & & D_2 & & & \\ & & & & & \ddots & & \\ & & & & & D_n \oplus & D_{n-r_n} & \\ & & & & & & \ddots & \\ & & & \cdots & & D_n \oplus D_{n-3} & D_{n-1} \oplus D_{n-2} & \\ & & & \cdots & & D_{n+1} \oplus D_{n-2} & D_n \oplus D_{n-1} & \Big\} r_n \end{vmatrix} \tag{3.14}$$

or in components

$$\mu = \otimes \det(\mathscr{D}),$$
$$\mathscr{D}_{ij} = D_{i-j+k} \oplus D_{i+j+k-1-2f_i},$$
$$r_1 + r_2 + \cdots + r_{k-1} < i \leqslant r_1 + r_2 + \cdots + r_k. \tag{3.15}$$

Then $\mu = \lambda$.

Proof: The proof again proceeds by induction of f_1. Suppose $\mu = \lambda$ for *all* λ such that $f_1 \leqslant p-1$; then we wish to prove that $\mu = \lambda$ for *all* λ such that $f_1 = p$. The expansion of the determinant along the first column yields tensor products of representations $D_{i-1+k} \oplus D_{i+k-2f_i}$ with minors. These minors are of the same form as the original determinant, and by the recurrence hypothesis equal to irreducible representation of which the first entry in the signature never exceeds $f_1 - 1$. The resulting formula for μ is the same as (2.16) but with d_α replaced with $D_\alpha \oplus D_{\alpha+1-2f_i}$. The definition of (generalized) signatures for the SO($2n+1$) again leads to a drastic simplification, analogous to the one that leads to (2.21) and we finally get

$$\mu = \sum_{\alpha=1}^{f_1+n-1} (D_\alpha \oplus D_{\alpha+1-2f_i}) \otimes (f_1 - \alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}. \tag{3.16}$$

We prove that $\mu = \lambda$ by explicit calculation of μ. As for the unitaries, the tensor products in (3.16) need not be worked out explicitly and we introduce the following shorthand:

$$(f_1, \{f_2, \ldots, f_n\}_\pm^p) = \oplus \sum_{1 < i_1 < i_2 \cdots < i_p} \sum_{\sigma_{i_j} = \pm 1} (f_1, \ldots, f_{i_1} + \sigma_{i_1}, \ldots, f_{i_p} + \sigma_{i_p}, \ldots, f_n), \tag{3.17a}$$

$$(f_1, \{f_2, \ldots, f_n\}_\pm^0) = (f_1, f_2, \ldots, f_n), \tag{3.17b}$$

$$(f_1, \{f_2, \ldots, f_n\}_\pm^p) = 0 \quad \text{if} \quad p < 0 \text{ or } p > n-1. \tag{3.17c}$$

We first evaluate the product of the functions $R(p)$ with an arbitrary character using (3.9):

$$R(p)\chi_{(f_1, f_2, \ldots, f_n)} = \chi_{(f_1+1, \{f_2, \ldots, f_n\}_\pm^{p-1})} + \chi_{(f_1, \{f_2, \ldots, f_n\}_\pm^p)} + \chi_{(f_1-1, \{f_2, \ldots, f_n\}_\pm^{p-1})}. \tag{3.18}$$

The weight diagram of D_p is computed using (3.5) and is given by

$$\chi_{D_p} = \sum_{\beta=0}^{[p/2]} C_{n-p+2\beta}^{\beta} R(p-2\beta) \tag{3.19}$$

for all $p \geqslant 0$. Next, we compute the tensor products occurring in (3.16) and make use of the shorthand introduced in (3.17)–(3.18)

$$D_\alpha \otimes (f_1 - \alpha, f_2, \ldots, f_n) = \sum_{\beta=0}^{[\alpha/2]} C_{n-\alpha+2\beta}^{\beta} (B_{\alpha-1}^{\beta} \oplus B_\alpha^\beta \oplus B_{\alpha+1}^{\beta+1}), \tag{3.20a}$$

$$D_{\alpha+1-2f_i} \otimes (f_1 - \alpha, f_2, \ldots, f_n) = \sum_{\beta=0}^{[(\alpha+1-2f_1)/2]} C_{n-\alpha-1+2f_1+2\beta}^{\beta} \times (\mathscr{K}_{\alpha-1}^{\beta} \oplus \mathscr{K}_\alpha^\beta \oplus \mathscr{K}_{\alpha+1}^{\beta+1}) \tag{3.20b}$$

with

$$B_\alpha^\beta = (f_1 - \alpha, \{f_2, \ldots, f_n\}_\pm^{\alpha-2\beta}), \tag{3.21a}$$
$$\mathscr{K}_\alpha^\beta = (f_1 - \alpha, \{f_2, \ldots, f_n\}_\pm^{\alpha-2\beta-2f_1+1}). \tag{3.21b}$$

We compute μ in two steps:

$$\mu_1 = \sum_{\alpha=1}^{f_1+n-1} D_\alpha \otimes (f_1 - \alpha, f_2, \ldots, f_n)(-1)^{\alpha-1} = \sum_{\alpha=1}^{f_1+n-1} \sum_{\beta=0}^{[\alpha/2]} C_{n-\alpha+2\beta}^{\beta} \times (B_{\alpha-1}^{\beta} \oplus B_\alpha^\beta \oplus B_{\alpha+1}^{\beta+1})(-1)^{\alpha-1}. \tag{3.22}$$

Upon performing the appropriate shifts in the summation variables, we find

$$\mu_1 = \sum_{\alpha=0}^{f_1+n-2} \sum_{\beta=0}^{[(\alpha+1)/2]} C_{n-\alpha-1+2\beta}^{\beta} B_\alpha^\beta (-1)^\alpha \ominus \sum_{\alpha=1}^{f_1+n-1} \sum_{\beta=0}^{[\alpha/2]} C_{n-\alpha+2\beta}^{\beta} B_\alpha^\beta (-1)^\alpha \oplus \sum_{\alpha=2}^{f_1+n} \sum_{\beta=1}^{[(\alpha+1)/2]} C_{n-\alpha-1+2\beta}^{\beta-1} B_\alpha^\beta (-1)^\alpha. \tag{3.23}$$

For α odd we have

$$B_\alpha^{[(\alpha+1)/2]} = (f_1 - \alpha, \{f_2, \ldots, f_n\}_\pm^{-1}) = 0 \tag{3.24}$$

so that the summation over β in the second term may be

extended from $[\alpha/2]$ to $[(\alpha+1)/2]$. Then we rearrange expression (3.23) as follows:

$$\begin{aligned}\mu_1 = {} & B_0^0 \ominus B_1^0 \ominus n B_1^1 \oplus B_1^0 \oplus (n+1) B_1^1 \\ & \oplus \sum_{\beta=0}^{[(f_1+n)/2]} C^{\beta}_{1-f_1+2\beta} B^{\beta}_{f_1+n-1} (-1)^{f_1+n} \\ & \oplus \sum_{\beta=1}^{[(f_1+n)/2]} C^{\beta-1}_{-f_1+2\beta} B^{\beta}_{f_1+n-1} (-1)^{f_1+n-1} \\ & \oplus \sum_{\beta=1}^{[(f_1+n+1)/2]} C^{\beta-1}_{-f_1-1+2\beta} B^{\beta}_{f_1+n} (-1)^{f_1+n} \\ & \oplus \sum_{\alpha=2}^{f_1+n-2} \sum_{\beta=0}^{[(\alpha+1)/2]} (C^{\beta}_{n-\alpha-1+2\beta} - C^{\beta}_{n-\alpha+2\beta} \\ & + C^{\beta-1}_{n-\alpha+2\beta}) B^{\beta}_{\alpha} (-1)^{\alpha}. \end{aligned} \tag{3.25}$$

The double sum in (3.25) vanishes due to Pascal's equality on binomial coefficients. Using (3.24) again for the term B_1^1 and Pascal's equality,

$$\begin{aligned}\mu_1 = {} & B_0^0 \oplus \sum_{\beta=0}^{[(f_1+n)/2]} C^{\beta}_{-f_1+2\beta} B^{\beta}_{f_1+n-1} (-1)^{f_1+n} \\ & \oplus \sum_{\beta=1}^{[(f_1+n+1)/2]} C^{\beta-1}_{-f_1-1+2\beta} B^{\beta}_{f_1+n} (-1)^{f_1+n}. \end{aligned} \tag{3.26}$$

The same sequence of manipulations may be applied to the expression for μ_2,

$$\mu_2 = \sum_{\alpha=1}^{f_1+n-1} D_{\alpha+1-2f_1} \otimes (f_1-\alpha, f_2, \ldots, f_n)(-1)^{\alpha-1}, \tag{3.27a}$$

and it yields

$$\begin{aligned}\mu_2 = {} & \sum_{\beta=f_1}^{[(n+f_1+1)/2]} (-1)^{f_1+n} C^{\beta-f_1}_{-f_1+2\beta-1} \mathscr{K}^{\beta-f_1}_{f_1+n-1} \\ & \oplus \sum_{\beta=f_1}^{[(n+f_1)/2]} (-1)^{f_1+n} C^{\beta-f_1}_{-f_1+2\beta} \mathscr{K}^{\beta-f_1+1}_{f_1+n} \end{aligned} \tag{3.27b}$$

Using the properties of the binomial coefficients, we see that the sums in (3.26) actually only start at $\beta = f_1$ instead of at $\beta = 0$ or $\beta = 1$. Taking this remark into account, we obtain the following result for μ:

$$\begin{aligned}\mu = {} & B_0^0 \oplus \sum_{\beta=f_1}^{[(f_1+n)/2]} C^{\beta}_{-f_1+2\beta} (B^{\beta}_{f_1+n-1} \\ & \oplus \mathscr{K}^{\beta-f_1+1}_{f_1+n})(-1)^{f_1+n} \\ & \oplus \sum_{\beta=f_1}^{[(f_1+n+1)/2]} C^{\beta-1}_{-f_1+2\beta-1} (B^{\beta}_{f_1+n} \\ & \oplus \mathscr{K}^{\beta-f_1}_{f_1+n-1})(-1)^{f_1+n}. \end{aligned} \tag{3.28}$$

From the definition of B and $\mathscr{K}$ in (3.20) and making use of the properties of generalized signatures we see that

$$\begin{aligned}\mathscr{K}^{\beta-f_1+1}_{f_1+n} & = (-n, \{f_2, \ldots, f_n\}_{\pm}^{f_1+n-2\beta-1}) \\ & = (-1)^{n-1}(\{f_2-1, f_3-1, \ldots, f_n-1\}_{\pm}^{f_1+n-2\beta-1}, -1). \end{aligned}$$

With the help of (3.8) this reduces to

$$\begin{aligned}\mathscr{K}^{\beta-f_1+1}_{f_1+n} & = (-1)^n(\{f_2-1, f_3-1, \ldots, f_n-1\}_{\pm}^{f_1+n-2\beta-1}, 0) \\ & = -(1-n, \{f_2, \ldots, f_n\}_{\pm}^{f_1+n-2\beta+1}). \end{aligned} \tag{3.29a}$$

Comparison with the definition of B yields

$$\mathscr{K}^{\beta-f_1+1}_{f_1+n} = -B^{\beta}_{f_1+n-1},$$

and similarly we have

$$\mathscr{K}^{\beta-f_1}_{f_1+n-1} = -B^{\beta}_{f_1+n}. \tag{3.29b}$$

As a consequence, the two sums in (3.28) cancel exactly, and we get

$$\mu = B_0^0 = (f_1, f_2, \ldots, f_n), \tag{3.30}$$

as announced in Theorem 3.

The decomposition formula for spinor representations

Examination of some simple examples again suggests that the correct building blocks for the decomposition formula are the D_k introduced in (3.13). The correct modification of (3.14) is then easily found, and will now be given. Let λ be a spinor representation of SO($2n+1$), with signature $(f_1, f_2, \ldots, f_n)$, and define the integers

$$r_i = f_i - f_{i+1}, \quad i = 1, \ldots, n-1, \quad r_n = f_n - \tfrac{1}{2}. \tag{3.31}$$

Theorem 4: Let

$$\mu = s \otimes \begin{vmatrix} D_1 & D_0 & 0 & & & & & & \\ D_2 & D_1 & D_0 & & & & & & \\ & D_1 & D_0 & 0 & & & & & \\ & D_3 & D_2 & D_1 & & & & & \\ & & & \ddots & & & & & \\ & & & & D_2 & & & & \\ & & & & & \ddots & & & \\ & & & & & & D_n \ominus D_{n-r_1} & & \\ & & & & & & & \ddots & \\ & & & & & & & & D_n \ominus D_{n-4} \\ & & & & & & & & D_{n+1} \ominus D_{n-3} \end{vmatrix} \begin{matrix} \otimes \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} r_1 \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} r_2 \\ \vdots \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} r_n \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ D_{n-1} \ominus D_{n-3} \\ D_n \ominus D_{n-2} \end{matrix} \tag{3.32}$$

or in components

$$\mu = s \otimes \det \mathscr{D}, \tag{3.33}$$

$$\mathscr{D}_{ij} = D_{i-j+k} \ominus D_{i-j+k-2-2f_i}$$

and k is defined by $r_1 + r_2 + \cdots + r_{k-1} < i \leqslant r_1 + r_2 + \cdots + r_k$. Then $\mu = \lambda$. Please note the difference in *sign* between (3.15) and (3.33) as well as the difference in index in the second term.

Outline of the proof: The proof proceeds by induction on f_1 as before. The definitions of generalized signatures and braces { } of (3.17) are extended to spinor representations and μ is again computed directly using the expansion of determinant (3.32) along the first column. Proceeding along the lines of the proof of Theorem 3, we find that we must calculate

$$\mu = \sum_{\alpha=1}^{f_1+n-1} (D_\alpha \ominus D_{-2f_1+\alpha}) \otimes (f_1-\alpha, f_2, ..., f_n)(-1)^{\alpha-1}. \tag{3.34}$$

The tensor products are evaluated with the help of the weight diagram of d_p in (3.5) and collected with the brace notation (3.17). After simplifications analogous to those made in the proof of Theorem 3, we find

$$\mu = (f_1, f_2, ..., f_n), \tag{3.35}$$

as announced.

In both Theorems 3 and 4, we have decomposed all representations of SO($2n+1$) in terms of $d_1, d_2, ..., d_n$ and s, even though d_n is not on the list of basis representations in (1.6b). We have done so because $d_1, d_2, ..., d_n$ and s form the natural set in terms of which the decomposition formulae are simplest. In addition, this presents no loss of generality since d_n itself is expressed in terms of the set of basis representations (1.6b) by a simple formula, which we shall now derive. From (3.10), we deduce

$$s \otimes s = \oplus \sum_{\sigma_i=0,1} (\sigma_1, \sigma_2, ..., \sigma_n). \tag{3.36}$$

Cancelling nondominant signatures leaves us with

$$s \otimes s = \oplus \sum_{p=0}^{n} d_p \tag{3.37}$$

so that

$$d_n = s \otimes s \ominus \sum_{p=0}^{n-1} d_p. \tag{3.38}$$

IV. THE SYMPLECTIC GROUP Sp(n)

The representation theory for the symplectic group is much simpler than that for SO($2n+1$), since there are no spinor representations. Moreover, the decomposition formuale as well as their proof are very similar to the case of SO($2n+1$). For this reason, we just quote the results for the decomposition formula; the reader should have no problem reconstructing the proof.

Multiplication of a basis representation with an arbitrary representation

The weight diagram of the representation d_p with signature $(\underbrace{1,1,...,1}_{p},0,...,0)$ (for $p = 1, ..., n$) is deduced from (1.4c) and can be conveniently expressed as

$$\chi_{d_p} = (\mathscr{A}_p - \mathscr{A}_{p-2})/\mathscr{A}_0. \tag{4.1}$$

The function $\mathscr{A}_p$ has been defined in (3.2), and the resulting weight diagram of d_p is found with the help of (3.4)

$$\chi_{d_p} = \sum_{\alpha=0}^{[p/2]} C^{\alpha}_{n-p+2\alpha} R(p-2\alpha) - \sum_{\alpha=0}^{[(p-2)/2]} C^{\alpha}_{n-p+1+2\alpha} R(p-2-2\alpha). \tag{4.2}$$

Here R is the function defined in (3.4b), and the product of R with the character of a representation with signature $(f_1, f_2, ..., f_n)$ is given by

$$R(q)\chi_\lambda = \sum_{i_1<i_2<\cdots<i_q} \sum_{\sigma_{i_j}=\pm 1} \chi_{(f_1,...,f_{i_1}+\sigma_{i_1},...,f_{i_q}+\sigma_{i_q},...,f_n)}. \tag{4.3}$$

The tensor product of d_p with the representation λ is then simply obtained combining (4.2) and (4.3).

The decomposition formula

Let λ be a representation with signature $(f_1, f_2, ..., f_n)$. Define the integers r_i by

$$r_i = f_i - f_{i+1}, \quad i = 1,...,n-1, \quad r_n = f_n \tag{4.4}$$

as well as the sequence of reducible representations

$$\begin{aligned} &0 \leqslant k \leqslant n, && \hat{D}_k = d_k \oplus d_{k-2} \oplus d_{k-4} \oplus \cdots \oplus \begin{cases} d_1 & \text{if } k \text{ odd}, \\ d_0 & \text{if } k \text{ even}, \end{cases} \\ &n \leqslant k \leqslant 2n, && \hat{D}_k = \hat{D}_{2n-k}, \\ &k > 2n \text{ or} \\ &k < 0, && \hat{D}_k = 0. \end{aligned} \tag{4.5}$$

Theorem 5: Define

$$\mu = \begin{vmatrix} \hat{D}_1\hat{D}_0 0 & & & & & & \\ \hat{D}_2\hat{D}_1\hat{D}_0 & & & & & & \\ \vdots \quad \ddots & & & & & & \\ \hat{D}_{r_1} \cdots \hat{D}_1 & & & & & & \\ & \hat{D}_3\hat{D}_2\hat{D}_1 \cdots & & & & & \\ & & \ddots & & & & \\ & & & \hat{D}_2 & & & \\ & & & & \ddots & & \\ & & & & \hat{D}_n \ominus \hat{D}_{n-4} & \hat{D}_{n-1} \ominus \hat{D}_{n-3} & \\ & & & & \hat{D}_{n+1} \ominus \hat{D}_{n-1} & \hat{D}_n \ominus \hat{D}_{n-2} & \end{vmatrix}^{\otimes} \tag{4.6}$$

(braces: r_1, r_2, r_n)

or in components

$$\begin{aligned} &\mu = \otimes \det \mathscr{D} \\ &\mathscr{D}_{ij} = \hat{D}_{i-j+k} \ominus \hat{D}_{i-j+k-2-2f_1} \end{aligned} \tag{4.7}$$

and k is defined by $r_1 + r_2 + \cdots + r_{k-1} < i \leqslant r_1 + r_2 + \cdots + r_l$. Then $\mu = \lambda$.

V. THE ORTHOGONAL GROUP SO($2n$)

According to whether $f_n = 0$ or $\neq 0$, the characters of the group SO(2n) defined in (1.4d) correspond to irreducible or reducible representations. When $f_n = 0$, the representations is non-self-associate, and the character is simple. When $f_n \neq 0$, the representation is self-associate, reducible into two associate irreducible representations of the same dimension, and the character is said to be a double character. We shall

show how to construct the tensor product of an arbitrary representation λ with a basis representation (whether self-associate or not) and if the basis representation is reducible, we shall also show how to find the product of its irreducible associate components with λ. A decomposition formula will be proven for both self-associate and non-self-associate representations. We have not found a decomposition formula for the irreducible components of a self-associate representation.

Multiplication of a generating representation with an arbitrary representation

We shall need the following generating representations

$$\begin{aligned} &d_p = (\underbrace{1,1,\ldots,1}_{p},0,\ldots,0), \qquad p = 1,\ldots,n-1, \\ &d_n^{\pm} = (1,1,\ldots,\pm 1), \quad s^{\pm} = (\tfrac{1}{2},\tfrac{1}{2},\ldots,\pm\tfrac{1}{2}), \\ &d_n = d_n^{+} \oplus d_n^{-}, \quad s = s^{+} \oplus s^{-}. \end{aligned} \tag{5.1}$$

The representations $\pm$ are associate to each other, whereas s and d_n are self-associate. When n is odd, $\pm$ are actually complex conjugates, whereas, for n even, both $+$ and $-$ are real. We first determine the weight diagrams of d_p and s and then indicate how those of $s^{\pm}$ and $d_n^{\pm}$ may be gotten.

From Weyl's *first* formula (1.4d) and using the definition of the function $\mathscr{A}_p$ in (3.2), we see that

$$\chi_{d_p} = \mathscr{A}_p / \mathscr{A}_0. \tag{5.2}$$

With the help of (3.4b), χ_{d_p} may be expressed in terms of R:

$$\chi_{d_p} = \sum_{\alpha=0}^{[p/2]} C^{\alpha}_{n-p+2\alpha} R(p-2\alpha). \tag{5.3}$$

The double character of the spinor representation is given by

$$\chi_s = \sum_{\sigma_i = \pm 1} \epsilon_1^{\sigma_1/2} \epsilon_2^{\sigma_2/2} \cdots \epsilon_n^{\sigma_n/2}. \tag{5.4}$$

The quantity σ defined as

$$\sigma = \prod_{i=1}^{n} \sigma_i \tag{5.5}$$

may take the values ± 1 in (5.4). The characters of s^{+} (resp. s^{-}) are also defined by (5.4), but now σ must be restricted to be 1 (resp. -1). The expression for the character of $d_n^{\pm}$ is more complicated, and we shall not give it here. It may be deduced from the relation

$$d_n^{\pm} = s^{\pm} \otimes s^{\pm} - d_{n-2} - d_{n-4} - \cdots. \tag{5.6}$$

For representations with $f_n = 0$ and self-associate representations, a generalized signature may be defined:

$$\chi_{(f_1,f_2,\ldots,f_n)} = \frac{|\epsilon^{l_1} + \epsilon^{-l_1}, \ldots, \epsilon^{l_n} + \epsilon^{-l_n}|}{|\epsilon^{l_1^0} + \epsilon^{-l_1^0}, \ldots, 1|} \tag{5.7}$$

even if $(f_1,f_2,\ldots,f_n)$ is not dominant. For the two irreducible associate representations into which a self-associate representation decomposes, similar generalized signatures may be defined, but we shall not need these here.

Tensor multiplication is effected using formulae (5.3)–(5.4); the product of R with the character of an arbitrary representation λ with signature $(f_1,f_2,\ldots f_n)$ is given by

$$R(q)\chi_\lambda = \sum_{i_1 < i_2 \cdots < i_q} \sum_{\sigma_{i_j} = \pm 1} G(\sigma) \chi_{(f_1,f_2,\ldots,f_{i_1}+\sigma_{i_1},\ldots,f_{i_q}+\sigma_{i_q},\ldots,f_n)} \tag{5.8}$$

and the product of χ_s and χ_λ is

$$\chi_s \chi_\lambda = \sum_{\sigma_i = \pm 1} F(\sigma) \chi_{(f_1+\sigma_1/2,\ldots,f_n+\sigma_n/2)}. \tag{5.9}$$

The integers $G(\sigma)$ and $F(\sigma)$ are present to obtain the correct counting of self-associate and non-self-associate representations. They are determined from (3.4b) and (5.7) using the invariance under the action of the Weyl group: first we make $(f_1,f_2,\ldots,f_n)$ *dominant*. For the integer G we have

$$\begin{aligned} &G(\sigma) = 1 \quad \text{if } i_q \neq n \text{ or } \sigma_n \neq -1, \\ &G(\sigma) = 1 \quad \text{if } \sigma_n = -1 \text{ and } f_n > 1, \\ &G(\sigma) = 2 \quad \text{if } \sigma_n = -1 \text{ and } f_n = 1, \\ &G(\sigma) = 0 \quad \text{if } \sigma_n = -1 \text{ and } f_n = 0. \end{aligned} \tag{5.10}$$

For the integer $F(\sigma)$ we have

$$\begin{aligned} &F(\sigma) = 1 \quad \text{if } \sigma_n = 1, \\ &F(\sigma) = 1 \quad \text{if } \sigma_n = -1 \text{ and } f_n \neq \tfrac{1}{2}, \\ &F(\sigma) = 2 \quad \text{if } \sigma_n = -1 \text{ and } f_n = \tfrac{1}{2}. \end{aligned} \tag{5.11}$$

Products with the representations s^{+} or s^{-} are obtained by making the appropriate restrictions on σ given in (5.9).

As an example of these multiplication rules, one may compute the following product for SO(10).[10] [We use the definition $R(q) = \operatorname{tr}\rho(q)$.]

$$\begin{aligned} &(1,1,1,1,0) \otimes (3,2,2,2,1) = [\rho(4) + 3\rho(2) + 10\rho(0)] \otimes (3,2,2,2,1), \\ &\rho(4) \otimes (3,2,2,2,1) = (4,\{2,2,2,1\}^3_{\pm}) \oplus (3,\{2,2,2,1\}^4_{\pm}) \oplus (2,\{2,2,2,1\}^3_{\pm}) \\ &\quad = (4,3,3,3,1) + (4,3,3,2,2) + 2(4,3,3,2,0) + 2(4,3,2,1,0) \\ &\qquad + (4,3,3,1,1) - (4,3,2,2,1) + (4,3,1,1,1) - 2(4,2,2,2,2) - 4(4,2,2,2,0) + 2(4,2,1,1,0) \\ &\qquad - (4,2,2,1,1) + (4,1,1,1,1) + (3,3,3,3,2) + 2(3,3,3,3,0) + 2(3,3,3,1,0) - (3,3,2,2,2) \\ &\qquad - 2(3,3,2,2,0) + 2(3,3,1,1,0) - 2(3,2,2,1,0) + 2(3,1,1,1,0) - 2(2,2,2,2,2) - 4(2,2,2,2,0) \\ &\qquad - (2,2,2,1,1) + (2,1,1,1,1) + 2(2,2,1,1,0), \\ &\rho(2) \otimes (3,2,2,2,1) = (4,\{2,2,2,1\}^1) + (3,\{2,2,2,1\}^2) + (2,\{2,2,2,1\}^1) \\ &\quad = (4,3,2,2,1) + (4,2,2,1,1) + (4,2,2,2,2) + 2(4,2,2,2,0) \\ &\qquad + (3,3,3,2,1) + (3,3,2,1,1) + (3,3,2,2,2) + 2(3,3,2,2,0) + 2(3,2,2,1,0) \\ &\qquad - (3,2,2,2,1) + (2,2,2,2,2) + (2,2,2,1,1) + 2(2,2,2,2,0) - (3,2,2,2,1) + (3,2,1,1,1). \end{aligned}$$

Putting all together, we obtain

$$(1,1,1,1,0)\otimes(3,2,2,2,1) = (4,3,3,3,1) + (4,3,3,2,2) + 2(4,3,3,2,0) + 2(4,3,2,1,0) + (4,3,3,1,1) + 2(4,3,2,2,1) + (4,3,1,1,1) + (4,2,2,2,2) + 2(4,2,2,2,0) + 2(4,2,1,1,0) + 2(4,2,2,1,1) + (4,1,1,1,1) + (3,3,3,3,2) + 2(3,3,3,3,0) + 2(3,3,3,1,0) + 2(3,3,2,2,2) + 4(3,3,2,2,0) + 2(3,3,1,1,0) + 4(3,2,2,1,0) + 2(3,1,1,1,0) + (2,2,2,2,2) + 2(2,2,2,2,0) + 2(2,2,2,1,1) + (2,1,1,1,1) + 3(3,3,2,1,1) + 4(3,2,2,2,1) + 3(3,3,3,2,1) + 2(2,2,1,1,0) + 3(3,2,1,1,1). \tag{5.12}$$

With the help of the tables of dimensions of representations,[9] we may check that dimensions work out correctly:

$$\begin{aligned}210\times 50\,688 = {} & 945\,945 + 660\,660 + 1\,698\,840 + 1\,048\,576 \\ & + 882\,882 + 2\times 848\,925 + 242\,550 + 90\,090 + 274\,560 + 143\,000 \\ & + 2\times 199\,017 + 17\,325 + 84\,942 + 165\,165 + 210\,210 + 128\,700 \\ & + 2\times 189\,189 + 73\,710 + 2\times 72\,765 + 8085 + 2772 + 8910 \\ & + 2\times 6930 + 1050 + 3\times 128\,700 + 4\times 50\,688 + 3\times 219\,648 + 5940 \\ & + 3\times 23\,040.\end{aligned}$$

The decomposition formula for nonspinor representations simple and double characters

Let λ be a representation with (dominant) signature $(f_1,f_2,\ldots,f_n)$[10] and define the sequence of representations

$$\begin{aligned} &0\leqslant k\leqslant n, && \delta_k = d_k \\ &n\leqslant k\leqslant 2n, && \delta_k = d_{2n-k}, \\ &k\geqslant 2n \text{ or } k<0, && \delta_k = 0.\end{aligned} \tag{5.13}$$

Note that all nonzero representations in this sequence are irreducible and that d_n is self-associate. We shall now prove the following decomposition theorem in the case of nonspinor representations.

Theorem 6: Let

$$\begin{aligned}&\mu = \otimes \det \mathscr{D}, \\ &\mathscr{D}_{ij} = (\delta_{i-j+k}\oplus\delta_{i+j+k-2f_i})/(1+\delta_{j,f_i})\end{aligned} \tag{5.14}$$

and let k be defined by $f_1 - f_k < i \leqslant f_1 - f_{k+1}$. Then $\mu = \lambda$.

Proof: As for the other three classical groups, the proof proceeds by induction on f_1. Expansion of determinant (5.14) along the first column and the use of generalized signatures reduce the calculation of μ to the evaluation of the following expression:

$$\begin{aligned}\mu = {} & \sum_{\alpha=1}^{f_1+n-1} (\delta_\alpha \oplus \delta_{\alpha-2f_1+2}) \\ & \otimes (f_1-\alpha, f_2,\ldots,f_n)(-1)^{\alpha-1}.\end{aligned} \tag{5.15}$$

To work out the products in (5.15), we use (5.8) and rearrange different contributions to the product of $R(q)$ and χ_λ with the help of the brace notation introduced in (3.17).

$$\begin{aligned}&\rho(q)\otimes(f_1-\alpha,f_2,\ldots,f_n) \\ &\quad = (f_1-\alpha+1,\{f_2,\ldots,f_n\}_{\pm}^{q-1}) \\ &\qquad \oplus (f_1-\alpha,\{f_2,\ldots,f_n\}_{\pm}^{q}) \\ &\qquad \oplus G(f_1-\alpha-1,\{f_2,\ldots,f_n\}_{\pm}^{q-1}),\end{aligned} \tag{5.16a}$$

where G is determined by the rules of (5.10):

$$\begin{aligned}&G=0 \quad \text{if } \alpha = f_1+n-1, \\ &G=2 \quad \text{if } \alpha = f_1+n-2, \\ &G=1 \quad \text{if otherwise.}\end{aligned} \tag{5.16b}$$

Then we make use of (5.3) and obtain

$$\begin{aligned}&\delta_\alpha\otimes(f_1-\alpha,f_2,\ldots,f_n) \\ &\quad = \sum_{\beta=0}^{[\alpha/2]} C^{\beta}_{n-\alpha+2\beta}(B^{\beta}_{\alpha-1}\oplus B^{\beta}_{\alpha}\oplus GB^{\beta+1}_{\alpha+1}),\end{aligned} \tag{5.17a}$$

where G is defined in (5.16b):

$$\begin{aligned}&\delta_{\alpha-2f_1+2}\otimes(f_1-\alpha,f_2,\ldots,f_n) \\ &\quad = \sum_{\beta=0}^{[\alpha/2]-f_1+1} C^{\beta}_{n-\alpha+2f_1-2+2\beta} \\ &\qquad \times(B^{\beta+f_1-1}_{\alpha-1}\oplus B^{\beta+f_1-1}_{\alpha}\oplus GB^{\beta+f_1}_{\alpha+1}).\end{aligned} \tag{5.17b}$$

Here we have made use of the quantity

$$B^{\beta}_{\alpha} = (f_1-\alpha,\{f_2,f_3,\ldots,f_n\}_{\pm}^{\alpha-2\beta}). \tag{5.18}$$

Shifts in summation variables, the use of Pascal's equality, and the explicit definition of G lead to

$$\begin{aligned}&\mu = \mu_1\oplus\mu_2, \\ &\mu_1 = B^0_0 \oplus \sum_{\beta=0}^{[(f_1+n)/2]} C^{\beta}_{-f_1+2\beta}B^{\beta}_{f_1+n-1}(-1)^{f_1+n} \\ &\qquad \ominus \sum_{\beta=0}^{[(f_1+n)/2]} C^{\beta-1}_{-f_1+2\beta}B^{\beta}_{f_1+n-1}(-1)^{f_1+n},\end{aligned} \tag{5.19a}$$

$$\begin{aligned}&\mu_2 = \sum_{\beta=0}^{[(-f_1+n)/2]+1} C^{\beta}_{f_1-2+2\beta}B^{\beta+f_1-1}_{f_1+n-1}(-1)^{f_1+n} \\ &\qquad \ominus \sum_{\beta=0}^{[(-f_1+n)/2]+1} C^{\beta-1}_{f_1-2+2\beta}B^{\beta+f_1-1}_{f_1+n-1}(-1)^{f_1+n}.\end{aligned} \tag{5.19b}$$

Making the substitution $\beta\to\beta-f_1+1$ in (5.19a) and using the properties of the binomial coefficients, it can be shown that the four β summations in (5.19) precisely cancel, leaving only $\mu = B^0_0 = (f_1,\ldots,f_n) = \lambda$ as announced.

The decomposition formula for spinor representations

Finally we shall exhibit a decomposition formula in the case of the (always self-associate) spinor representations. The proof is completely analogous to the proofs of Theorems 4 and 6 and will not be given here. We define the same sequence of representations δ_k in (5.13), let λ be a representation with signature $(f_1,\ldots,f_n)$.

Theorem 7: Let

$$\mu = \otimes \det \mathscr{D} \otimes s,$$

$$\mathscr{D}_{ij} = \delta_{i-j+k} \ominus \delta_{i+j+k-2f_1-1}, \tag{5.20}$$

and let k be defined by $f_1 - f_k < i \leqslant f_1 - f_{k+1}$. Then $\mu = \lambda$. Please note the difference in sign and the difference in indices between (5.20) and (5.14).

VI. APPLICATIONS

A. Multiplication of arbitrary representations

Several algorithms exist in the literature for the decomposition into irreducible representations of the tensor product of two irreducible representations.[2,3,5] If the weight diagram of one of the representations is known, Weyl's first formula can be used to obtain the irreducible components.[2,3] However, the determination of the weight diagram is a notoriously difficult problem. Želobenko[3] and Murnaghan[5] use Weyl's second formula, respectively for the unitary and orthogonal groups. The knowledge of the product of a symmetric representation with an arbitrary representation then suffices to perform the product of two arbitrary representations. This method is very attractive for the unitary groups,[3] but appears quite involved for the orthogonal groups.[5]

With the Theorems 2–7, we dispose of decomposition formulae in terms of the *basis* representations. In proving these relations, we have also shown how to perform the tensor product of any of these basis representations with an arbitrary representation. Thus, we dispose of an algorithm that allows us to compute the tensor product of two arbitrary representations, and the rules of this algorithm seem rather convenient, even though the calculations remain lengthy.

To demonstrate the practicality of these rules, we shall work out an example of intermediate difficulty: the tensor product in Sp(4) of the representations α and β with signatures (2,1,1,0) and (3,2,2,1). To do so, we use Theorem 5 for α:

$$(2,1,1,0) \otimes (3,2,2,1) = \begin{vmatrix} \hat{D}_1 & \hat{D}_0 \\ \hat{D}_4 - \hat{D}_0 & \hat{D}_3 - \hat{D}_1 \end{vmatrix}^{\otimes} \otimes (3,2,2,1)$$
$$= \hat{D}_1 \otimes (\hat{D}_3 - \hat{D}_1) \otimes (3,2,2,1) - (\hat{D}_4 - \hat{D}_0) \otimes (3,2,2,1),$$

$$(\hat{D}_3 - \hat{D}_1) \otimes (3,2,2,1) = (4,3,3,1) + (4,3,1,1) + (4,3,2,2) + (4,3,2,0)$$
$$+ (4,2,1,0) + (4,2,2,1) + (4,1,1,1) + (3,3,3,2) + (3,3,3,0) + (3,3,1,0)$$
$$+ (3,2,2,2) + (3,2,2,0) + (3,1,1,0) + (2,2,1,0) + (2,2,2,1) + (2,1,1,1)$$
$$+ 2(3,3,2,1) + 2(3,2,1,1),$$

$$(\hat{D}_4 - \hat{D}_0) \otimes (3,2,2,1) = (4,3,3,2) + (4,3,3,0) + (4,3,1,0) + (4,2,2,2)$$
$$+ (4,2,2,0) + (4,1,1,0) + (2,2,2,2) + (2,2,2,0) + (2,1,1,0) + 2(4,3,2,1)$$
$$+ 2(4,2,1,1) + 2(3,3,3,1) + 2(3,3,1,1) + 2(3,3,2,2) + 2(3,3,2,0) + 2(3,2,1,0)$$
$$+ 2(3,1,1,1) + 3(3,2,2,1) + 2(2,2,1,1).$$

Putting all together, we obtain

$$(2,1,1,0) \otimes (3,2,2,1) = (5,3,3,1) + (4,4,3,1) + (5,3,1,1) + (4,4,1,1)$$
$$+ 5(4,3,2,1) + (5,3,2,2) + (4,4,2,2) + 2(4,3,3,2) + (5,3,2,0)$$
$$+ (4,4,2,0) + 2(4,3,3,0) + 3(4,3,1,0) + (5,2,1,0) + 3(4,2,2,0)$$
$$+ (4,2,0,0) + (5,2,2,1) + 4(4,2,1,1) + 2(4,2,2,2) + (5,1,1,1)$$
$$+ 2(4,1,1,0) + (3,3,3,3) + 3(3,3,3,1) + 4(3,3,2,0) + (3,3,0,0)$$
$$+ (2,2,2,2) + 3(3,3,2,2) + 2(2,2,2,0) + 5(3,2,1,0) + 5(3,2,2,1)$$
$$+ 2(2,1,1,0) + (3,1,0,0) + 3(3,1,1,1) + (2,2,0,0) + (1,1,1,1)$$
$$+ 4(3,3,1,1) + 3(2,2,1,1).$$

It is also useful to check the dimensions using the tables[9]:

$$315 \times 6237 = 213\,444 + 122\,850 + 96\,228 + 41\,250 + 5 \times 65\,536$$
$$+ 142\,155 + 67\,760 + 2 \times 56\,628 + 146\,250 + 66\,528 + 2 \times 42\,042$$
$$+ 3 \times 29\,106 + 36\,864 + 3 \times 16\,848 + 4914 + 63\,063 + 4 \times 14\,300$$
$$+ 2 \times 13\,728 + 9009 + 2 \times 3696 + 4719 + 3 \times 12\,012 + 4 \times 10\,010$$
$$+ 2184 + 594 + 3 \times 9009 + 2 \times 825 + 5 \times 4096 + 5 \times 6237$$
$$+ 2 \times 315 + 594 + 3 \times 1155 + 308 + 42 + 4 \times 7020 + 3 \times 792.$$

B. A relation between the dimensions of the representations of Sp(n) and spinor representations of SO($2n+1$)

Formulae (3.32) and (4.6) have the same formal structure except for the overall tensor product with the spinors in (3.32) and the difference in definition for D_k and $\hat{D}_k$. In particular, the value of the characters of D_k and $\hat{D}_k$ at the identity of the group can be shown to be equal. Indeed, upon using formulae (3.4)–(3.6) and (3.13) on one hand and formulae (4.2) and (4.5) on the other, we find that

$$\chi_{D_k}(e) = \chi_{\hat{D}_k}(e) = \sum_{\alpha=0}^{[k/2]} C^{\alpha}_{n-k+2\alpha} 2^{k-2\alpha} C^{k-2\alpha}_{n} \tag{6.1}$$

Substitution of (6.1) into (3.32) and (4.6) yields the following result. Let λ be an arbitrary representation of Sp(n) with dominant signature $(f_1, f_2, \ldots, f_n)$, let Λ be the representation of SO($2n+1$) with signature $(f_1 + \frac{1}{2}, f_2 + \frac{1}{2}, \ldots, f_n + \frac{1}{2})$, and let s be the fundamental spinor of SO($2n+1$) with signature $(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$. Then we have

$$\dim \Lambda = \dim s \dim \lambda. \tag{6.2}$$

In fact, it is also clear from (3.5) and (3.13) that, in the canonical basis, a more general relation holds

$$\chi_{\Lambda}(h) = \chi_s(h)\chi_{\lambda}(h), \tag{6.3}$$

where h is an element of the Cartan subgroup, parametrized by the angles $\phi_1, \ldots, \phi_n$.

C. Restriction of a representation to a subgroup

Let G be any of the four classical groups, and let G_0 be any of its classical nontrivial subgroups. We wish to determine the irreducible components of the restriction of the representation λ of G to G_0. If, by classical methods, we can derive the restriction of the basis representations of G to the subgroup G_0, then we can calculate the restriction of any representation by Theorems 2–7.

We shall treat the following simple example:

$$G = \mathrm{SU}(2n), \quad G_0 = \mathrm{SO}(2n).$$

The restriction of the basis representations of SU($2n$) to SO($2n$) are *irreducible* and given by[11]

$$d^{A}_{k}|_{\mathrm{SO}(2n)} = \delta^{D}_{k}, \quad k = 1, \ldots, 2n, \tag{6.4}$$

where δ_k is defined by (5.13). The restriction of a representation of SU($2n$) is then given by determinant (2.14) in which d^{A}_{k} is replaced by δ^{D}_{k}. It is usually not necessary to fully work out the products in this new determinant, as often irreducible representations of SO($2n$) may be recognized in it. Consider, e. g., the restriction of the representation (2,2,2,1,1,0,0) of SU(8) to SO(8),

$$(2,2,2,1,1,0,0,0)_A\,|_{\mathrm{SO}(8)} = \begin{vmatrix} d^{A}_{3} & d^{A}_{2} \\ d^{A}_{6} & d^{A}_{5} \end{vmatrix}_{\mathrm{SO}(8)} = \begin{vmatrix} \delta^{D}_{3} & \delta^{D}_{2} \\ \delta^{D}_{2} & \delta^{D}_{3} \end{vmatrix},$$

$$\delta^{D}_{3} \otimes \delta^{D}_{3} = (2,2,2,0) + (2,2,1,1) + (2,2,0,0) + 2(2,1,1,0) + (2,0,0,0) + (1,1,1,1) + 2(1,1,0,0) + (0,0,0,0),$$

$$\delta^{D}_{2} \otimes \delta^{D}_{2} = (2,2,0,0) + (2,1,1,0) + (2,0,0,0) + (1,1,1,1) + (0,0,0,0) + (1,1,0,0),$$

$$(2,2,2,1,1,0,0,0)_A\,|_{\mathrm{SO}(8)} = (2,2,2,0)_D + (2,2,1,1)_D + (2,1,1,0)_D + (1,1,0,0)_D. \tag{6.5}$$

Using the tables,[9] we can easily check that the dimensions work out:

$$2352_A = 840_D + 567_D + 567_D + 350_D + 28_D. \tag{6.6}$$

In a completely analogous fashion, the restrictions of the basis representation of SU($2n+1$) to SO($2n+1$) are also irreducible, and can be used to calculate the restrictions of arbitrary representations to SO($2n+1$).

The restrictions of the basis representations of SU($2n$)to Sp(n) are reducible and can be easily derived using conventional methods:

$$d^{A}_{k}|_{\mathrm{Sp}(n)} = \hat{D}^{C}_{k}, \tag{6.7}$$

where $\hat{D}_k$ has been defined in (4.5). We shall illustrate this restriction with an extremely simple example, the decomposition of the representation α of SU(6) with signature (2,2,1,1,0) to Sp(3):

$$\alpha|_{\mathrm{Sp}(3)} = \begin{vmatrix} d^{A}_{2} & d^{A}_{1} \\ d^{A}_{5} & d^{A}_{4} \end{vmatrix}^{\otimes}_{\mathrm{Sp}(3)} = \begin{vmatrix} \hat{D}^{C}_{2} & \hat{D}^{C}_{1} \\ \hat{D}^{C}_{1} & \hat{D}^{C}_{2} \end{vmatrix}^{\otimes}.$$

Working out these products, one finds

$$(2,2,1,1,0,0)_A\,|_{\mathrm{Sp}(3)} = (2,2,0)_C \oplus (2,1,1)_C \oplus 2(1,1,0)_C \otimes (0,0,0)_C \tag{6.8}$$

with dimensions

$$189_A = 90_C + 70_C + 2\times 14_C + 1_C.$$

The peculiar property of this algorithm is that we only need to know the restrictions of a *finite* number of representations to compute that of all representations. The procedure can be easily generalized to arbitrary classical groups G and G_0.

Note added in manuscript: The problem of decomposing a given representation into a finite set of basis representations has also been discussed by A. J. Feingold, Proc. Am. Math. Soc. **70**, 109 (1978). I thank Professor J. Patera for drawing my attention to this work.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor Bob Sharp, Dr. Yvan Saint-Aubin, and Dr. Baha Balantekin for stimulating discussions and for helpful remarks on the manuscript.

[1] For a review of the theory of compact Lie groups, see Refs. 2 and 3.

[2] H. Weyl, *The Classical Groups* (Princeton U. P., Princeton, NJ, 1973), and references therein.

[3] D. P. Želobenko, *Compact Lie Groups and Their Representations* (American Mathematical Society, Providence, RI, 1973).

[4] Variants of Weyl's second formula have been studied in N. E. Samra and R. C. King, J. Phys. A. Gen. **12**, 2305 (1979); R. C. King, J. Math. Phys. **12**, 1588 (1971); M. J. Newell, Proc. Roy. Irish Acad., 153 (1951).

[5] F. D. Murnaghan, *The Theory of Group Representations* (Baltimore, 1938).

[6] It is understood that, in the expansion of the determinant, all products are tensor products and all sums are direct sums. Whenever a minus sign occurs, it is understood that the representation is multiplied by -1, and will cancel an identical term with a $+$ sign.

[7] A formula essentially the same as (2.14) appears in the following references for the representations of the symmetric group and for that of the unitary group, respectively: D. E. Littlewood, *The Theory of Group Characters and Matrix Representations of Groups* (Oxford U. P., Oxford, 1940); M. J. Newell, Proc. Roy. Irish Acad., 345 (1949). (I am grateful to A. B. Balantekin for bringing the latter reference to my attention.)

[8] By definition, a character vanishes identically on the group if and only if its signature vanishes.

[9] W. G. McKay and J. Patera, *Table of Dimensions, Indices and Branching Rules for Representations of Simple Lie Algebras* (Marcel Dekker, New York, 1981).

[10] It is understood that in the case $f_n \neq 0$, the representation is self-associate and that its character is double.

[11] In this section, we shall indicate with a superscript to which group the representation d_k belongs. The superscripts are A, B, C, and D for respectively SU(n), SO($2n+1$), Sp(n), and SO($2n$).

On classes of integrable systems and the Painlevé property[a)]

John Weiss
La Jolla Institute, 8950 Villa La Jolla Drive, Suite 2150, La Jolla, California 92037 and Institute for Pure and Applied Physical Science, University of California, San Diego, La Jolla, California 92093

(Received 14 March 1983; accepted for publication 9 September 1983)

The Caudrey–Dodd–Gibbon equation is found to possess the Painlevé property. Investigation of the Bäcklund transformations for this equation obtains the Kuperschmidt equation. A certain transformation between the Kuperschmidt and Caudrey–Dodd–Gibbon equation is obtained. This transformation is employed to define a class of p.d.e.'s that identically possesses the Painlevé property. For equations within this class Bäcklund transformations and rational solutions are investigated. In particular, the sequences of higher order KdV, Caudrey–Dobb–Gibbon, and Kuperschmidt equations are shown to possess the Painlevé property.

PACS numbers: 02.30. + g

1. INTRODUCTION

In Ref. 1 the Painlevé property for partial differential equations was defined. Briefly, we say that a partial differential equation has the Painlevé property when the solutions of the p.d.e. are "single-valued" about the movable, singularity manifold and the singularity manifold is "noncharacteristic." To be precise, if the singularity manifold is determined by

$$\varphi(z_1,z_2,\ldots,z_n)=0 \tag{1.1}$$

and $u=u(z_1,\ldots,z_n)$ is a solution of the p.d.e., then we require that

$$u=\varphi^{\alpha}\sum_{j=0}^{\infty}u_j\varphi^{j}, \tag{1.2}$$

where $u_0\neq 0$, $\varphi=\varphi(z_1,\ldots,z_n)$, $u_j=u_j(z_1,\ldots,z_n)$ are analytic functions of (z_j) in a neighborhood of the manifold (1.1), and α is an integer. The requirement that the manifold (1.1) be noncharacteristic insures that the expansion (1.2) will be well defined, in the sense of the Cauchy–Kowalevsky theorem. Substitution of (1.2) into the p.d.e. determines the value(s) of α, and defines the recursion relations for u_j, $j=0,1,2,\ldots$. When the anzatz (1.2) is correct, the p.d.e. is said to possess the Painlevé property and is conjectured to be integrable. The "Painlevé conjecture," as originally formulated by Ablowitz *et al.*,[2] states that when all the ordinary differential equations obtained by exact similarity transforms from a given partial differential equation have the Painlevé property, then the partial differential equation is "integrable." The above definition of the "Painlevé property" allows this conjecture to be stated directly for the partial differential equation.

In Ref. 3 Bäcklund transformations were obtained by truncating the expansion (1.2) at the "constant" level term. That is, we set

$$u=u_0\varphi^{-N}+u_1\varphi^{-N+1}+\cdots+u_N \tag{1.3}$$

and find, from the recursion relations for u_j, an overdetermined system of equations for $(\varphi,u_j,j=0,1,\ldots,N)$, where u_N will satisfy the (original) p.d.e. Upon solving the overdetermined system, it was found, for those equations considered, that φ satisfied an equation formulated in terms of the Schwarzian derivative:

$$\{\varphi;x\}=\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right)-\frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2. \tag{1.4}$$

The invariance of (1.4) under the Moebius group

$$\varphi=\frac{a\psi+b}{c\psi+d},\quad \{\varphi;x\}=\{\psi;x\} \tag{1.5}$$

motivates the substitution

$$\varphi=v_1/v_2, \tag{1.6}$$

by which the Lax pairs may be found.[3]

Investigation of a certain class of equations formulated in terms of the Schwarzian derivatives revealed that these equations have the Painlevé property about movable, singularity manifolds of order -1. However, the occurrence of an additional type of movable singularity prevents this class of equations from identically possessing the Painlevé property. Hence, nonintegrable behavior can arise.[2]

In this paper a restriction (symmetry) is imposed that allows one to conclude that, when an equation is formulated in terms of the Schwarzian derivative and has this "symmetry," the equation identically possesses the Painlevé property. Within this class of equations are found the KdV, Caudrey–Dodd–Gibbon and Kuperschmidt equations. Furthermore, the "symmetry" property and invariance under the Moebius group allow effective Bäcklund transforms to be defined for these equations. In particular, rational or algebraic [in (x,t)] solutions can be generated iteratively.

In the next section, the Painlevé property and Bäcklund transformation for the KdV equation are reviewed for later reference.

In Sec. 3 the Painlevé property and Bäcklund transforms for the Caudrey–Dodd–Gibbon equation are presented. From these considerations the Kuperschmidt equation is found. The transformation between the Caudrey–Dodd–Gibbon and Kuperschmidt equations can be regarded as a

[a)] This work supported by Department of Energy Contract DOE DE-AC03-81ER10923 and AFOSR Grant No. AFOSR 83-0095.

 0022-2488/84/010013-12$02.50

certain "symmetry" under which these equations are "dual."

In Sec. 4, the "symmetry" discovered in Sec. 3 is employed to define a class of p.d.e.'s that possess the Painlevé property. The KdV equation is shown to be contained in this class of equations and self-dual w.r.t. this symmetry. Then, the sequences of higher order KdV, Caudrey–Dodd–Gibbon, and Kuperschmidt equations are found to be within this identically Painlevé class of equations and Bäcklund transformations are obtained for these sequences of equations.

In Sec. 5 rational [in (x,t)] solutions are constructed for several equations. In Appendix A the Lax pair for the Caudrey–Dodd–Gibbon equation is derived. In Appendix B further considerations relating to the seventh-order equations are presented.

2. THE KORTEWEG–DE VRIES EQUATION

The KdV equation

$$u_t + uu_x + u_{xxx} = 0 \tag{2.1}$$

possesses the Painlevé property.[1] The expansion about the singularity manifold has the form

$$u = \varphi^{-2} \sum_{j=0}^{\infty} u_j \varphi^j. \tag{2.2}$$

The "resonances" occur at

$$j = -1,4,6, \tag{2.3}$$

and (φ,u_4,u_6) are arbitrary functions of (x,t) in the expansion (2.2). We now assume the following "Bäcklund" transformation:

$$u = u_0/\varphi^2 + u_1/\varphi + u_2 \tag{2.4}$$

and find the following overdetermined system of equations,

$$\begin{aligned}
&\text{(i)} && u_0 = -12\varphi_x^2,\\
&\text{(ii)} && u_1 = 12\varphi_{xx},\\
&\text{(iii)} && \varphi_x\varphi_t + \varphi_x^2 u_2 + 4\varphi_x\varphi_{xxx} - 3\varphi_{xx}^2 = 0,\\
&\text{(iv)} && \varphi_{xt} + \varphi_{xx}u_2 + \varphi_{xxxx} = 0,\\
&\text{(v)} && u_{2t} + u_2u_{2x} + u_{2xxx} = 0,
\end{aligned} \tag{2.5}$$

$$u = 12\frac{\partial^2}{\partial x^2}\ln\varphi + u_2 \tag{2.6}$$

and, by eliminating u_2 in (2.5 iii,iv),

$$\varphi_t/\varphi_x + \{\varphi;x\} = \lambda, \tag{2.7}$$

where

$$\{\varphi;x\} = \frac{\partial}{\partial_x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\frac{\varphi_{xx}^2}{\varphi_x^2} \tag{2.8}$$

is the Schwarzian derivative of φ. Equation (2.7) is invariant under the Moebius group:

$$\varphi = \frac{a\psi + b}{c\psi + d}, \qquad \{\varphi;x\} = \{\psi;x\}. \tag{2.9}$$

The substitution[2]

$$\varphi = v_1/v_2, \quad \text{where } (v_1,v_2)\text{satisfy} \tag{2.10}$$

$$v_{xx} = av, \quad v_t = bv_x + cv, \tag{2.11}$$

readily obtains the Lax pair:

$$\begin{aligned}
a &= -\tfrac{1}{6}(u_2 + \lambda),\\
b &= -u_2/3 + \tfrac{2}{3}\lambda,\\
c &= u_x/6.
\end{aligned} \tag{2.12}$$

As noted in Ref. 2, Eq. (2.7) has an expansion

$$\varphi = \psi^{-1}\sum_{j=0}^{\infty}\varphi_j\psi^j \tag{2.13}$$

about a singularity manifold

$$\psi(x,t) = 0. \tag{2.14}$$

The resonances occur at

$$j = -1,0,1 \tag{2.15}$$

and the compatability conditions at $j = 0$ and 1 are satisfied identically. Thus, Eq. (2.7) has the Painlevé property about singularities of the form (2.13). However, we note that the vanishing of φ_x in (2.7) introduces the possibility of new, movable, singularities. This point will be resolved in Sec. 4.

The most general form of the Bäcklund transform defined by the expression

$$\varphi = \varphi_0/\psi + \varphi_1 \tag{2.16}$$

can be shown to be equivalent to the Moebius transformation (2.9). Again, an "effective" Bäcklund transformation for equation (2.7) will be defined in Sec. 4.

3. THE CAUDREY–DODD–GIBBON EQUATION

The Caudrey–Dodd–Gibbon equation[4,5]

$$u_t + \frac{\partial}{\partial x}(u_{xxxx} + 30uu_{xx} + 60u^3) = 0 \tag{3.1}$$

possesses the Painlevé property. The expansion about the singularity manifold is of the form

$$u = \varphi^{-2}\sum_{j=0}^{\infty}u_j\varphi^j. \tag{3.2}$$

There are found to be two solution branches.

Branch i: $u_0 = -\varphi_x^2$: The resonances occur at

$$j = -1,2,3,6,10. \tag{3.3}$$

Branch ii: $u_0 = -2\varphi_x^2$: The resonances occur at

$$j = -2,-1,5,6,12. \tag{3.4}$$

Both branches of the solution possess the Painlevé property.

The Bäcklund transformation defined for the "branch i" form of the solution is

$$u = u_0/\varphi^2 + u_1/\varphi + u_2. \tag{3.5}$$

The resulting overdetermined system of equations for (φ,u_0,u_1,u_2) is found to be

$$\begin{aligned}
&\text{(i)} && u_0 = -\varphi_x^2,\\
&\text{(ii)} && u_1 = \varphi_{xx},
\end{aligned} \tag{3.6}$$

$$\text{(iii)}\quad \frac{\varphi_t}{\varphi_x} + 6\frac{\varphi_{xxxxx}}{\varphi_z} - 15\frac{\varphi_{xx}\varphi_{xxxx}}{\varphi_x^2} + 10\frac{\varphi_{xxx}^2}{\varphi_x^2} + 30\left\{u_{2xx} + 4\left(\frac{\varphi_{xxx}}{\varphi_x} - 3\frac{\varphi_{xx}^2}{\varphi_x^2}\right)u_2 + 6u_2^2\right\} = 0, \tag{3.7}$$

(iv) $$\frac{\varphi_{xt}}{\varphi_x}+\frac{\varphi_{xxxxxx}}{\varphi_x}+30\left\{\frac{\varphi_{xx}}{\varphi_x}u_{2xx}+\frac{\varphi_{xxxx}}{\varphi_x}u_2+60\frac{\varphi_{xx}}{\varphi_x}u_2^2\right\}=0, \tag{3.8}$$

(v) $$u_{2t}+\frac{\partial}{\partial x}(u_{2xxxx}+30u_2u_{2xx}+60u_2^3)=0.$$

Using (3.6), Eq. (3.5) is

$$u=\frac{\partial^2}{\partial x^2}\ln\varphi+u_2. \tag{3.9}$$

We note that if

$$\varphi=1/\psi, \tag{3.10}$$

then

$$u_2=\frac{\partial^2}{\partial x^2}\ln\psi+u \tag{3.11}$$

and

$$W=u_2+\frac{1}{4}\frac{\varphi_{xx}^2}{\varphi_x^2}=u+\frac{1}{4}\frac{\psi_{xx}^2}{\psi_x^2}. \tag{3.12}$$

To employ this invariance, we let

$$u_2=W-\frac{1}{4}\frac{\varphi_{xx}^2}{\varphi_x^2} \tag{3.13}$$

and find

(i) $$\frac{\varphi_t}{\varphi_x}+6\frac{\partial^2}{\partial x^2}\{\varphi;x\}+19\{\varphi;x\}^2+30[W_{xx}+6W^2+4\{\varphi;x\}W]=0, \tag{3.14}$$

(ii) $$\frac{\partial}{\partial x}\left(\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+\frac{13}{2}\{\varphi;x\}^2\right)+30W\frac{\partial}{\partial x}\{\varphi;x\}=0, \tag{3.15}$$

where $\{\varphi;x\}$ is the Schwarzian derivative. To simplify these expressions, we let

$$\vartheta=\{\varphi;x\}+6W \tag{3.16}$$

and find

(i) $$\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2+5(\vartheta_{xx}+\vartheta^2+2\{\varphi;x\}\vartheta)=0, \tag{3.17}$$

(ii) $$\frac{\partial}{\partial x}\left(\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2\right)+5\vartheta\frac{\partial}{\partial x}\{\varphi;x\}=0. \tag{3.18}$$

From the consistency of (3.17) and (3.18)

$$\vartheta\vartheta_{xx}-\frac{\vartheta_x^2}{2}+\frac{2}{3}\vartheta^3+\{\varphi;x\}\vartheta^2=C. \tag{3.19}$$

Herein, we shall consider only the trivial solution

$$\vartheta=C=0, \tag{3.20}$$

$$\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2=0, \tag{3.21}$$

$$u_2=-\frac{1}{6}\frac{\varphi_{xxx}}{\varphi_x}. \tag{3.22}$$

It can be shown that Eqs. (3.21) and (3.22) imply that u_2 satisfies the Caudrey–Dodd–Gibbon equation. Actually, as is explained in Appendix A, (3.21) and (3.22) constitute a Lax pair for the Caudrey–Dodd–Gibbon equation.

We now let

$$\varphi=v_1/v_2, \quad \text{where } (v_1,v_2) \text{ satisfy} \tag{3.23}$$

$$v_{xx}=-\tfrac{1}{2}av, \quad v_t=bv_x+cv. \tag{3.24}$$

Equations (3.21), (3.23), and (3.24) imply that

$$a_t+\frac{\partial}{\partial x}\left(a_{xxxx}+\frac{45}{2}a_x^2+30aa_{xx}+60a^3\right)=0. \tag{3.25}$$

Equation (3.25) is known as the Kuperschmidt equation.[6] Analysis reveals that it possesses the Painlevé property. The expansion about the singularity manifold is of the form

$$a=\psi^{-2}\sum_{j=0}^{\infty}a_j\psi^j. \tag{3.26}$$

Again, there are two branches.

Branch i: $a_0=-\psi_x^2/2$: The resonances occur at

$$j=-1,3,5,6,7. \tag{3.27}$$

Branch ii: $a_0=-4\psi_x^2$: The resonances occur at

$$j=-7,-1,6,10,12. \tag{3.28}$$

We define the Bäcklund transformation about branch i:

$$a=a_0/\psi^2+a_1/\psi+a_2 \tag{3.29}$$

and find that

$$a_0=-\frac{\psi_x^2}{2}, \quad a_1=\frac{\psi_{xx}}{2}, \tag{3.30}$$

$$a_2=-\frac{1}{6}\{\psi;x\}-\frac{1}{8}\frac{\psi_{xx}^2}{\psi_x^2}, \tag{3.31}$$

$$\frac{\psi_t}{\psi_x}+\frac{\partial^2}{\partial x^2}\{\psi;x\}+\frac{1}{4}\{\psi;x\}^2=0. \tag{3.32}$$

We note that on account of the resonance structure, (3.27), (3.29)–(3.32) is *not* an overdetermined system.

Letting

$$\psi=W_1/W_2, \quad \text{where } (W_1,W_2) \text{ satisfy} \tag{3.33}$$

$$W_{xx}=-6uW, \quad W_t=bW_x+cW, \tag{3.34}$$

it is found that u satisfies Eq. (3.1).

Furthermore, if

$$v=\varphi_{xx}/\varphi_x=-\tfrac{1}{2}\psi_{xx}/\psi_x, \tag{3.35}$$

where φ satisfies Eq. (3.21) and ψ satisfies Eq. (3.32), then

$$v_t+\frac{\partial}{\partial x}(v_{xxxx}+5v_xv_{xx}-5v^2v_{xx}-5vv_x^2+v^5)=0. \tag{3.36}$$

The above implies the nonlinear transformation found in Ref. 6. For our purposes we note that (3.35) provides the transformation:

$$\psi_x=\varphi_x^{-2}. \tag{3.37}$$

Equation (3.37) indicates that Eqs. (3.21) and (3.32)

identically possess the Painlevé property. Each equation has the Painlevé property about "poles" or order -1,[2] and, about the possible movable singularities (where $\psi_x = 0$ or $\varphi_x = 0$), the transformation (3.37) provides the appropriate representation of the solution. For instance, (3.37) refers the behavior of φ at points where $\varphi_x = 0$ to the expansion of ψ at points where $\psi_x \to \infty$ (the poles of ψ). And, as is explained in the next section, this allows us to conclude that φ is single-valued at these points.

4. AN INTEGRABLE CLASS OF PARTIAL DIFFERENTIAL EQUATIONS

An equation

$$\varphi_t/\varphi_x + B(\{\varphi;x\}) = 0, \tag{4.1}$$

where $B(\{\varphi;x\})$ is a constant coefficient multinomial in $(\partial^j/\partial x^j)\{\varphi;x\}$, will identically possess the Painlevé property when there exists a transformation

$$\varphi_x = \psi_x^m, \tag{4.2}$$

where m is rational and negative and ψ satisfies an equation of the form (4.1). The form of Eq. (4.1) is sufficient to guarantee the existence of "meromorphic" expansions about the "poles" of order -1. That is,

$$\varphi = \vartheta^{-1} \sum_{j=0}^{\infty} \varphi_j \vartheta^j, \tag{4.3}$$

where the resonances occur at $j = -1, 0, 1, \ldots, n+1$ and n is the order of the highest derivative (of the Schwarzian) appearing in B. The transformation (4.2) provides a representation of the solution in a neighborhood of the points where $\varphi_x = 0$ ($\psi_x = 0$) by associating these points with the behavior of solutions of the "dual" equation in a neighborhood of their singularities.

To see the validity of the expansion (4.3), we observe that for singularities of the form (4.3) the expansion for the Schwarzian derivative begins at the constant level (is nonsingular). And, consequently, the (n) derivatives of the Schwarzian merely "shift" the recursion relations to the appropriate higher coefficient, φ_{n+2}, adding one resonance for each derivative. For particular equations of the form (4.1) higher order poles (φ^{-m}) can occur. We shall find that these singularities can be "reduced" to (4.3) through the invariance of (4.1) under the "symmetry" (4.2) and the Moebius group.

Consider forms of $B(\{\varphi;x\})$ that are linear in the highest order derivative of the Schwarzian and order the terms defining $B(\{\varphi;x\})$ into expressions that are homogeneous of the same degree under the change of variable

$$x \to a^{-1}x, \tag{4.4}$$

$$\{\varphi;x\} = \frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\left(\frac{\varphi_{xx}^2}{\varphi_x^2}\right) \Rightarrow a^2\{\varphi;x\}. \tag{4.5}$$

These are

$$\text{(i)}\quad \{\varphi;x\},$$

$$\text{(ii)}\quad \frac{\partial}{\partial x}\{\varphi;x\},$$

$$\text{(iii)}\quad \frac{\partial^2}{\partial x^2}\{\varphi;x\} + \lambda\{\varphi;x\}^2, \tag{4.6}$$

$$\text{(iv)}\quad \frac{\partial^3}{\partial x^3}\{\varphi;x\} + \lambda\{\varphi;x\}\frac{\partial}{\partial x}\{\varphi;x\},$$

$$\text{(v)}\quad \frac{\partial^4}{\partial x^4}\{\varphi;x\} + \alpha\{\varphi;x\}\frac{\partial^2}{\partial x^2}\{\varphi;x\} + \beta\left(\frac{\partial}{\partial x}\{\varphi;x\}\right)^2 + \lambda\{\varphi;x\}^3,$$

etc.

We consider equations (4.6i,ii,iii,v). Therefore, let

$$\frac{\varphi_t}{\varphi_x} + \{\varphi;x\} = \lambda \tag{4.7}$$

and

$$\varphi_x = \psi_x^m. \tag{4.8}$$

Then

$$\{\varphi;x\} = m\frac{\psi_{xxx}}{\psi_x} - \left(\frac{m^2}{2} + m\right)\frac{\psi_{xx}^2}{\psi_x^2} \tag{4.9}$$

and

$$m\psi_x^{m-1}\psi_{xt} + \frac{\partial}{\partial x}\psi_x^m\left(m\frac{\psi_{xxx}}{\psi_x} - \left(\frac{m^2}{2} + m\right)\frac{\psi_{xx}^2}{\psi_x^2}\right) = 0. \tag{4.10}$$

Direct calculation obtains

$$m\frac{\partial}{\partial x}\left(\psi_t + \psi_{xxx} - \frac{3\psi_{xx}^2}{2\psi_x} - \lambda\psi_x\right) + \left(2m - \frac{m^3}{2} - \frac{3m}{2}\right)\frac{\psi_{xx}^3}{\psi_x^2} = 0. \tag{4.11}$$

For Eq. (4.11) to be of the form (4.1),

$$2m - m^3/2 - 3m/2 = 0 \tag{4.12}$$

or

$$m = 0, \pm 1. \tag{4.13}$$

Then, if

$$\varphi_x = \psi_x^{-1}, \tag{4.14}$$

ψ will satisfy

$$\psi_t/\psi_x + \{\psi;x\} = \lambda, \tag{4.15}$$

assuming the constant of integration introduced in expression (4.11) is to vanish. For instance, we can assume that all solutions approach time-independent constants when x approaches $-\infty$.

Thus, Eqs. (4.11), (4.14), and (4.15) define a Bäcklund transformation that will be employed, with the invariance under the Moebius group, in Sec. 5 to generate rational solutions. Equation (4.7) is directly related to the KdV equation (Sec. 2).

Next, it can be readily shown that the equation

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial}{\partial x}\{\varphi;x\} = 0 \tag{4.16}$$

does not have a transformation

$$\varphi_x = \psi_x^m$$

that remains within the class (4.1). This equation, studied in Ref. 3, is transformable to an equation with complex reson-

ances (self-similar natural boundary)[7] and is thought to be nonintegrable.

It is useful to observe that, by Eq. (4.9), a transformation of type (4.2) does not change the degree of homogeneity (4.4) of the expressions in (4.6). Thus, if a transformation exists, it can only effect the value of the coefficients in (4.6).

Equations of the form

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial^2}{\partial x^2}\{\varphi;x\} + \lambda\{\varphi;x\}^2 = 0 \tag{4.17}$$

have a transformation

$$\varphi_x = \psi_x^m \tag{4.18}$$

that preserves the formulation (4.1) when

$$\begin{aligned} &\text{(i)} \quad m = -1, \quad \lambda = \tfrac{3}{2}, \\ &\text{(ii)} \quad m = -2, \quad \lambda = \tfrac{1}{4}, \\ &\text{(iii)} \quad m = -\tfrac{1}{2}, \quad \lambda = 4. \end{aligned} \tag{4.19}$$

Equation (4.19i) is (essentially) the first higher-order (fifth degree) KdV equation.[2] Equation (4.19i,ii) are (obtained from) the Kuperschmidt and Caudrey–Dodd–Gibbon equations, respectively (see Sec. 3). Then, the Kuperschmidt equation and Caudrey–Dodd–Gibbon equation are, in a sense, "dual" under the transformation

$$\psi_x = \varphi_x^{-2}. \tag{4.20}$$

The KdV equation (4.7) and fifth-degree higher-order KdV equation (4.19i) are then "self-dual."

We note that the property of possessing a transformation within the class (4.1) is additive (by construction) for expressions with the same value of exponent m.

Thus, by (4.19i) and (4.14) the equation

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial^2}{\partial x^2}\{\varphi;x\} + \frac{3}{2}\{\varphi;x\}^2 + \lambda\{\varphi;x\} = 0 \tag{4.21}$$

has, for any λ, an (auto) Bäcklund transform

$$\varphi_x = \psi_x^{-1}. \tag{4.22}$$

Finally, the equation

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial^4}{\partial x^4}\{\varphi;x\} + \alpha\{\varphi;x\}\frac{\partial^2}{\partial x^2}\{\varphi;x\} + \beta\left(\frac{\partial}{\partial x}\{\varphi;x\}\right)^2 + \lambda\{\varphi;x\}^3 \tag{4.23}$$

has a transformation

$$\varphi_x = \psi_x^m \tag{4.24}$$

preserving the form of Eq. (4.23) when

$$\begin{aligned} &\text{(i)} \quad m = -1, \quad \alpha = 5, \quad \beta = \tfrac{5}{2}, \quad \lambda = \tfrac{5}{2}, \\ &\text{(ii)} \quad m = -2, \quad \alpha = \tfrac{3}{2}, \quad \beta = \tfrac{3}{4}, \quad \lambda = \tfrac{1}{6}, \\ &\text{(iii)} \quad m = -\tfrac{1}{2}, \quad \alpha = 12, \quad \beta = 6, \quad \lambda = \tfrac{32}{5}. \end{aligned} \tag{4.25}$$

These are higher order KdV, Kuperschmidt, and Caudrey–Dodd–Gibbon equations, respectively. Further information concerning Eq. (4.23) is contained in Appendix B.

We now consider the sequence of higher-order KdV equations determined by the "Lenard recursion relation"[8]

$$\frac{\partial}{\partial x} b^{n+1} = b^n_{xxx} + 2ub^n_x + u_x b^n, \tag{4.26}$$

where

$$u_t + \frac{\partial}{\partial x} b^{n+1}(u) = 0 \tag{4.27}$$

for $n = 1,2,3,\ldots$ are the sequence of higher-order KdV equations and

$$\begin{aligned} b^0 &= 1, \\ b^1 &= u, \\ b^2 &= u_{xx} + \tfrac{3}{2}u^2, \\ b^3 &= u_{xxxx} + 5uu_{xx} + \tfrac{5}{2}u_x^2 + \tfrac{5}{2}u^3. \end{aligned} \tag{4.28}$$

Now inspection of Eqs. (4.7), (4.17), and (4.23) leads us to formulate the following.

Theorem 1: The sequence of higher-order KdV equations

$$u_t + \frac{\partial}{\partial x} b^{n+2}(u) = 0 \tag{4.29}$$

for $n = 0,1,2,\ldots$ has the following Bäcklund transformation:

$$u = 4\frac{\partial^2}{\partial x^2}\ln\varphi + u_2, \tag{4.30}$$

$$u_2 = -\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2, \tag{4.31}$$

$$\frac{\varphi_t}{\varphi_x} + b^{n+1}(\{\varphi;x\}) = 0. \tag{4.32}$$

Furthermore,

$$\omega = \{\varphi;x\} \tag{4.33}$$

(and u_2) satisfies Eqs. (4.29) and (4.32) is invariant under the transformation

$$\varphi_x = \psi_x^{-1}. \tag{4.34}$$

Note: To simplify the statement of the above results, we require the sequence of b^n to be defined by precisely Eq. (4.26). "Scalings" in the argument "u" of Eq. (4.26) is essential for the definition of Eq. (4.32), but not for Eq. (4.29).

Proof: We prove the above by the following observations: For each n, let

$$V = \varphi_{xx}/\varphi_x. \tag{4.35}$$

Then Eq. (4.32) obtains the "higher-order modified KdV equation"

$$V_t + \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x} + V\right) b^{n+1}(V_x - \tfrac{1}{2}V^2) = 0, \tag{4.36}$$

where

$$\omega = \{\varphi;x\} = V_x - \tfrac{1}{2}V^2. \tag{4.37}$$

From Eqs. (4.36) and (4.37) we find that

$$\omega_t + \left(\frac{\partial^3}{\partial x^3} + 2\omega\frac{\partial}{\partial x} + \omega_x\right) b^{n+1}(\omega) = 0, \tag{4.38}$$

or, using Eq. (4.26),

$$\omega_t + \frac{\partial}{\partial x} b^{n+2}(\omega) = 0. \tag{4.39}$$

This equation (4.32) implies that ω is a solution of Eq. (4.29). From Eqs. (4.31) and (4.35)

$$u_2 = -V_x - \tfrac{1}{2}V^2. \tag{4.40}$$

Now, if

$$\varphi_x \to \varphi_x^{-1}, \tag{4.41}$$

then

$$V \to -V, \quad u_2 \to \omega, \tag{4.42}$$

$$w \to u_2. \tag{4.43}$$

Hence, both u_2 and ω will be solutions of Eq. (4.29) if Eq. (4.32) is invariant under (4.41), or equivalently, if Eq. (4.36) is invariant under (4.42).

To see this, we let

$$D = \frac{\partial}{\partial x}, \tag{4.44}$$

$$M_v = D(D+V), \tag{4.45}$$

$$L_v = D^{-1}(D-V)M_v, \tag{4.46}$$

and find that the Lenard relationship (4.26) becomes

$$b^{n+2}(V_x - \tfrac{1}{2}V^2) = L_v b^{n+1}(V_x - \tfrac{1}{2}V^2) \tag{4.47}$$

while Eq. (4.36) is

$$V_t + M_v b^{n+1}(V_x - \tfrac{1}{2}V^2) = 0. \tag{4.48}$$

The condition of invariance of (4.48) under (4.42) reads

$$M_v b^{n+1}(V_x - \tfrac{1}{2}V^2) + M_{-v} b^{n+1}(-V_x - \tfrac{1}{2}V^2) = 0. \tag{4.49}$$

We verify (4.49) by induction. Previous calculations demonstrate (4.49) for $n = 0,1$. We assume (4.49) with $n = 0,1,2,\ldots,m-1$. Then with $n = m$ and, using (4.47), Eq. (4.49) is

$$M_v L_v b^m(V_x - \tfrac{1}{2}V^2) + M_{-v} L_{-v} b^m(-V_x - \tfrac{1}{2}V^2) = 0. \tag{4.50}$$

However, from (4.46),

$$M_v L_v = I_v M_v, \tag{4.51}$$

where

$$I_v = D(D+V)D^{-1}(D-V). \tag{4.52}$$

Using the identity for constants a,b,

$$(D+aV)D^{-1}(D+bV) = (D+bV)D^{-1}(D+aV), \tag{4.53}$$

it is found that

$$I_v = I_{-v} \tag{4.54}$$

and, with (4.51), Eq. (4.50) is

$$I_v\{M_v b^m(V_x - \tfrac{1}{2}V^2) + M_{-v} L_{-v} b^m(-V_x - \tfrac{1}{2}V^2)\} = 0. \tag{4.55}$$

Since the term in brackets vanishes by assumption, (4.49) is verified for $n = m$. We note that (4.52) is a recursion operator for the higher-order modified KdV equations.

Equation (4.32) and the invariance (4.34) obtain that (ω,u_2) are solutions of Eq. (4.29). We now show that Eqs. (4.32), (4.31), and (4.30) imply that u [defined in Eq. (4.30)] will be a solution of Eq. (4.29), completing the proof of the existence of the Bäcklund transform.

To begin, we note that Eq. (4.32) is invariant under the Moebius group.

Letting

$$\varphi = 1/\psi, \tag{4.56}$$

we find that ψ satisfies Eq. (4.32) and that

$$u_2 = 4\frac{\partial^2}{\partial x^2}\ln\psi + u, \tag{4.57}$$

$$u_2 = -\frac{\partial}{\partial x}\left(\frac{\psi_{xx}}{\psi_x}\right) - \frac{1}{2}\left(\frac{\psi_{xx}}{\psi_x}\right)^2 + 4\frac{\partial^2}{\partial x^2}\ln\psi, \tag{4.58}$$

or

$$u = -\frac{\partial}{\partial x}\left(\frac{\psi_{xx}}{\psi_x}\right) - \frac{1}{2}\left(\frac{\psi_{xx}}{\psi_x}\right)^2. \tag{4.59}$$

By the previous calculation Eq. (4.59) implies u satisfies Eq. (4.29), completing the proof.

Remark 1: Equation (4.32) effectively defines three distinct solutions of Eq. (4.29). That is,

$$u_2,\omega = \{\varphi;x\}$$

and

$$u = 4\frac{\partial^2}{\partial x^2}\ln\varphi + u_2. \tag{4.60}$$

Remark 2: If we consider the stationary solutions of a higher-order KdV equation[8] Theorem 1 defines Bäcklund transformations for the associated ordinary differential equations. Furthermore, to construct solutions of the $(n+2)$ equation

$$\frac{\partial}{\partial x} b^{n+2}(u) = 0, \tag{4.61}$$

we integrate the $(n+1)$ equation

$$b^{n+1}(\omega) = 0 \tag{4.62}$$

and set

$$\omega = \{\varphi;x\}. \tag{4.63}$$

Then

$$\varphi = V_1/V_2, \tag{4.64}$$

where V_1 and V_2 satisfy

$$V_{xx} = -\tfrac{1}{2}wV, \tag{4.65}$$

defines the solutions (u,u_2) of (4.61).

Thus the solution of the $(n+1)$ equation is the "potential" in a associated linear, Schrödinger equation, that defines the solutions and Bäcklund transforms for the $(n+2)$ equation. Further consideration of these Bäcklund transforms for (Painlevé) ODE's and the iterative construction of solutions seems warranted.

We now generalize Theorem 1 to allow for the inclusion of a spectral parameter, λ.

Theorem 2: The sequence of higher-order KdV equations

$$u_t + \frac{\partial}{\partial x} b^{n+2}(u) = 0 \tag{4.66}$$

for $n = 0,1,2,\cdots$ has the Bäcklund transformation

$$u = 4\frac{\partial^2}{\partial x^2}\ln\varphi + u_2, \tag{4.67}$$

$$u_2 = -\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2 + \lambda, \tag{4.68}$$

$$\frac{\varphi_t}{\varphi_x} + \alpha_{n+1,j}\, b^j(\{\varphi;x\}) = 0, \tag{4.69}$$

where $\alpha_{n+1,j} = \alpha_{n+1,j}(\lambda)$, with a summation convention over $j = 0,1,\ldots,n+1$.

Furthermore,

$$w = \{\varphi;x\} + \lambda \tag{4.70}$$

satisfies equation (4.66) and equation (4.69) is invariant under the transform

$$\varphi_x = \psi_x^{-1}. \tag{4.71}$$

Proof: By a previous remark invariance (4.71) follows immediately from (4.34). Now, let

$$V = \varphi_{xx}/\varphi_x. \tag{4.72}$$

Then

$$V_t + \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x} + V\right)\alpha_{n+1,j}b^j\left(V_x - \frac{1}{2}V^2\right) \tag{4.73}$$

and

$$\omega_t + \left(\frac{\partial^3}{\partial x^3} + 2(\omega-\lambda)\frac{\partial}{\partial x} + \omega_x\right)\alpha_{n+1,j}b^j(\omega-\lambda) = 0. \tag{4.74}$$

By (4.26)

$$b_x^{j+1}(\omega-\lambda) = \left(\frac{\partial^3}{\partial x^3} + 2(\omega-\lambda)\frac{\partial}{\partial x} + \omega_x\right)b^j(\omega-\lambda). \tag{4.75}$$

Lemma 1:

$$b^j(\omega-\lambda) = \sum_{k=0}^{j} a_{jk}b^k(\omega),$$

where

$$a_{jj} = 1,$$

$$a_{j0} = -\lambda\,((2j-1)/j)a_{j-1,0}$$

and

$$a_{j,k} = a_{j-1,k-1} - 2\lambda a_{j-1,k}, \quad \text{where } k < j.$$

Proof: By induction, using (4.26).

Now using Eqs. (4.70), (4.71), Lemma 1, and requiring that ω satisfy Eq. (4.66) determines, for each n, the $\alpha_{n+1,j}$, $j = 0,1,\ldots,n+1$.

We find the following triangular system of linear equations for

$$\alpha_{n+1} = \begin{pmatrix} \alpha_{n+1,n+1} \\ \alpha_{n+1,n} \\ \alpha_{n+1,n-1} \\ \vdots \\ \alpha_{n+1,0} \end{pmatrix}, \tag{4.76}$$

$$\begin{pmatrix} 1 & 0 & & 0 \\ a_{n+2,m+1} & 1 & & \\ a_{n+2,m}\ a_{m+1,m} & & 1 & \\ \vdots & & \ddots & \\ a_{m+2,k}\ a_{m+1,k} & & & 1\ \ 0 \\ a_{m+2,1}\ a_{n+1,1} & & & 1 \end{pmatrix} a_{n+1} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.77}$$

Since the system (4.77) is always solvable, α_{n+1} exists for each n, and the Bäcklund transformations are well defined, completing the proof.

For reference, we present the following tables:

	a_{jk}				
j/k	0	1	2	3	4
0	1				
1	$-\lambda$	1			
2	$\frac{3}{2}\lambda^2$	-3λ	1		
3	$-\frac{5}{2}\lambda^3$	$\frac{15}{2}\lambda^2$	-5λ	1	
4	$\frac{35}{8}\lambda^4$	$\frac{-35}{2}\lambda^3$	$\frac{35}{2}\lambda^2$	-7λ	1

	$\alpha_{j,k}$			
j/k	0	1	2	3
1	3λ	1		
2	$\frac{15}{2}\lambda^2$	5λ	1	
3	$\frac{35}{2}\lambda^3$	$\frac{35}{2}\lambda^2$	7λ	1

We next consider the sequence of higher-order Caudrey–Dodd–Gibbon and Kuperschmidt equations. Again, to avoid unnecessary complexity, we consider these equations with a specific scaling. With reference to Sec. 3, we let

$$u \to u/12, \qquad a \to a/3, \tag{4.78}$$

and find the Caudrey–Dodd–Gibbon equation

$$u_t + \frac{\partial}{\partial x}\left(u_{xxxx} + \frac{5}{2}uu_{xx} + \frac{5}{12}u^3\right) = 0 \tag{4.79}$$

and the Kuperschmidt equation

$$a_t + \frac{\partial}{\partial x}\left(a_{xxxx} + 10aa_{xx} + \frac{15}{2}a_x^2 + \frac{20}{3}a^3\right) = 0. \tag{4.80}$$

From Ref. 9 the sequences of conserved covariants (functional gradients of conserved densities) are given by

$$G_{n+2} = J_1(u)\Theta_1(u)G_n, \tag{4.81}$$

$$H_{n+2} = J_2(a)\Theta_2(a)H_n \tag{4.82}$$

for the Caudrey–Dodd–Gibbon and Kuperschmidt equations, respectively, where

$$\Theta_1 = D^3 + 2uD + u_x,$$
$$J_1 = D^3 + \tfrac{1}{2}D^2uD^{-1} + \tfrac{1}{2}D^{-1}uD^2 + \tfrac{1}{8}(u^2D^{-1} + D^{-1}u^2), \tag{4.83}$$

and

$$\Theta_2 = D^3 + 2uD + u_x,$$
$$J_2 = D^3 + 3(uD + Du) + 2(D^2uD^{-1} + D^{-1}uD^2) + 8(u^2D^{-1} + D^{-1}u^2). \tag{4.84}$$

With the normalization that we employ,

$$G_0 = 1, \quad H_0 = 1,$$
$$G_1 = u_{xx} + \tfrac{1}{4}u^2, \quad H_1 = a_{xx} + 4a^2, \tag{4.85}$$

and Eqs. (4.79) and (4.80) are

$$u_t + \Theta_1 G_1(u) = 0,$$
$$a_t + \Theta_2 H_1(a) = 0. \tag{4.86}$$

Furthermore, the respective sequences of higher-order equations are given by

$$u_t + \Theta_1 G_n(u) = 0,$$
$$a_t + \Theta_2 H_n(a) = 0. \tag{4.87}$$

For what follows it is convenient, as was the case for the KdV equations, to "factorize" the recursion operators. That is,

$$\Theta_1 = (D - W)D(D + W), \tag{4.88}$$
$$J_1 = D^{-1}\{(D - W/2)(D + W/2) \times D(D - W/2)(D + W/2)\}D^{-1},$$

and

$$\Theta_2 = (D - V)D(D + V), \tag{4.89}$$
$$J_2 = D^{-1}\{(D - 2V)(D - V) \times D(D + V)(D + 2V)\}D^{-1},$$

where

$$u = W_x - \tfrac{1}{2}W^2,$$
$$a = V_x - \tfrac{1}{2}V^2. \tag{4.90}$$

We now formulate the following.

Theorem 3: The sequences of higher-order Caudrey–Dodd–Gibbon and Kuperschmidt equations

$$u_t + \Theta_1 G_n(u) = 0,$$
$$a_t + \Theta_2 H_n(a) = 0, \tag{4.91}$$

for $n = 1,2,3,\cdots$, have the following Bäcklund transformations:

$$u = 12\frac{\partial^2}{\partial x^2}\ln\varphi + u_2,$$
$$a = \frac{3}{2}\frac{\partial^2}{\partial x^2}\ln\psi + a^2, \tag{4.92}$$

where

$$u_2 = -2\frac{\varphi_{xxx}}{\varphi_x},$$
$$a_2 = -\frac{1}{2}\left(\frac{\psi_{xxx}}{\psi_x} - \frac{3}{4}\frac{\psi_{xx}^2}{\psi_x^2}\right) \tag{4.93}$$

and

$$\frac{\varphi_t}{\varphi_x} + H_n(\{\varphi;x\}) = 0,$$
$$\frac{\psi_t}{\psi_x} + G_n(\{\psi;x\}) = 0. \tag{4.94}$$

Furthermore, Eqs. (4.94) possess the symmetry

$$\psi_x = \varphi_x^{-2}, \tag{4.95}$$

and

$$u_3 = \{\psi;x\},$$
$$a_3 = \{\phi;x\} \tag{4.96}$$

are solutions of Eq. (4.91), respectively.

Proof: (i) The sequences of higher-order modified Caudrey–Dodd–Gibbon and Kuperschmidt equations are given by

$$W_t + M_w G_n(W_x - \tfrac{1}{2}W^2) = 0,$$
$$V_t + M_v H_n(V_x - \tfrac{1}{2}V^2) = 0, \tag{4.97}$$

respectively, where

$$W = \psi_{xx}/\psi_x,$$
$$V = \varphi_{xx}/\varphi_x, \tag{4.98}$$

and

$$M_v = D(D + V). \tag{4.99}$$

Since

$$u_3 = W_x - \tfrac{1}{2}W^2,$$
$$a_3 = V_x - \tfrac{1}{2}V^2, \tag{4.100}$$

the factorizations (4.88) and (4.89) show

$$u_{3t} + \Theta_1 G_n(u_3) = 0,$$
$$a_{3t} + \Theta_2 H_n(a_3) = 0. \tag{4.101}$$

(ii) Now *if* (4.95) is valid, then, as is readily verified,

$$u_2 = \{\psi;x\},$$
$$a_2 = \{\varphi;x\}, \tag{4.102}$$

and by the above (u_2,a_2) solve Eqs. (4.91). Now, the invariance of Eqs. (4.94) under the Moebius group, (4.92) and (4.93) imply

$$u = -2\tilde{\varphi}_{xxx}/\tilde{\varphi}_x,$$
$$a = -\tfrac{1}{2}(\tilde{\psi}_{xxx}/\tilde{\psi}_x - \tfrac{3}{4}\tilde{\psi}_{xx}^2/\tilde{\psi}_x^2), \tag{4.103}$$

where

$$\tilde{\varphi} = 1/\varphi, \quad \tilde{\psi} = 1/\psi \tag{4.104}$$

and $(\tilde{\varphi},\tilde{\psi})$ are solutions of (4.94). By the above, (u,a) are solutions of (4.91), and (4.92) is well defined if (4.95) is verified.

(iii) By (4.98), (4.95) is equivalent to the condition

$$W = -2V, \tag{4.105}$$

or, using (4.97), to

$$2M_v H_n(V_x - \tfrac{1}{2}V^2) + M_{-2v}G_n(-2V_x - 2V^2) = 0. \tag{4.106}$$

We verify (4.106) by induction. Previous calculations demonstrate (4.106) for $n = 1,2$. We assume (4.106) valid for $n = 1,2,\ldots,m$; then, by (4.81) and (4.82),

$$2M_v H_{m+1}(V_x - \tfrac{1}{2}V^2) + M_{-2v}G_{m+1}(-2V_x - 2V^2) = 2M_v J_2(a)\Theta_2(a)H_{m-1}(V_x - \tfrac{1}{2}V^2) + M_{-2v}J_1(\tilde{u})\Theta_1(\tilde{u})G_{m-1}(-2V_x - 2V^2), \tag{4.107}$$

where

$$a = V_x - \tfrac{1}{2}V^2, \quad \tilde{u} = -2V_x - 2V^2.$$

However, (4.88) and (4.89) readily obtain

$$M_v J_1 \Theta_1 = \lambda_v M_v, \quad M_v J_2 \Theta_2 = \Phi_v M_v, \tag{4.108}$$

where

$$\lambda_v = D(D+V)D^{-1}\{(D-V/2)(D+V/2) \times D(D-V/2)(D+V/2)\}D^{-1}(D-V) \tag{4.109}$$

and

$$\Phi_v = D(D+V)D^{-1}\{(D-2V)(D-V) \times D(D+V)(D+2V)\}D^{-1}(D-V). \tag{4.110}$$

The identity [by (4.53)]

$$\lambda_{-2v} = \Phi_v \tag{4.111}$$

and (4.106) for $n = m-1$ imply that (4.107) vanishes, verifying (4.106), (4.105) and completing the proof.

We note that, in another context, the method of factorization of operators has been used to derive Miura transformations and Hamiltonian structures.[10–12]

Remark 3: It is not known whether the sequences of KdV, Caudrey–Dodd–Gibbon, and Kuperschmidt equations exhaust the equations in the class (4.1). Presumably, there may exist a sequence of equations for every index pair, $(m, 1/m)$, $m = -1, -2, -3, \cdots$.

We conclude this section with some remarks concerning the nature of the higher-order poles for the class of equations considered herein. For instance, the sequence of KdV equations, (4.32), can have singularities of the form

$$\varphi = \varphi_0 \epsilon^{-N} + \varphi_1 \epsilon^{-N+1} + \cdots, \tag{4.112}$$

where it is not assumed that (4.112) is Painlevé.

For simplicity we employ the "reduced" expansion[1]

$$\epsilon = x - \psi(t), \quad \varphi_j = \varphi_j(t). \tag{4.113}$$

Now, since (4.32) is invariant under the Moebius group, the transformation

$$\psi = 1/\varphi \tag{4.114}$$

produces a solution which has an expansion

$$\psi = \psi_0 \epsilon^N + \psi_1 \epsilon^{N+1} + \cdots. \tag{4.115}$$

Furthermore, the symmetry

$$\varphi_x = \psi_x^{-1}$$

obtains

$$\varphi_x = \varphi_0 \epsilon^{-N+1} + \cdots, \tag{4.116}$$

$$\varphi = \varphi_0 \epsilon^{-N+2} + \cdots. \tag{4.117}$$

If N is an odd integer, after a finite number of steps, there results

$$\varphi = \varphi_0 \epsilon^{-1} + \cdots. \tag{4.118}$$

However, singularities of the form (4.118) identically possess the Painlevé property. Now $\ln \epsilon$ terms could arise in going from (4.112) to (4.118) [but do not, since (4.118) is Painlevé with the complete set of "arbitrary functions"]. However, no $\ln \epsilon$ terms can occur in going from (4.118) to (4.112). Thus, (4.112), as reconstructed from (4.118), has the Painlevé property (when N is odd). [Note in going from (4.118) to (4.112) Taylor, not Laurent, series are integrated.]

Let us now assume that

$$\varphi \approx \varphi_0 \epsilon^{m+1}. \tag{4.119}$$

Then,

$$\{\varphi; x\} \approx -\tfrac{1}{2}m(m+2)\epsilon^{-2}, \tag{4.120}$$

and using the Lenard formula (4.26) with

$$b^n(\{\varphi; x\}) \approx P^n(m)\epsilon^{-2n}$$

obtains

$$(2n+2)P^{n+1}(m) = 2(2n+1)(\lambda + n(n+2))P^n(m) \tag{4.121}$$

where $\lambda = -\tfrac{1}{2}m(m+2)$.

Thus, each higher-order equation of order $(n+1)$ acquires two new leading orders

$$\lambda = -\tfrac{1}{2}m(m+2) = -n(n+2) \tag{4.122}$$

or

$$m = 2n, -2-2n,$$

where

$$\varphi \approx \varphi_0 \epsilon^{2n+1} \tag{4.123}$$

or

$$\varphi \approx \varphi_0 \epsilon^{-2n-1}.$$

The higher-order KdV equations (in the Schwarzian formulation) can have only odd integral leading orders, and by the previous remarks these have the Painlevé property.

Considerations of a similar nature determine that the higher-order singularities of the Caudrey–Dodd–Gibbon and Kuperschmidt sequences, again, "reduce" to singularities of the (Painlevé) form (4.118). Thus, these equations identically possess the Painlevé property.

5. ITERATIVE CONSTRUCTION OF RATIONAL SOLUTIONS

For the KdV equation

$$u_t + \frac{\partial}{\partial x}\left(\frac{u^2}{2} + u_{xx}\right) = 0, \tag{5.1}$$

the Bäcklund transform

$$u = 12\frac{\partial^2}{\partial x^2}\ln\varphi + u_2 \tag{5.2}$$

implies that

$$\frac{\varphi_t}{\varphi_x} + \{\varphi; x\} = \lambda. \tag{5.3}$$

Equation (5.3) is invariant under:

(i) The Moebius group

$$\varphi = \frac{a\psi + b}{c\psi + d} \tag{5.4}$$

and the transformation

$$\text{(ii)} \quad \varphi_x = \psi_x^{-1}. \tag{5.5}$$

Combining Eq. (5.4), i.e.,

$$\psi = -1/\varphi, \tag{5.6}$$

and Eq. (5.5), there is defined the Bäcklund transformation

$$\varphi_{n+1,x} = \varphi_n^2/\varphi_{n,x}. \tag{5.7}$$

Without loss of generality (modulo a Galilean transformation) we set $\lambda = 0$ in Eq. (5.3). Then setting

$$\varphi_0 = x, \tag{5.8}$$

it is found from Eqs. (5.7) and (5.3) that

$$\varphi_1 = x^3/3 + 4t. \tag{5.9}$$

We normalize (5.9) by setting

$$\varphi_1 = x^3 + 12t. \tag{5.10}$$

From Eq. (5.7) it is found that (after normalization)

$$\varphi_2 = (x^6 + 60tx^3 + ex - 720t^2)/x \tag{5.11}$$

and

$$\varphi_3 = 1/\varphi_1[x^{10} + 180tx^7 + 302\,400t^3x + 7e(x^5 - 60tx^3 - e/3) + f\varphi_1], \tag{5.12}$$

where (e,f) are constants of integration.

Equations (5.8)–(5.12) suggest that

$$\varphi_n = P_n/P_{n-2}, \tag{5.13}$$

where the P_j are polynomials in (x,t). Substitution of (5.13) into (5.7) obtains

$$P_{n-1}P_{n+1,x} - P_{n-1,x}P_{n+1} = P_n^2, \tag{5.14}$$

where

$$\begin{aligned} P_0 &= x, \\ P_1 &= x^3 + 12t, \\ P_2 &= x^6 + 60tx^3 - 720t^2 + ex. \end{aligned} \tag{5.15}$$

The solutions obtained from (5.14) and (5.15) are (essentially) those rational solutions of the KdV equation found by Ablowitz and Segur,[13] using Hirota's method, and are equivalent to rational solutions of Airault, McKean and Moser.[14]

From Eqs. (5.13) and (5.2) we find that

$$\begin{aligned} u &= 12\frac{\partial^2}{\partial x^2}\ln P_n, \\ u_2 &= 12\frac{\partial^2}{\partial x^2}\ln P_{n-2} \end{aligned} \tag{5.16}$$

define rational solution of the KdV equations.

For the Caudrey–Dodd–Gibbon equation (3.1) and the Kuperschmidt equation (3.25), there are defined the following Bäcklund transformations:

$$u = \frac{\partial^2}{\partial x^2}\ln\varphi + u_2 \tag{5.17}$$

and

$$a = \frac{1}{2}\frac{\partial^2}{\partial x^2}\ln\psi + a_2, \tag{5.18}$$

respectively, where

$$\frac{\varphi_t}{\varphi_x} + \frac{\partial^2}{\partial x^2}\{\varphi;x\} + 4\{\varphi;x\}^2 = 0 \tag{5.19}$$

and

$$\frac{\psi_t}{\psi_x} + \frac{\partial^2}{\partial x^2}\{\psi;x\} + \frac{1}{4}\{\psi;x\}^2 = 0. \tag{5.20}$$

Using the transformation

$$\psi_x = \varphi_x^{-2} \tag{5.21}$$

and invariance under the Moebius group, we find the following Bäcklund transformation:

$$\varphi_{n,x} = \psi_n/\psi_{n,x}^{1/2}, \tag{5.22}$$

$$\psi_{n,x} = \varphi_{n-1}^4/\varphi_{n-1,x}^2. \tag{5.23}$$

Letting

$$\varphi_n = P_n/P_{n-1} \tag{5.24}$$

and

$$\psi_n = Q_n/Q_{n-1} \tag{5.25}$$

obtains

$$P_{n-1}P_{n,x} - P_{n-1,x}P_n = Q_n, \tag{5.26}$$

$$Q_{n-1}Q_{n,x} - Q_{n-1,x}Q_n = P_{n-1}^4. \tag{5.27}$$

It is readily found that

$$\begin{aligned} &P_0 = 1, \quad Q_0 = 1, \\ &P_1 = x, \quad Q_1 = 1, \\ &P_2 = x^5 - 720t, \quad Q_2 = x^5 + 180t. \end{aligned} \tag{5.28}$$

are the first terms (after normalization) that satisfy Eqs. (5.26) and (5.27) and define (rational) solutions of Eqs. (5.19) and (5.20).

APPENDIX A: LAX PAIR AND BÄCKLUND TRANSFORMATIONS FOR THE CAUDREY–DODD–GIBBON EQUATION

In Sec. 3 the Caudrey–Dodd–Gibbon equation

$$u_t + \frac{\partial}{\partial x}(u_{xxxx} + 30uu_{xx} + 60u^3) = 0 \tag{A1}$$

was found to have the Bäcklund transformation

$$u = \frac{\partial^2}{\partial x^2}\ln\varphi + u_2, \tag{A2}$$

where u_2 satisfies (A1) and

$$\text{(i)}\quad u_2 = -\frac{1}{6}\frac{\varphi_{xxx}}{\varphi_x}, \tag{A3}$$

$$\text{(ii)}\quad \frac{\varphi_t}{\varphi_x} + \frac{\partial^2}{\partial x^2}\{\varphi;x\} + 4\{\varphi;x\}^2 = 0. \tag{A4}$$

Equations (A3) and (A4) may be rewritten as the following "Lax pair":

$$\varphi_{xxx} + 6u_2\varphi_x = 0, \tag{A5}$$

$$\varphi_t = -18u_{2x}\varphi_{xx} + 6(u_{2xx} - 6u_2^2)\varphi_x. \tag{A6}$$

With the exception that the spectral parameter vanishes, this is the Lax pair found in Ref. 4.

To obtain a Lax pair with the spectral parameter, it is necessary to generalize the procedures introduced in Ref. 2. That is, we define a Bäcklund transformation (A2), where (u,u_2) satisfy (A1). In Sec. 3 the resulting expressions were ordered according to the inverse powers of φ, i.e., (3.6iii, iv, and v). Herein, other than requiring that u_2 satisfy (A1) the various terms are collected into a single equation, obtaining

$$\frac{\partial^2}{\partial x^2}\left(\frac{\varphi_t}{\varphi}\right)+\frac{\partial}{\partial x}\left(\frac{H_5}{\varphi}+\frac{H_4}{\varphi^2}\right)=0, \tag{A7}$$

where

$$H_4=-\varphi_x^2\left\{\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2 + 5(\vartheta_{xx}+\vartheta^2+2\{\varphi;x\}\vartheta)\right\}, \tag{A8}$$

$$H_5=\varphi_x\left\{\frac{\partial^3}{\partial x^3}\{\varphi;x\}+4\frac{\partial}{\partial x}\{\varphi;x\}^2+5\vartheta\frac{\partial}{\partial x}\{\varphi;x\}\right\} + \vartheta_{xx}\left\{\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2+5\vartheta_{xx} + 5\vartheta^2+10\{\varphi;x\}\vartheta\right\}, \tag{A9}$$

$$\vartheta=\{\varphi;x\}+6W, \tag{A10}$$

and

$$W=u_2+\frac{1}{4}\frac{\varphi_{xx}^2}{\varphi_x^2}. \tag{A11}$$

Now, letting

$$\vartheta=6\lambda\varphi/\varphi_x, \tag{A12}$$

it is found from (A10) and (A11) that

$$\varphi_{xxx}+6u_2\varphi_x=6\lambda\varphi. \tag{A13}$$

From (A7)–(A9) and (A12) there results

$$\frac{\partial^2}{\partial x^2}\left\{\frac{\varphi_t}{\varphi}+\frac{\varphi_x}{\varphi}\left(\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2+30\lambda\frac{\varphi\varphi_{xxx}}{\varphi_x^2} - 30\lambda\frac{\varphi\varphi_{xx}^2}{\varphi_x^3}-30\lambda\frac{\varphi_{xx}}{\varphi_x}-180\lambda^2\frac{\varphi^2}{\varphi_x^2}\right)\right\}=0. \tag{A14}$$

Setting the term inside the bracket equal to 0,

$$\frac{\varphi_t}{\varphi_x}+\frac{\partial^2}{\partial x^2}\{\varphi;x\}+4\{\varphi;x\}^2+30\lambda\frac{\varphi\varphi_{xxx}}{\varphi_x^2} - 30\lambda\frac{\varphi\varphi_{xx}^2}{\varphi_x^3}-30\lambda\frac{\varphi_{xx}}{\varphi_x}-180\lambda^2\frac{\varphi^2}{\varphi_x^2}=0. \tag{A15}$$

Using (A13),

$$\varphi_t=(54\lambda-18u_{2x})\varphi_{xx}+6(u_{2xx}-6u_2^2)\varphi_x+216\lambda u_2\varphi. \tag{A16}$$

Equations (A13) and (A16) constitute the Lax pair for the Caudrey–Dodd–Gibbon equation,[4] where λ is the spectral parameter. We note that Eq. (A15) is not invariant under the Moebius group.

APPENDIX B: SOME SEVENTH-ORDER EQUATIONS

We consider when the equation

$$\frac{\varphi_t}{\varphi_x}+\frac{\partial^4}{\partial x^4}\{\varphi;x\}+\alpha\{\varphi;x\}\frac{\partial^2}{\partial x^2}\{\varphi;x\} +\beta\left(\frac{\partial}{\partial x}\{\varphi;x\}\right)^2+\lambda\{\varphi;x\}^3=0 \tag{B1}$$

has a transformation

$$\varphi_x=\psi_x^m \tag{B2}$$

preserving the form of (B1).

Directly,

$$\{\varphi;x\}=m\frac{\psi_{xxx}}{\psi_x}-\left(\frac{m^2}{2}+m\right)\frac{\psi_{xx}^2}{\psi_x^2} \tag{B3}$$

and

$$\varphi_{xt}=m\psi_x^{m-1}\psi_{xt}. \tag{B4}$$

We note that

$$m\psi_x^{m-1}\psi_{xt}=\frac{\partial}{\partial x}(\psi_x^m F) =\psi_x^m\frac{\partial}{\partial x}F+m\psi_x^{m-1}\psi_{xx}F \tag{B5}$$

or

$$m\psi_{xt}=\psi_x\frac{\partial}{\partial x}F+m\psi_{xx}F. \tag{B6}$$

Therefore, for Eq. (B6) to be of the form (B1)

$$\psi_x\frac{\partial}{\partial x}F+m\psi_{xx}F=\frac{\partial}{\partial x}G, \tag{B7}$$

where G is a functional of ψ_x. Expressions on the lhs of (B7) that are not "gradients" must vanish. In this case, we find:

(i) Term $\psi_{xx}\psi_{xxxx}^2/\psi_x^2$ obtains the condition

$$2m+7+2m(\alpha-\beta)=0. \tag{B8}$$

(ii) Term $\psi_{xx}\psi_{xxx}^3/\psi_x^3$ obtains the condition

$$17m+42+\tfrac{1}{2}\alpha m(9m+28)-6\beta m(m+3)-3\lambda m^2=0. \tag{B9}$$

(iii) Term $\psi_{xxx}^3\psi_{xxx}^2/\psi_x^4$ obtains the condition

$$-39m-84+\alpha m(3m^2-\tfrac{9}{2}m-25) -2\beta m(m^2-5m-16)+3\lambda m^2(m+2)=0. \tag{B10}$$

(iv) Term ψ_{xx}^7/ψ_x^6 obtains the condition

$$60(m+2)-\tfrac{1}{2}\alpha m(13m^2-8m-68)+2\beta m(2m^2-7m-22) +\tfrac{3}{4}\lambda m^2(m^3+m^2-8m-12)=0. \tag{B11}$$

Equation (B8)–(B11) have the following solutions:

(i) $m=-1,\quad \alpha=\beta+\frac{5}{2},\quad 6\lambda=5\beta+\frac{5}{2},$ (B12)

(ii) $m=-2,\quad \alpha=\beta+\frac{3}{4},\quad 6\lambda=\beta+\frac{1}{4},$ (B13)

(iii) $m=-\frac{1}{2},\quad \alpha=12,\quad \beta=6,\quad \lambda=\frac{32}{3},$ (B14)

(iv) $m=-\frac{1}{3},\quad \alpha=26,\quad \beta=\frac{33}{2},\quad \lambda=48,$ (B15)

(v) $m=-\frac{2}{5},\quad \alpha=5,\quad \beta=\frac{3}{4},\quad \lambda=\frac{3}{4}.$ (B16)

Further calculation obtains that Eq. (B6) will be of the form (B1) when

(i) $m=-1,\quad \alpha=5,\quad \beta=\frac{5}{2},\quad \lambda=\frac{5}{2},$

(ii) $m=-2,\quad \alpha=\frac{3}{2},\quad \beta=\frac{3}{4},\quad \lambda=\frac{1}{6},$ (B17)

(iii) $m=-\frac{1}{2},\quad \alpha=12,\quad \beta=6,\quad \lambda=\frac{32}{3}$

The transformations defined by (B15) and (B16) do not preserve the form of Eq. (B1).

[1] J. Weiss, M. Tabor, and G. Carnevale, "The Painlevé property for partial differential equations," J. Math. Phys. **24**, 522 (1983).

[2] M. J. Ablowitz, A. Ramani, and H. Segur, "A connection between nonlinear evolution equations and ordinary differential equations of *P*-type. I," J. Math. Phys. **21**, 715 (1980).

[3]J. Weiss, "The Painlevé property for partial differential equations. II Bäcklund transformation, Lax pairs and the Schwarzian derivative," J. Math. Phys. **24**, 1405 (1983).
[4]P. J. Caudrey, R. K. Dodd, and J. D. Gibbon, "A New Heirarchy of Korteweg–de Vries Equations," Proc. Roy Soc. Lond. A **351**, 407 (1976).
[5]R. K. Dodd and J. D. Gibbon, "The Prolongation Structure of a Higher Order Korteweg–de Vries Equation," Proc. Roy. Soc. Lond. A **358**, 287 (1977).
[6]A. P. Fordy and John Gibbons, "Some Remarkable Nonlinear Transformations," Phys. Lett. A **75**, 325 (1980).
[7]Y. F. Chang, J. M. Greene, M. Tabor, and J. Weiss, "The Analytic Structure of Dynamical Systems and Self-Similar Natural Boundaries," Physica D **8**, 183 (1983).
[8]P. Lax, "Almost Periodic Solutions of the KdV Equation," SIAM Rev. **18**, 351 (1976).
[9]B. Fuchssteiner and W. Oevel, "The bi-Hamiltonian structure of some nonlinear fifth- and seventh-order differential equations and recursion formulas for their symmetries and conserved covariants," J. Math. Phys. **23**, 358 (1982).
[10]A. Fordy and J. Gibbons, "Factorization of operators. I. Miura transformations," J. Math. Phys. **21**, 2508 (1980).
[11]A. Fordy and J. Gibbons, "Factorization of operators. II," J. Math. Phys. **22**, 1170 (1981).
[12]B. Kuperschmidt and G. Wilson, "Modifying Lax Equations and the Second Hamiltonian Structure," Invent. Math. **62**, 403 (1981).
[13]M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM Stud. Appl. Math. (SIAM, Philadelphia, 1981).
[14]H. Airault, H. P. McKean, and J. Moser, "Rational and Elliptic Solutions of the Korteweg-de Vries Equation and a Related Many-Body Problem," Comm. Pure Appl. Math. **30**, 95 (1977).

Expansions over the "squared" solutions and difference evolution equations

V. S. Gerdjikov[a] and M. I. Ivanov[a]
Joint Institute for Nuclear Research, Dubna, USSR

P. P. Kulish
Leningrad Branch of the Steklov Mathematical Institute, Leningrad, USSR

(Received 3 February 1981; accepted for publication 3 December 1982)

The completeness relation for the system of "squared" solutions of the discrete analog of the Zakharov–Shabat problem is derived. It allows one to rederive the known statements concerning the class of difference evolution equations related to this linear problem and to obtain additional results. These include: (i) the expansion of the potential and its variations over the system of "squared" solutions, the expansion coefficients being the scattering data and their variations, respectively; thus the interpretation of the inverse scattering transform (IST) as a generalized Fourier transform becomes obvious; (ii) compact expressions for the trace identities through the operator Λ, for which the "squared" solutions are eigenfunctions; (iii) brief exposition of the spectral theory of the operator Λ; (iv) direct calculation of the action-angle variables based on the symplectic form of the completeness relation; (v) the generating functional of the M operators in the Lax representation; (vi) the quantum version of the IST.

PACS numbers: 02.30.Hq

I. INTRODUCTION

The intensive development of the inverse scattering transform (IST) has led to the discovery of a vast number of completely integrable Hamiltonian systems. For such physically important nonlinear evolution equations (NLEE) as the KdV, nonlinear Schrödinger, sine–Gordon equations, etc., the classes of soliton solutions, the infinite series of conserved quantities, the Bäcklund transformations, the explicit form of the action-angle variables, etc. (see Ref. 1 and the review papers, Refs. 2–4) have been constructed and investigated.

The investigation of the class of NLEE, related to the one-dimensional Zakharov–Shabat system

$$\left[i\sigma_3\frac{d}{dx}+\begin{pmatrix}0 & q(x)\\ r(x) & 0\end{pmatrix}-\lambda\right]\psi(x,\lambda)=0,$$
$$\sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \tag{1.1}$$

has revealed the importance of: (i) the expansions over the "squared" solutions of (1.1)[2,5–8] and (ii) the operator for which the "squared" solutions of (1.1) are eigenfunctions. The spectral theory of the operator Λ [7] enables one to justify the suggested in Ref. 2 interpretation of the IST as a generalized Fourier transform, linearizing the corresponding NLEE. An important property of the operator Λ consists also of the fact that it generates the hierarchy of Hamiltonian structures for the NLEE.[9]

Besides the NLEE there also exist a number of important difference evolution equations (DEE), solvable by the IST.[1,3] An example of such system is the Toda chain.[10]

The main result of the present paper consists in the derivation of the complete integrability, the construction of the hierarchy of symplectic structures and the quantization of the DEE, related to the discrete analog of the Zakharov–Shabat system[11]:

$$\psi(n+1,z)=L(n,z)\psi(n,z),\quad L(n,z)=E(z)+Q(n), \tag{1.2}$$
$$E(z)=\begin{pmatrix}z & 0\\ 0 & z^{-1}\end{pmatrix},\quad Q(n)=\begin{pmatrix}0 & q(n)\\ r(n) & 0\end{pmatrix}.$$

Our construction is based on the completeness relation for the "squared" solutions of the system (1.2).

Ablowitz and Ladik have considered in Ref. 11 the more general at first sight system (we put it in the form, proposed in Ref. 12):

$$u(n+1,\zeta)=\mathscr{L}(n,\zeta)u(n,\zeta),$$
$$\mathscr{L}(n,\zeta)=(\mu_n\nu_n)^{-1/2}\begin{pmatrix}1 & S_n\\ T_n & 1\end{pmatrix}\begin{pmatrix}\zeta & Q_n\\ R_n & \zeta^{-1}\end{pmatrix},$$
$$\mu_n=1-Q_nR_n,\quad \nu_n=1-S_nT_n. \tag{1.3}$$

The class of DEE related to (1.3) includes the discrete analogs of the nonlinear Schrödinger, KdV, sine–Gordon equations, etc. For these DEE the soliton solutions, conservation laws, the Bäcklund transformations, Hamiltonian structure, and the asymptotic of the solutions for $t\to\infty$ are known.[3,11–15]

It comes out that the systems (1.2) and (1.3) are equivalent. (The authors are grateful to I. T. Khabibulin for this remark.) Indeed, it is easy to see that if we relate the potentials and the solutions of these problems by

$$S_n=q(2n+1),\quad Q_n=q(2n),$$
$$T_n=r(2n+1),\quad R_n=r(2n), \tag{1.4}$$
$$u(n,\zeta)|_{\zeta=z^2}=\prod_{k=-\infty}^{2n-1}h(k)E^{-1/2}(z)\psi(n,z)E^{1/2}(z),$$

where $h(k)=1-q(k)r(k)$, we obtain

$$\mathscr{L}(n,\zeta)|_{\zeta=z^2}=[h(2n)h(2n+1)]^{-1/2}E^{-1/2}(z)L(2n+1,z)$$
$$\times L(2n,z)E^{1/2}(z). \tag{1.5}$$

As a result all the objects related to the system (1.2) such as

[a] On leave of absence from the Institute of Nuclear Energy and Nuclear Research, Sofia, Bulgaria.

 0022-2488/84/010025-10$02.50

DEE, conservation laws, Hamiltonian structures, etc. transfer to the corresponding objects of the system (1.3). Therefore, we confine ourselves to the system (1.2).

The present paper is a further development of our preprint.[16] We regret that when writing this preprint we were not aware of Ref. 12. We thank the referee for calling our attention to this paper.

In Sec. II we derive the completeness relation for the "squared" solutions of (1.2). Starting from it, we easily reproduce the statements from Refs. 11–14, and also obtain additional results. These include: (i) the expansion of the potential of (1.2) and its variation over the "squared" solutions, which justify the interpretation of the IST as a Fourier transform (Sec. III); (ii) compact expressions for the trace identities (Sec. III); (iii) brief exposition of the spectral theory of the operator Λ (2.21) (Sec. II); (iv) direct calculation of the action-angle variables based on the symplectic completeness relation[7,8] (Sec. IV); (v) the generating functional of the M operators in the Lax representation (Sec. III). In Sec. V it is shown that the DEE related to (1.2) with the natural reduction $r(n) = \pm q^+(n)$ may be quantized through the quantum IST.[17–19]

II. COMPLETENESS RELATION OF THE "SQUARED" SOLUTIONS

Let us start with some known facts (see Refs. 3 and 11) from the direct and inverse scattering problem for the system (1.2). In order to make the exposition simpler, we consider the case when the potential $w(n) = \binom{q(n)}{r(n)} \in \mathfrak{S}(\mathbf{Z},\mathbb{C}^2)$, the space of complex-valued vector sequences such that

$$\lim_{n\to\infty} n^k w(n) = 0 \quad \text{for all } k = 0,1,2,\cdots . \tag{2.1}$$

This together with the condition

$$0 < \prod_{k=-\infty}^{\infty} |h(k)| < \infty, \quad h(k) = 1 - q(k)r(k) \tag{2.2}$$

ensures the existence and the analyticity properties of the Jost solutions of (1.2), introduced by

$$\lim_{n\to\infty} \psi(n,z)E^{-n}(z) = \mathbf{1}, \quad \lim_{n\to\infty} \phi(n,z)E^{-n}(z) = \mathbf{1},$$
$$\psi(n,z) = \|\psi^-,\psi^+\|, \quad \phi(n,z) = \|\phi^+,\phi^-\|,$$

where $\psi^+, \phi^+, (\psi^-, \phi^-)$ are analytic for $|z| > 1$ ($|z| < 1$). The transition matrix is introduced by

$$\phi(n,z) = \psi(n,z)S(z), \quad S(z) = \begin{pmatrix} a^+, & -b^- \\ b^+, & a^- \end{pmatrix},$$
$$\det S(z) = v = \prod_{k=-\infty}^{\infty} h(k). \tag{2.3}$$

We shall denote by χ^+ (χ^-) the fundamental solutions of (1.2), analytic for $|z| > 1$ ($|z| < 1$):

$$\chi^+(n,z) = \|\phi^+,\psi^+\|, \quad \chi^-(n,z) = \|\psi^-,\phi^-\|,$$
$$\chi^+(n,z) = \psi S^- = \phi S^+, \quad \chi^-(n,z) = \psi T^+ = \phi T^-, \tag{2.4}$$
$$S^+(z) = \begin{pmatrix} 1 & b^-/v \\ 0 & a^+/v \end{pmatrix}, \quad S^-(z) = \begin{pmatrix} a^+ & 0 \\ b^+ & 1 \end{pmatrix},$$
$$T^+(z) = \begin{pmatrix} 1 & -b^- \\ 0 & a^- \end{pmatrix}, \quad T^-(z) = \begin{pmatrix} a^-/v & 0 \\ -b^+/v & 1 \end{pmatrix};$$

obviously $S^-\hat{S}^+ = T^+\hat{T}^- = S(z)$. Here and in what follows by $\hat{X}$ we shall denote the matrix inverse to X, i.e., $\hat{X} \equiv X^{-1}$. The solutions χ^+ and χ^- satisfy the following relations:

$$\chi^+(n,z)E^{-n}(z) = \chi^-(n,z)E^{-n}(z)G(n,z), \quad |z| = 1,$$
$$G(n,z) = E^n(z)\hat{T}(z)S^-(z)E^{-n}(z), \tag{2.5}$$

on the unit circle S^1. If we consider $G(z)$ as a given matrix-valued function of $z \in S^1$, then this relation may be interpreted as a noncanonical Riemann problem.[20]

The continuous spectrum of the problem (1.2) has multiplicity 2 and fills up S^1. The discrete spectrum $\Delta = \Delta^+ \cup \Delta^-$ is located at the zeroes of $a^\pm(z)$,

$$\Delta^\pm \equiv \{z_{j\pm} : a^\pm(z_{j\pm}) = a^\pm(-z_{j\pm}) = 0, \quad |z_{j\pm}| \geqslant 1,\ j = 1,\ldots,N^\pm\}. \tag{2.6}$$

Here for simplicity we asume that $n^+ = n^- = N$. The fact, that $a^\pm(z)(b^\pm(z))$ are even (odd) functions of z follows from

Remark 1: If $\psi(n,z)$ is a solution of (1.2), then $(-1)^n\sigma_3\psi(n,-z)\sigma_3$ will also be a solution of (1.2).

From the analyticity of $\chi^\pm$ it follows that $a^\pm(z)$ will also be analytic functions of z for $|z| \geqslant 1$. One is able to derive the following dispersion relation for them:

$$\ln a^+(z) = \frac{1}{4\pi i}\oint_{S^1} \frac{d\zeta^2}{\zeta^2 - z^2} \ln[1 + \rho^+\rho^-(\zeta)] + \sum_{j=1}^{N} \ln \frac{z^2 - z_{j+}^2}{z^2 - z_{j-}^2}, \quad |z| > 1,$$
$$-\ln a^-(z) = \frac{1}{4\pi i}\oint_{S^1} \frac{d\zeta^2 \cdot z^2}{\zeta^2(\zeta^2 - z^2)} \ln[1 + \rho^+\rho^-(\zeta)] + \sum_{j=1}^{N} \ln \frac{(z^2 - z_{j+}^2)z_{j-}^2}{(z^2 - z_{j-}^2)z_{j+}^2}, \quad |z| < 1, \tag{2.7}$$

where $\rho^\pm(z) = b^\pm(z)/a^\pm(z)$ are the reflection coefficients for the system (1.2).

We shall not discuss the solution of the inverse scattering problem in detail; see Refs. 3, 11, and 20. Note only that the set of independent scattering data $\mathcal{T} = \mathcal{T}^+ \cup \mathcal{T}^-$

$$\mathcal{T}^\pm \equiv \{\rho^\pm(z) = -\rho^\pm(-z), z \in S^1;$$
$$c_j^\pm, z_{j\pm}, |z_{j\pm}| \geqslant 1,\ j = 1,\ldots,N\},$$
$$\rho^\pm = b^\pm/a^\pm(z), \quad c_j^\pm = b_j^\pm/\dot{a}_j^\pm,$$
$$\dot{a}_j^\pm = \left.\frac{da^\pm}{dz}\right|_{z=z_{j\pm}}, \tag{2.8}$$
$$b_j^\pm: \ \phi^\pm(n,z_{j\pm}) = b_j^\pm\psi^\pm(n,z_{j\pm}),$$

and the dispersion relation (2.7) allow one to reconstruct uniquely the functions $a^+(z)$ ($a^-(z)$) for all z, $|z| \geqslant 1$ ($|z| \leqslant 1$), and also $b^\pm(z)$ for $|z| = 1$.

It is instructive to consider the interrelations between the potential $w(n)$ and the set of scattering data $\mathcal{T}$, (2.8), following from the formulas

$$\hat{\chi}^{\pm}(n,z)\sigma_3\chi^{\pm}(n,z)|_{n=-\infty}^{\infty}$$
$$= 2\sum_{n=-\infty}^{\infty}\hat{\chi}^{\pm}(n+1,z)\sigma_3 Q(n)\chi^{\pm}(n,z), \tag{2.9}$$
$$\hat{\chi}(n,z)\delta\chi^{\pm}(n,z)|_{n=-\infty}^{\infty}$$
$$= \sum_{n=-\infty}^{\infty}\hat{\chi}^{\pm}(n+1,z)\delta Q(n)\chi^{\pm}(n,z),$$

which are direct consequences of (1.2). The lhs of (2.9) are expressed easily through the scattering data $\mathcal{T}$, (2.8), and their variations. Inserting the first line of (2.4) into (2.9) for the matrix elements of the rhs of (2.9) one obtains expressions of the type:

$$\sum_{n=-\infty}^{\infty}\tilde{\Phi}(n,z)w(n)h^{-1}(n),$$
$$\sum_{n=-\infty}^{\infty}\tilde{\Phi}^{\pm}(n,z)\sigma_3\delta w(n)h^{-1}(n), \tag{2.10}$$

where

$$\Phi^{\pm}(n,z) = v(n)\phi^{\pm}(n,z)\circ\phi^{\pm}(n+1,z), \quad \tilde{\Phi} = (\Phi_2, -\Phi_1), \tag{2.11}$$
$$\phi(n,z)\circ\psi(m,z) \overset{\text{def}}{=} \begin{pmatrix}\phi_1(n,z)\psi_1(m,z)\\ \phi_2(n,z)\psi_2(m,z)\end{pmatrix}, \quad v(n) = \prod_{k=n}^{\infty}h(k).$$

If we introduce in the space $\mathfrak{S}(\mathbf{Z},\mathbf{C}^2)$ the skew-scalar product, $X,Y\in\mathfrak{S}(\mathbf{Z},\mathbf{C}^2)$:

$$[X,Y] = \sum_{n=-\infty}^{\infty}\tilde{X}(n)Y(n)$$
$$= \sum_{n=-\infty}^{\infty}[X_2(n)Y_1(n) - X_1(n)Y_2(n)], \tag{2.12}$$

then the matrix elements of the rhs of (2.9) can be interpreted as expansion coefficients of $w(n)$ and $\sigma_3\delta w(n)$ over the "squared" solutions $\Phi^{\pm}(n,z)$ of (1.2), i.e., the terms (2.10) will have the form $[\Phi^{\pm}(n),w(n)h^{-1}(n)]$, $[\Phi^{\pm}(n),\sigma_3\delta w(n)h^{-1}(n)]$.

Let us introduce the system $\{\Phi\}$, $\{\Psi\}$ of "squared" solutions of (1.2) by

$$\{\Phi\}\equiv\{\Phi^{\pm}(n,z), z\in S^1; \Phi_j^{\pm}(n), \dot{\Phi}_j^{\pm}(n), j=1,\ldots,N\},$$
$$\{\Psi\}\equiv\{\Psi^{\pm}(n,z), z\in S^1; \Psi_j^{\pm}(n), \dot{\Psi}_j^{\pm}(n), j=1,\ldots,N\}, \tag{2.13}$$
$$\Psi^{\pm}(n,z) = v(n)\psi^{\pm}(n,z)\circ\psi^{\pm}(n+1,z),$$
$$\Psi_j^{\pm}(n) = \Psi^{\pm}(n,z_{j\pm}),$$
$$\dot{\Psi}_j^{\pm}(n) = \lim_{z\to z_{j\pm}}\frac{d}{dz}\Psi^{\pm}(n,z);$$

$\Phi_j^{\pm}(n)$ and $\dot{\Phi}_j^{\pm}(n)$ are obtained analogously from the definition of $\Phi^{\pm}(n,z)$ in (2.11). The completeness of the systems $\{\Psi\}$, $\{\Phi\}$ is proved by introducing the Green function $G = G^{\pm}(n,m,z)$, $|z|\gtrless 1$,

$$G^{\pm}(n,m,z) = \{2/[a^{\pm}(z)]^2\}\{\Psi^{\pm}(n,z)\tilde{\Phi}(m,z)\theta(n-m)$$
$$+\theta(m-n)[2(\phi^{\pm}(n,z)\circ\psi^{\pm}(n+1,z))$$
$$\times(\phi^{\pm}(m+1,z)\circ\psi^{\pm}(m,z))\tilde{}\,v(n)v(m)$$
$$-\Phi^{\pm}(n,z)\tilde{\Psi}^{\pm}(m,z)]\},$$
$$\theta(n-m) = \begin{cases}1, & n>m,\\ \frac{1}{2}, & n=m,\\ 0, & n<m,\end{cases} \tag{2.14}$$

and applying the contour integration method to the integral,

$$\frac{1}{2\pi i}\oint_{\gamma_+}\frac{dz}{z}G^{+}(n,m,z) - \frac{1}{2\pi i}\oint_{\gamma_-}\frac{dz}{z}G^{-}(n,m,z).$$

Here the contours $\gamma_+ = S^1\cup\bar{S}^{\infty}$, $\gamma_- = S^1\cup\bar{S}^{0}$, where S^1 is the positively oriented unit circle and $\bar{S}^{\infty}$ and $\bar{S}^{0}$ the negatively oriented circles with infinitely large and infinitely small radii resp. The result is

$$h(n)\delta(n-m) = \frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}\left[\frac{\Psi^{+}(n,z)\tilde{\Phi}^{+}(m,z)}{[a^{+}(z)]^2} - \frac{\Psi^{-}(n,z)\tilde{\Phi}^{-}(m,z)}{[a^{-}(z)]^2}\right]$$
$$-2\sum_{j=1}^{N}[X_j^{+}(n,m)+X_j^{-}(n,m)], \tag{2.15}$$
$$X_j^{\pm}(n,m) = \frac{1}{z_{j\pm}(a_j^{\pm})^2}[\Psi_j^{\pm}(n)\dot{\tilde{\Phi}}_j^{\pm}(m) + \dot{\Psi}_j^{\pm}(n)\tilde{\Phi}_j^{\pm}(m)]$$
$$-\frac{\dot{a}_j^{\pm}+z_{j\pm}\ddot{a}_j^{\pm}}{z_{j\pm}^2(\dot{a}_j^{\pm})^3}\Psi_j^{\pm}(n)\tilde{\Phi}_j^{\pm}(m).$$

This completeness relation may be rewritten in the so-called symplectic form:

$$h(n)\delta(n-m) = \oint_{S^1}\frac{dz}{z}[P(n,z)\tilde{Q}(m,z) - Q(n,z)\tilde{P}(m,z)]$$
$$+2\sum_{j=1}^{N}[P_j^{+}(n)\tilde{Q}_j^{+}(m) - Q_j^{+}(n)\tilde{P}_j^{+}(m)$$
$$+P_j^{-}(n)\tilde{Q}_j^{-}(m) - Q_j^{-}(n)\tilde{P}_j^{-}(m)], \tag{2.16}$$

where

$$P(n,z) = -(1/2\pi)(\rho^{+}\Psi^{+}+\rho^{-}\Psi^{-})(n,z)$$
$$= -(1/2\pi v)(\sigma^{+}\Phi^{+}+\sigma^{-}\Phi^{-})(n,z),$$
$$Q(n,z) = (iv/b^{+}b^{-})\left(\rho^{+}\Psi^{+} - \frac{\sigma^{+}}{v}\Phi^{+}\right)(n,z)$$
$$= (iv/2b^{+}b^{-})\left(\frac{\sigma^{-}}{v}\Phi^{-} - \rho^{-}\Psi^{-}\right)(n,z), \tag{2.17}$$
$$P_j^{\pm}(n) = \mp(ic_j^{\pm}/z_{j\pm})\Psi_j^{\pm}(n),$$
$$Q_j^{\pm}(n) = \mp\tfrac{1}{2}i[m_j^{\pm}\dot{\Phi}_j^{\pm}(n) - c_j^{\pm}\dot{\Psi}_j^{\pm}(n)],$$
$$\sigma^{\pm}(z) = b^{\mp}(z)/a^{\pm}(z), \quad m_j^{\pm} = (b_j^{\pm}a_j^{\pm})^{-1}.$$

The two systems $\{\Psi\}$ and $\{\Phi\}$ are biorthogonal with respect to the skew-scalar product (2.12). Indeed, using (1.2), one can verify the following biquadratic relations between any two solutions $\phi(n,z)$ and $\psi(n,\zeta)$ of (1.2):

$$[\Phi(n,z),\Psi(n,\zeta)h^{-1}(n)]$$
$$= \frac{\zeta}{z}\cdot\frac{v^2(n)}{\zeta^2-z^2}$$
$$\times[z\phi_2(n,z)\psi_1(n,\zeta) - \zeta\phi_1(n,z)\psi_2(n,\zeta)]^2|_{n=-\infty}^{\infty}. \tag{2.18}$$

Making use of (2.4) and of the fact that

$$\text{P.v.}\lim_{n\to\infty}\frac{(z/\zeta)^n}{\zeta - z} = \pi\delta(\arg z - \arg\zeta), \quad z,\zeta\in S^1,$$

we obtain

$$[\Phi^{\pm}(n,\zeta),\Psi^{\pm}(n,\zeta)h^{-1}(n)] = \mp 2\pi[a^{\pm}(\zeta)]^2\delta(\arg z - \arg\zeta), \quad z,\zeta\in S^1,$$
$$[\Phi_j^{\pm}(n),\Psi_k^{\pm}(n)h^{-1}(n)] = 0,$$
$$[\Phi_j^{\pm}(n),\dot{\Psi}_k^{\pm}(n)] = -\tfrac{1}{2}(\dot{a}_j^{\pm})^2 z_{j\pm}\delta_{jk}, \tag{2.19}$$
$$[\dot{\Phi}_j^{\pm}(n),\Psi_k^{\pm}(n)h^{-1}(n)] = -\tfrac{1}{2}(\dot{a}_j^{\pm})^2 z_{j\pm}\delta_{jk},$$
$$[\dot{\Phi}_j^{\pm}(n),\dot{\Psi}_k^{\pm}(n)h^{-1}(n)] = -\tfrac{1}{2}(\ddot{a}_j^{\pm} z_{j\pm} + \dot{a}_j^{\pm})\dot{a}_j^{\pm}\delta_{jk}.$$

From (2.18) and (1.29) we also have

$$[Q(n,z),P(n,\zeta)h^{-1}(n)] = -i\delta(\arg z - \arg\zeta), \quad z,\zeta\in S^1,$$
$$[Q_j^{\pm}(n),P_k^{\pm}(n)h^{-1}(n)] = \tfrac{1}{2}\delta_{jk}, \tag{2.20}$$
$$[Q_j^{\pm}(n),P_k^{\mp}(n)h^{-1}(n)] = 0.$$

Relations (2.19) and (2.20) allow one to conclude that the systems $\{\Psi\}$, $\{\Phi\}$, and $\{P,Q\}$ consist of a linearly independent element.

Now it is natural to introduce the operators $\Lambda_{\pm}$, for which the elements of $\{\Psi\}$ and $\{\Phi\}$ are eigenfunctions, i.e.,

$$(\Lambda_+ - z^2)\Psi^{\pm}(n,z) = 0, \quad (\Lambda_- - z^2)\Phi^{\pm}(n,z) = 0, \quad z\in S^1\cup\Delta,$$
$$(\Lambda_+ - z_{j\pm}^2)\dot{\Psi}_j^{\pm}(n) = 2z_{j\pm}\Psi_j^{\pm}(n), \tag{2.21}$$
$$(\Lambda_- - z_{j\pm}^2)\dot{\Phi}_j^{\pm}(n) = 2z_{j\pm}\Phi_j^{\pm}(n).$$

The explicit form of $\Lambda_{\pm}$ has been known.[11,12,14] For us it will be convenient to factorize them in the form

$$\Lambda_{\pm}X(n) = \Lambda_2^{\pm}\Lambda_1^{\pm}X(n), \quad X(n)\in\mathfrak{S}(\mathbb{Z},\mathbb{C}^2), \tag{2.22}$$

where the operators $\Lambda_i^{\pm}$, $i = 1,2$, are defined by

$$\Lambda_1^{+}\Psi^{\pm}(n,z) = z\overline{\Psi}^{\pm}(n,z), \quad \Lambda_1^{-}\Phi^{\pm}(n,z) = z\overline{\Phi}^{\pm}(n,z),$$
$$z = S^1\cup\Delta, \tag{2.23}$$
$$\Lambda_2^{+}\overline{\Psi}^{\pm}(n,z)z\Psi^{\pm}(n,z), \quad \Lambda_2^{-}\overline{\Phi}^{\pm}(n,z) = z\Phi^{\pm}(n,z),$$
$$\overline{\Psi}^{\pm}(n,z) = v(n)\psi^{\pm}(n,z)\circ\psi^{\pm}(n,z),$$
$$\overline{\Phi}^{\pm}(n,z) = v(n)\phi^{\pm}(n,z)\circ\phi^{\pm}(n,z).$$

The explicit form of $\Lambda_i^{\pm}$, $i = 1,2$ and their inverse is given by

$$\Lambda_1^{\pm}X(n) = \begin{pmatrix} X_1(n) \\ X_2(n-1)\end{pmatrix} \pm \begin{pmatrix} q(n) \\ -r(n-1)\end{pmatrix} \times\sum_{n}^{\pm}[r(k)X_1(k) + q(k)X_2(k)]h^{-1}(k),$$

$$\Lambda_2^{\pm}X(n) = h(n)\begin{pmatrix} X_1(n+1) \\ X_2(n)\end{pmatrix} \pm \begin{pmatrix} q(n) \\ -r(n)\end{pmatrix} \times\sum_{n+1}^{\pm}[r(k-1)X_1(k) + q(k)X_2(k)],$$

$$\hat{\Lambda}_1^{\pm}X(n) = h(n)\begin{pmatrix} X_1(n) \\ X_2(n+1)\end{pmatrix} \mp \begin{pmatrix} q(n) \\ -r(n)\end{pmatrix} \times\sum_{n+1}^{\pm}[r(k)X_1(k) + q(k-1)X_2(k)],$$

$$\hat{\Lambda}_2^{\pm}X(n) = \begin{pmatrix} X_1(n-1) \\ X_2(n)\end{pmatrix} \mp \begin{pmatrix} q(n-1) \\ -r(n)\end{pmatrix} \times\sum_{n}^{\pm}[r(k)X_1(k) + q(k)X_2(k)]h^{-1}(k), \tag{2.24}$$

where

$$\sum_n^{+} \equiv \sum_{k=n}^{\infty}, \quad \sum_n^{-} \equiv \sum_{k=-\infty}^{n-1}.$$

The condition (2.1) ensures that $\Lambda_i^{\pm}X\in\mathfrak{S}(\mathbb{Z},\mathbb{C}^2)$ for any $X\in\mathfrak{S}(\mathbb{Z},\mathbb{C}^2)$.

The operators $\Lambda_i^{\pm}$, $i = 1,2$, and $\Lambda_{\pm}$ satisfy conjugationlike relations with respect to the skew-scalar product (2.12):

$$[Y(n),\Lambda_1^{+}X(n)h(n)] = [\Lambda_2^{-}Y(n),X(n)],$$
$$[Y(n),\Lambda_2^{+}X(n)] = [\Lambda_1^{-}h(n)Y(n),X(n)], \tag{2.25}$$
$$[Y(n),\Lambda_{+}h(n)X(n)] = [\Lambda_{-}h(n)Y(n),X(n)].$$

The first two lines of (2.25) follow directly from the explicit form of $\Lambda_i^{\pm}$, (2.24), and from the definition of $[\ ,\ ]$, (2.12); the third line is a consequence of the first two and (2.22).

The spectral theory of the operators $\Lambda_{\pm}$ can be constructed analogously to Refs. 7 and 21. Here we will only show the interrelation between the Green function (2.14) and the operator Λ_+. Applying the contour integration method to the integral

$$\frac{1}{2\pi i}\oint_{\gamma_+}\frac{d\zeta}{\zeta}\frac{\zeta^2+z^2}{\zeta^2-z^2}G^{+}(n,m,\zeta) - \frac{1}{2\pi i}\oint_{\gamma_-}\frac{d\zeta}{\zeta}\frac{\zeta^2+z^2}{\zeta^2-z^2}G^{-}(n,m,\zeta),$$

we obtain the following spectral decomposition for G:

$$G(n,m,z) = \frac{i}{2\pi}\oint_{S^1}\frac{d\zeta}{\zeta}\frac{\zeta^2+z^2}{\zeta^2-z^2}\left\{\frac{\Psi^{+}(n,\zeta)\widetilde{\Phi}^{+}(m,\zeta)}{[a^{+}(\zeta)]^2} - \frac{\Psi^{-}(n,\zeta)\widetilde{\Phi}^{-}(m,\zeta)}{[a^{-}(\zeta)]^2}\right\} - 2\sum_{j=1}^{N}[Y_j^{+}(n,m) + Y_j^{-}(n,m)],$$
$$Y_j^{\pm}(n,m) = \lim_{\zeta\to z_{j\pm}}\frac{d}{d\zeta}\left[\frac{(\zeta - z_{j\pm})^2(\zeta^2+z^2)}{\zeta(\zeta^2-z^2)[a^{\pm}(\zeta)]^2}\Psi^{\pm}(n,\zeta)\widetilde{\Phi}^{\pm}(m,\zeta)\right]. \tag{2.26}$$

From (2.26), (2.21), and (2.15) it follows that

$$(\Lambda_+ + z^2)^{-1}(\Lambda_+ - z^2)G(n,m,z)h^{-1}(m) = \delta(n-m),$$

i.e., $G(n,m,z)$ is the Green function for the operator $(\Lambda_+ + z^2)^{-1}(\Lambda_+ - z^2)$. This result is esentially different from the one related to the Zakharov–Shabat system (1.1); there the continuous analogs of G and Λ_+ are related by $(\Lambda_+ - \lambda)G(x,y,\lambda) = \delta(x-y)$.

III. THE IST AS A FOURIER TRANSFORM

Let us start by deriving the expansions for $w(n)$ and $\sigma_3\delta w(n)$ over the systems of "squared" solutions $\{\Psi\}$ and $\{\Phi\}$. To do this, we multiply the completeness relations (2.15) and (2.16) by $w(m)h^{-1}(m)$ and $\sigma_3\delta w(m)h^{-1}(m)$ from the right and sum over m. Thus the corresponding expansion coefficients have the form (2.10) and through (2.9) are easily expressed in terms of the scattering data $\mathcal{T}$. The result is

$$w(n) = \frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}(\rho^+\Psi^+ + \rho^-\Psi^-)(n,z) - 2\sum_{j=1}^{N}\left[\frac{c_j^+}{z_{j+}}\Psi_j^+(n) - \frac{c_j^-}{z_{j-}}\Psi_j^-(n)\right], \quad (3.1a)$$

$$w(n) = -i\oint_{S^1}\frac{dz}{z}P(n,z) - 2i\sum_{j=1}^{N}[P_j^+(n) + P_j^-(n)] \quad (3.1b)$$

and

$$\sigma_3\delta w(n) = -\frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}(\delta\rho^+\Psi^+ - \delta\rho^-\Psi^-)(n,z) + 2\sum_{j=1}^{N}[Y_j^+(n) + Y_j^-(n)], \quad (3.2a)$$

$$Y_j^\pm(n) = \delta\left(\frac{c_j^\pm}{z_{j\pm}}\right)\Psi_j^\pm(n) + c_j^\pm\delta\ln z_{j\pm}\,\dot\Psi_j^\pm(n),$$

$$\sigma_3\delta w(n) = \oint_{S^1}\frac{dz}{z}[Q(n,z)\,\delta\hat p(z) - P(n,z)\,\delta\hat q(z)] + 2\sum_{j=1}^{N}[Z_j^+(n) + Z_j^-(n)], \quad (3.2b)$$

$$Z_j^\pm(n) = Q_j^\pm(n)\,\delta\hat p_j^\pm - P_j^\pm(n)\,\delta\hat q_j^\pm,$$

where

$$\hat p(z) = -(1/2\pi)\ln[1+\rho^+\rho^-(z)],$$
$$\hat q(z) = -\tfrac{1}{2}i\ln[b^+(z)/b^-(z)], \quad z\in S^1,$$
$$\hat p_j^\pm = \mp i\ln z_{j\pm}, \quad \hat q_j^\pm = \mp i\ln(b_j^\pm/\sqrt{v}), \quad (3.3)$$
$$\delta\hat p(z) = -[P(n,z),\sigma_3\delta w(n)h^{-1}(n)],$$
$$\delta\hat q(z) = -[Q(n,z),\sigma_3\delta w(n)h^{-1}(n)].$$

Now the parallel between the IST and the Fourier transform is obvious: The expansion coefficients in (3.1) and (3.2) are simply the scattering data $\mathcal{T}$, (2.8), and their variations. As a generalization of the usual "discrete exponent" z^n, one should consider $\{\Psi\}$ or $\{P,Q\}$; the role of the shift operator will be played by the operator Λ_+ (2.21).

From (3.1) and (3.2) there follows a more rigorous proof of the theorem, concerning the description of the DEE related to (1.2).

Theorem 1: Let $f(z^2)$ be a meromorphic function with poles lying outside of a certain neighborhood of the spectrum $S^1\cup\Delta$ of (1.2). Then $w(n,t)$ satisfies the DEE

$$\sigma_3\frac{dw}{dt} + f(\Lambda_+)w(n,t) = 0 \quad (3.4)$$

if and only if the scattering data $\mathcal{T}$, (2.8), satisfy the linear equations:

$$\frac{d\rho^\pm}{dt} \mp f(z^2)\rho^\pm(z,t) = 0,$$
$$\frac{dc_j^\pm}{dt} \mp f(z_{j\pm}^2)c_j^\pm(t) = 0, \quad (3.5)$$
$$\frac{dz_{j\pm}}{dt} = 0.$$

Proof: Let us insert the expansion of $w(n)$, (3.1a), and $\sigma_3(dw/dt)$ over the system $\{\Psi\}$ in the lhs of (3.4). The latter is obtained from (3.2a) by considering variations of the form $\sigma_3\delta w(n) = \sigma_3(dw/dt)\delta t + O((\delta t)^2)$, and differs from (3.2a) only in that the coefficients $\delta\rho^\pm,\cdots$ are replaced by $d\rho^\pm/dt,\cdots$; the same is true also for (3.2b). This gives

$$\sigma_3\frac{dw}{dt} + f(\Lambda_+)w(n,t) = \frac{1}{2\pi i}\oint_{S^1}\frac{dz}{z}\{[\rho_t^+ - f(z^2)\rho^+]\Psi^+(n,z) - (\rho_t^- + f(z^2)\rho^-)\Psi^-(n,z)\} + 2\sum_{j=1}^{N}[U_j^+(n) + U_j^-(n)], \quad (3.6)$$

$$U_j^\pm(n) = \left[\frac{d}{dt}\left(\frac{c_j^\pm}{z_{j\pm}}\right) - f(z_{j\pm}^2)\frac{c_j^\pm}{z_{j\pm}}\right]\Psi_j^\pm(n) + \frac{c_j^\pm}{z_{j\pm}}\frac{dz_{j\pm}}{dt}\dot\Psi_j^\pm(n).$$

In obtaining the rhs of (3.6) we have made use of (2.21). It remains to be noted that the lhs of (3.6) vanishes if and only if all the expansion coefficients on the rhs of (3.6) vanish, which readily gives (3.5). This last step follows also from the fact that the systems $\{\Psi\}$ and $\{\Phi\}$ are biorthogonal [see (2.19)].

Analogously, using the symplectic expansion (2.16), we can prove

Theorem 2: $w(n,t)$ satisfies (3.4) if and only if the set $\{\hat p,\hat q\}$ in (3.3) satisfies the linear equations:

$$\frac{d\hat p(z,t)}{dt} = 0, \quad i\frac{d\hat q(z,t)}{dt} = f(z^2),$$
$$\frac{d\hat p_j^\pm(t)}{dt} = 0, \quad i\frac{d\hat q_j^\pm(t)}{dt} = f(z_{j\pm}^2). \quad (3.7)$$

From (3.5) and (3.7) it follows that the DEE (3.4) has an infinite series of conserved quantities $C^{(p)}, p = 0, \pm 1,\cdots$. As a generating functional of $C^{(p)}$ it is natural to consider $\mathcal{A}(z)$:

$$\mathscr{A}(z) = \ln a^+(z), \quad |z| > 1,$$
$$\mathscr{A}(z) = -\ln a^-(z), \quad |z| < 1, \tag{3.8}$$

$C^{(p)}$ being the expansion coefficients of $\mathscr{A}(z)$:

$$\mathscr{A}(z) = \sum_{p=1}^{\infty} C^{(p)} z^{-2p}, \quad |z| \gg 1,$$
$$\mathscr{A}(z) = -\sum_{p=0}^{\infty} C^{(p)} z^{2p}, \quad |z| \ll 1. \tag{3.9}$$

To derive compact expressions for $C^{(p)}$ as functionals of $w(n)$, we start with the relation

$$z\frac{d\mathscr{A}}{dz} = \tfrac{1}{2}\operatorname{tr}\{[z\hat{\chi}^+(n,z)\dot{\chi}^+(n,z) - \dot{n}\sigma_3](\mathbf{1}+\sigma_3)\}|_{n=-\infty}^{\infty}, \quad |z| > 1, \tag{3.10}$$

which follows from (2.4), (3.8), and (1.2). Using (1.2) once more for the rhs of (3.10), we obtain

$$\sum_{n=-\infty}^{\infty} \{\tfrac{1}{2}\operatorname{tr}[\hat{\chi}^+(n,z)\sigma_3\chi^+(n,z)(\mathbf{1}+\sigma_3)] - 1\} = \sum_{n=-\infty}^{\infty} \sum\nolimits_n^+ \operatorname{tr}[\hat{\chi}^+(k+1,z)Q(k)\sigma_3\chi^+(k,z)\sigma_3], \tag{3.11}$$

which can be put into the form

$$z\frac{d\mathscr{A}}{dz} = -\sum_{n=-\infty}^{\infty} \sum\nolimits_n^+ v(k+1)\tilde{w}(k) \times \frac{\psi^+(k+1)\circ\phi^+(k) + \psi^+(k)\circ\phi^+(k+1)}{a^+(z)}. \tag{3.12}$$

Let us now expand $[v(k)/a^+(z)][\psi^+(k+1)\circ\phi^+(k) + \psi^+(k)\circ\phi^+(k+1)]$ over the system $\{\Psi\}$. The corresponding expansion coefficients are expressed through the scattering data $\mathscr{T}$ by using (2.18). Thus we obtain

$$\frac{v(n)}{a^+(z)}[\psi^+(n+1,z)\circ\phi^+(n,z) + \psi^+(n,z)\circ\phi^+(n+1,z)] = \frac{i}{2\pi}\oint_{S^1}\frac{d\zeta}{\zeta}\frac{\zeta^2+z^2}{\zeta^2-z^2}[\rho^+\Psi^+ + \rho^-\Psi^-](n,z) - 2\sum_{j=1}^{N}\left[\frac{c_j^+}{z_{j+}}\cdot\frac{z_{j+}^2+z^2}{z_{j+}^2-z^2}\Psi_j^+(n) - \frac{c_j^-}{z_{j-}}\cdot\frac{z_{j-}^2+z^2}{z_{j-}^2-z^2}\Psi_j^-(n)\right]$$
$$= (\Lambda_+ + z^2)(\Lambda_+ - z^2)^{-1}w(n). \tag{3.13}$$

Inserting the rhs of (3.13) into (3.12), we arrive at

$$z\frac{d\mathscr{A}}{dz} = -\sum_{n=-\infty}^{\infty}\sum\nolimits_n^+\frac{\tilde{w}(k)}{h(k)}(\Lambda_+ + z^2)(\Lambda_+ - z^2)^{-1}w(k), \tag{3.14}$$

which proves to be valid both for $|z| > 1$ and $|z| < 1$ (the considerations for $|z| < 1$ are analogous). Comparing (3.14) and (3.9) for $C^{(p)}$, we obtain

$$C^{(p)} = \frac{1}{p}\sum_{n=-\infty}^{\infty}\sum\nolimits_n^+\frac{\tilde{w}(k)\Lambda_+^p w(k)}{h(k)}, \quad p = \pm 1, \pm 2, \cdots. \tag{3.15}$$

The dispersion relations (2.7) allow one to express $C^{(p)}$ as functionals of the scattering data $\mathscr{T}$:

$$C^{(p)} = \frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}z^{2p}\ln[1+\rho^+\rho^-(z)] - \frac{1}{p}\sum_{j=1}^{N}(z_{j+}^{2p} - z_{j-}^{2p}), \quad p \neq 0, \tag{3.16}$$
$$C^{(0)} = -\ln v = \frac{i}{2\pi}\oint_{S^1}\frac{dz}{z}\ln[1+\rho^+\rho^-(z)] - \sum_{j=1}^{N}\ln\frac{z_{j+}^2}{z_{j-}^2}.$$

The desired trace identities are obtained after equating the rhs of (3.15) and (3.16). Through the same pattern one can derive compact formulas for the variations of $\delta C^{(p)}$.[12] For this it is enough to note that

$$\delta\mathscr{A}(z) = \tfrac{1}{2}tr[\hat{\chi}^+(n,z)\delta\hat{\chi}(n,z)(\mathbf{1}+\sigma_3)]|_{n=-\infty}^{\infty} = \frac{1}{2}\sum_{n=-\infty}^{\infty}[\sigma_3\delta w(n)]\tilde{}\,\frac{v(n+1)}{a^+(z)} \times [\phi^+(n,z)\circ\psi^+(n+1,z) + \phi^+(n+1,z)\circ\psi^+(n,z)], \quad |z| > 1,$$

which with (3.13) directly leads to

$$\delta C^{(p)} = [\sigma_3\delta w(n)h^{-1}(n), \Lambda_+^p w(n)]. \tag{3.17}$$

We end this paragraph by reproducing in compact form the formulas from the traditional approach[2,3] to the DEE (3.4) as a consistency condition,

$$\frac{dL(n,z)}{dt} + L(n,z)M(n,z) - M(n+1,z)L(n,z) = 0, \tag{3.18}$$

of two linear problems: (1.2) and

$$\frac{d\psi(n,z)}{dt} = M(n,z)\psi(n,z). \tag{3.19}$$

Choosing $M(n,z) = \Sigma_k z^k M^{(k)}(n)$ as polynomial of z and z^{-1} and inserting in (3.18), one obtains recurrent relations for the coefficients $M^{(k)}(n)$ [11]; trying to solve them, one, after somewhat tedious calculations, naturally obtains the Λ operator.

Here we shall use another approach, developed for the NLEE by Gel'fand and Dickey[22]; see also Ref. 21. Let us introduce the resolvent of the system (1.2)[23]

$$\mathscr{R}(n,m,z)=\mathscr{R}^{\pm}(n,m,z),\quad |z|\gtrless 1,$$
$$\mathscr{R}^{\pm}(n,m,z)=\chi^{\pm}(n,z)\Theta(n-m)\hat{\chi}^{\pm}(m+1,z), \tag{3.20}$$
$$\Theta^{+}(n-m)=\begin{cases}\mathrm{diag}(-1,0), & m>n,\\ \mathrm{diag}(0,1), & m<n,\end{cases}$$
$$\Theta^{-}(n-m)=\begin{cases}\mathrm{diag}(0,-1), & m>n,\\ \mathrm{diag}(1,0), & m<n,\end{cases}$$

and define its "diagonal" as
$R(n,z)=\mathscr{R}(n,n-1,z)-\frac{1}{2}=-\frac{1}{2}\chi^{\pm}(n,z)\sigma_3\hat{\chi}^{\pm}(n,z)$. It is easy to verify that $R(n,z)$ satisfies the equation

$$L(n,z)R(n,z)-R(n+1,z)L(n,z)=0. \tag{3.21}$$

Since $R^{\pm}(n,z)$ is analytic in z for $|z|\gtrless 1$, one may consider the asymptotic expansions

$$R^{+}(n,z)=-\tfrac{1}{2}\sigma_3+\sum_{p=1}^{\infty}R^{(p)}(n)z^{-p},\quad |z|\gtrgtr 1,$$
$$R^{-}(n,z)=\tfrac{1}{2}\sigma_3+\sum_{p=1}^{\infty}R^{(-p)}(n)z^{p},\quad |z|\lessgtr 1, \tag{3.22}$$

Note that $R^{(p)}(n)$ and $M^{(p)}(n)$, $p\neq 0$, satisfy the same recurrent relations. From the definition of $R(n)$ and (2.23) and (2.24) we have

$$\sigma_3\begin{pmatrix}R_{12}^{\pm}(n)\\ R_{21}^{\pm}(n)\end{pmatrix}=\pm\frac{v(n)\phi^{\pm}(n,z)\circ\psi^{\pm}(n,z)}{a^{\pm}(z)}$$
$$=\tfrac{1}{2}z\hat{\Lambda}_{2}^{+}\frac{v(n)}{a^{\pm}(z)}[\phi^{\pm}(n+1,z)\circ\psi^{\pm}(n,z)$$
$$+\phi^{\pm}(n,z)\circ\psi^{\pm}(n+1,z)+w(n)], \tag{3.23}$$
$$R_{11}^{\pm}(n)=-R_{22}^{\pm}(n)$$
$$=\mp\frac{1}{2}\frac{v(n)}{a^{\pm}(z)}[\phi_1^{\pm}(n,z)\psi_2^{\pm}(n,z)$$
$$+\phi_2^{\pm}(n,z)\psi_1^{\pm}(n,z)].$$

Making use of (3.12) and (3.13), one obtains compact expressions for $R^{(p)}(n,z)$ through the operator Λ_+:

$$R^{(2p)}(n)-\sigma_3\sum_{n}^{+}\frac{\tilde{w}(k)}{h(k)}\Lambda_{+}^{p}w(k),\quad p=\pm 1,\pm 2,\cdots,$$
$$R^{(2p-1)}(n)=\begin{pmatrix}0, & R_{12}^{(2p-1)}(n)\\ R_{21}^{(2p-1)}(n), & 0\end{pmatrix}, \tag{3.24}$$
$$\begin{pmatrix}R_{12}^{(2p-1)}(n)\\ R_{21}^{(2p-1)}(n)\end{pmatrix}=-\sigma_3\hat{\Lambda}_{\epsilon}^{+}\Lambda_{+}^{p}w(n),\quad \epsilon=\begin{cases}2, & p>0,\\ 1, & p<0,\end{cases}$$

The M operators for the DEE (3.4) are simple linear combinations of $R^{(p)}(n)$. We write down the M operator only for the simplest case, when in (3.4) $f(z^2)=\nu z^{2N}$, $N>0$, $\nu=\mathrm{const}$:

$$M^{(N)}(n,z)=\nu\left[\sum_{p=0}^{2N-1}z^{2N-p}R^{(p)}(n)+\tfrac{1}{2}R^{(2N)}(1+\sigma_3)\right]. \tag{3.25}$$

Thus $R(n,z)$ may be considered as a generating functional of the M operators and also of the conserved quantities of the DEE (3.4). The last statement is obtained by comparing (3.24) and (3.11), which gives

$$z\frac{d\mathscr{A}}{dz}=\mp\sum_{n=-\infty}^{\infty}[\mathrm{tr}(R(n,z)\sigma_3)\pm 1],\quad |z|\gtrless 1. \tag{3.26}$$

IV. HIERARCHIES OF HAMILTONIAN STRUCTURES

The proof of the Hamiltonian structure of the DEE (3.4) is now easy. For this one should introduce the following symplectic form[12]:

$$\Omega^{(0)}=2i\sum_{n=-\infty}^{\infty}\frac{\delta q(n)\wedge\delta r(n)}{h(n)}$$
$$\equiv i[\sigma_3\delta w(n),\ \text{exterior product},\ \sigma_3\delta w(n)h^{-1}(n)], \tag{4.1}$$

where $\delta q\wedge\delta r=\delta_1 q\delta_2 r-\delta_2 q\delta_1 r$ is the usual exterior product. In order that the Hamiltonian equations of motion

$$\Omega^{(0)}\left(\sigma_3\frac{dw}{dt},\cdot\right)=\delta H_f(\cdot) \tag{4.2}$$

coincide with (3.4), one should choose H_f in the form

$$H_f=-i\sum_p f_p C^{(p)}$$
$$=-if_0\sum_{n=-\infty}^{\infty}\ln h(n)$$
$$-i\sum_{n=-\infty}^{\infty}\sum_{n}^{+}\tilde{w}(n)h^{-1}(n)F(\Lambda_+)w(n), \tag{4.3}$$

where

$$f(z^2)=\sum_p f_p z^{2p},\quad F(z^2)=\int^{z^2}\frac{ds}{s}[f(s)-f_0]. \tag{4.4}$$

The complete integrability of the DEE (3.4) becomes obvious after recalculating $\Omega^{(0)}$ and H_f in terms of the scattering data variations. Most simply $\Omega^{(0)}$ is calculated by inserting the symplectic expansion (2.16) into (4.1) and using the third line in (3.3). This immediately casts $\Omega^{(0)}$ in canonical form:

$$\Omega^{(0)}=2i\oint_{S^1}\frac{dz}{z}\delta\hat{p}(z)\wedge\delta\hat{q}(z)$$
$$+4i\sum_{j=1}^{N}[\delta\hat{p}_j^{+}\wedge\delta\hat{q}_j^{+}+\delta\hat{p}_j^{-}\wedge\delta\hat{q}_j^{-}], \tag{4.5}$$

which means that $\{\hat{p},\hat{q}\}$ is a set of canonical coordinates and momenta. From (3.16) and (4.3) we see that

$$H_f=-\oint_{S^1}\frac{dz}{z}f(z^2)\hat{p}(z)+i\sum_{j=1}^{N}[F_1(z_{j+}^2)-F_1(z_{j-}^2)],$$
$$F_1(s)=\int^{s}\frac{ds'}{s'}f(s'),\quad z_{j\pm}^2=\exp(\pm 2i\hat{p}_j^{\pm}), \tag{4.6}$$

i.e., H_f depends only on the set of the new momenta $\{\hat{p}\}$. Thus $\{\hat{p},\hat{q}\}$ in (3.3) is the set of the action-angle variables for the DEE (3.4).[12]

The symplectic structure $\Omega^{(0)}$ is not unique. One can introduce a one-parameter family of symplectic forms $\Omega^{(m)}$, generated from $\Omega^{(0)}$ (4.1) by the operator Λ_+:

$$\Omega^{(m)}=i[\sigma_3\delta w(n)h^{-1}(n),\Lambda_{m}^{+}\sigma_3\delta w(n)]. \tag{4.7}$$

The proof that $\Omega^{(m)}$ are symplectic is most easily performed as in Ref. 9 after recalculating $\Omega^{(m)}$ in terms of the scattering data variations, which now gives

$$\Omega^{(m)} = 2i \oint_{S^1} \frac{dz}{z} z^{2m}\, \delta\hat{p}(z) \wedge \delta\hat{q}(z) + 4i \sum_{j=1}^{N} [z_{j+}^{2m}\, \delta\hat{p}_j^{+} \wedge \delta\hat{q}_j^{+} + z_{j-}^{2m}\, \delta\hat{p}_j^{-} \wedge \delta\hat{q}_j^{-}]. \tag{4.8}$$

From (4.8) it is obvious that $\{\Omega^{(2m)}, m = 0, \pm 1, \pm 2, \cdots\}$ is a hierarchy of compatible symplectic forms, which generate a hierarchy of Hamiltonian structures for the DEE (3.4). Indeed, the choice $\Omega = \Omega^{(m)}, H = H_{f^{(m)}}$ in (4.2) with $f^{(m)}(z^2) = z^{2m} f(z^2)$ lead to the same DEE (3.4) as $\Omega = \Omega^{(0)}, H = H_f$.

In complete analogy to Refs. 7 and 8, one can define the Lagrange manifold for the DEE (3.4) by

$$\mathfrak{m}(t) \equiv \{X(n,t) \in \mathfrak{m}: [X(n,t), P(n,t,z)] = 0,\ z \in S^1 \cup \Delta\}.$$

Let us list without proof the main properties of $\mathfrak{m}(t)$:

(i) if $X \in \mathfrak{m}$, then $\Lambda_+ X = \Lambda_- X \in \mathfrak{m}$;

(ii) $\dim \mathfrak{m} = \operatorname{codim} \mathfrak{m}$;

(iii) $\sigma_3 \delta w(n) \in \mathfrak{m}$ if and only if $\delta\hat{p}(z) = 0$ for all $z \in S^1 \cup \Delta$, i.e., the restriction of $\Omega^{(m)}|_{\mathfrak{m}} \equiv 0$ for all $m = 0, \pm 1, \cdots$.

Remark 2: From (2.17), (2.9), and (2.4) one verifies that $w(n,t) \in \mathfrak{m}(t)$. This together with the property (i) of $\mathfrak{m}$ gives $f(\Lambda_+)w = f(\Lambda_-)w$, i.e., the operators Λ_+ and Λ_- generate the same DEE (3.4).

Remark 3: If $w(n,t)$ satisfies any of the DEE (3.4), then $\sigma_3(dw/dt) \in \mathfrak{m}(t)$ for all t.

At the end of this paragraph let us consider two particular examples of soluble DEE. They are related to the system (1.2) with simple reductions of the potential, which naturally requires a recalculation of the action-angle variables.

A. The difference nonlinear Schrödinger equation (DNLS)

$$i\frac{dq(n,t)}{dt} = -[1 - \epsilon q^*(n)q(n)][q(n+1) + q(n-1)] + 2q(n), \quad \epsilon = \pm 1, \tag{4.9}$$

is obtained from (3.4) with $f(z^2) = i(2 - z^2 - z^{-2})$ provided the reduction $r(n) = \epsilon q^*(n)$ holds. This reduction imposes the following restrictions on the scattering data:

$$a^+(z) = a^-(1/z^*), \quad b^+(z) = -\epsilon b^{-*}(1/z^*), \quad z_{j+} = 1/z_{j-}^*, \quad c_j^- = \epsilon (c_j^+)^*/(z_{j+}^*)^2. \tag{4.10}$$

As a Hamiltonian and 2-form, generating (4.9), one can choose

$$\Omega_{\mathrm{DNLS}} = \epsilon \Omega^{(0)}|_{q = \epsilon r^*} = -\frac{\epsilon}{\pi} \int_{-\pi}^{\pi} d\tau\, \delta(\arg b^+(e^{i\tau})) \wedge \delta \times \ln[1 - \epsilon|\rho^+(e^{i\tau})|^2] - 4\epsilon \sum_{j=1}^{N} [\delta\zeta_j \wedge \delta\rho_j + \delta\omega_j \wedge \delta\xi_j], \tag{4.11}$$

where $\ln z_{j+}^2 = \zeta_j + i\omega_j$, $\ln(b_j^+/\sqrt{v}) = \xi_j + i\rho_j$;

$$H_{\mathrm{DNLS}} = \epsilon(2C^{(0)} - C^{(1)} - C^{(-1)})|_{q = \epsilon r^*} = -\sum_{n=-\infty}^{\infty} \{q^*(n)[q(n+1) + q(n-1)] + 2\epsilon \ln[1 - \epsilon q^*(n)q(n)]\} = 2\epsilon\left\{-\frac{1}{\pi}\int_{-\pi}^{\pi} d\tau \sin^2\tau \ln[1 - \epsilon|\rho^+(e^{i\tau})|^2] + 2\sum_{j=1}^{N} [\cos\omega_j \sinh\zeta_j - \xi_j]\right\}. \tag{4.12}$$

The explicit form of the action-angle variables is obvious from (4.11). Note that from (4.10) and (3.16) one obtains $C^{(p)} = C^{(-p)*}$

B. The difference modified KdV equation (DMKdV)

$$i\frac{dq(n,t)}{dt} = -[1 - \epsilon q^2(n)][q(n+1) - q(n-1)], \quad \epsilon = \pm 1, \tag{4.13}$$

is obtained from (3.4) with $f(z^2) = z^{-2} - z^2$ provided that the reduction $q(n) = \epsilon r(n)$ holds, which means that

$$a^+(z) = a^-(1/z), \quad b^+(z) = -\epsilon b^-(1/z), \quad z_{j+} = 1/z_{j-}, \quad c_j^+ = \epsilon c_j^- z_{j+}^2. \tag{4.14}$$

As it has been noted in Ref. 12, $\Omega^{(0)}$ vanishes identically if this reduction is imposed. Therefore, we should use another symplectic structure from the hierarchy, e.g.,

$$\Omega_{\mathrm{MDKdV}} = i\Omega^{(-1)}|_{q = \epsilon r} = -2\epsilon \times \sum_{n=-\infty}^{\infty} \left[2\delta q(n) \wedge \delta q(n+1) + \delta \ln h(n) \wedge \delta \times \left(\sum\nolimits_n^+ q(k)q(k-1)\right)\right] = -\frac{2}{\pi}\int_0^{\pi} d\tau \sin 2\tau\, \delta \times \ln[1 - \epsilon\rho^+(e^{i\tau})\rho^+(e^{-i\tau})] \wedge \delta \times \left[-\frac{i}{2}\ln\frac{b^+(e^{i\tau})}{b^+(e^{-i\tau})}\right] + 4\sum_{j=1}^{N} \delta\cosh(\zeta_j + i\omega_j) \wedge \delta(\xi_j + i\rho_j), \tag{4.15}$$

$\ln z_{j+}^2 = \zeta_j + i\omega_j$, $\ln(b_j^+/\sqrt{v}) = \xi_j + i\rho_j$.

The corresponding Hamiltonian is

$$H_{\mathrm{MDKdV}} = C^{(0)} - C^{(2)} = -\sum_{n=-\infty}^{\infty} \{\ln[1 - \epsilon q^2(n)] + \epsilon q(n)q(n-2)[1 - q^2(n-1)] - \tfrac{1}{2}q^2(n)q^2(n-1)\} = -\frac{2}{\pi}\int_0^{\pi} d\tau \sin^2 2\tau \times \ln[1 - \epsilon\rho^+(e^{i\tau})\rho^+(e^{-i\tau})] - 2\sum_{j=1}^{N} [\zeta_j + i\omega_j - \tfrac{1}{2}\sinh 2(\zeta_j + i\omega_j)].$$

If besides $q(n) = \epsilon r(n)$ we require $q(n) = \epsilon r^*(n)$, then the scattering data $\mathcal{T}$, (2.7), will satisfy both (4.10) and (4.14). In this case the eigenvalues appear either in four tuples $(z_{j+}, z_{j+}^*,$

$-z_{j+}$, $-z_{j+}^*$)or pairwise if among z_{j+} there occur real or pure imaginary numbers. Let us introduce the notations:

$$\begin{aligned} z_{j+}^2 &= e^{\xi_j + i\omega_j}, \quad b_j^+/\sqrt{v} = e^{\xi_j + i\rho_j}, \quad j = 1,...,N_1, \\ z_{\alpha+}^2 &= e^{\epsilon_\alpha}, \quad b_\alpha^+/\sqrt{v} = e^{\gamma_\alpha}, \quad \alpha = 1,...,N_2, \\ z_{\beta+}^2 &= -e^{\eta_\beta}, \quad b_\beta^+/\sqrt{v} = e^{\theta_\beta}, \quad \beta = 1,...,N_3, \\ & 2N_1 + N_2 + N_3 = N. \end{aligned} \tag{4.16}$$

Then the 2-forms $i\Omega^{(m)} = -i\Omega^{(-m)}$ become real and equal to

$$\begin{aligned} i\Omega^{(m)} = \frac{2}{\pi}\int_0^\pi d\tau \sin 2m\tau\, \delta(\ln[1-\epsilon|\rho^+(e^{i\tau})|^2]) \\ \wedge \delta(\arg b^+(e^{i\tau})) \\ - \frac{8}{m}\sum_{j=1}^{N_1}\{\delta[\cos(m\omega_j)\cosh(m\xi_j)]\wedge\delta\beta_j \\ - \delta[\sin(m\omega_j)\sinh(m\xi_j)]\wedge\delta\rho_j\} \\ - \frac{4}{m}\sum_{\alpha=1}^{N_2}\delta\cosh(m\epsilon_\alpha)\wedge\delta\gamma_\alpha \\ - \frac{4(-1)^m}{m}\sum_{\beta=1}^{N_3}\delta\cosh m\eta_\beta\wedge\delta\theta_\beta. \end{aligned} \tag{4.17}$$

From (4.17) with $m = 1$ we easily get the action-angle variables for the MDKdV (4.13) with real-valued $q(n)$. If in (4.13) we change the variables to $u(n) = \operatorname{arctanh} q(n)$ for $\epsilon = 1$, and $u(n) = \operatorname{arctanh} q(n)$ for $\epsilon = -1$, we obtain another interesting DEE:

$$\begin{aligned} \frac{du(n,t)}{dt} &= \tan u(n+1) - \tan u(n-1), \quad \epsilon = 1, \\ \frac{du(n,t)}{dt} &= \tanh u(n+1) - \tanh u(n-1), \quad \epsilon = -1. \end{aligned} \tag{4.18}$$

The equivalence of (4.13) and (4.18) is obvious only for $\epsilon = -1$; for $\epsilon = 1$ the change of the variables $u(n) = \arctan q(n)$ is singular.

There are more examples of interesting DEE which can be obtained from (3.4). Obviously for all of them one can calculate the Hamiltonian structures and the action-angle variables, following the above considerations.

V. QUANTUM DIFFERENCE NONLINEAR EQUATIONS

The nonlinear DEE mentioned above can be solved by a quantum version of IST. Let us consider quantum DNS (4.9) where now the quantities $q(n)$ and $q^+(n)$ are operators with commutation relations ($m,n = 0, \pm 1, \pm 2,..., \pm N$)

$$[q(m),q^+(n)] = \hbar[1-\epsilon q^+(n)q(n)]\delta(n-m). \tag{5.1}$$

Hereafter we shall use the normal ordering with respect to q and q^+. For finite N we can realize these operators in the state space $\mathscr{H}^{(N)}$:

$$\begin{aligned} \mathscr{H}^N &= \bigotimes_{n=-N+1}^{N}\mathscr{H}_n, \\ \mathscr{H}_n &= \mathscr{L}\{|0\rangle_n,|1\rangle_n,|2\rangle_n,\cdots\}, \end{aligned} \tag{5.2}$$

where $\mathscr{L}$ denotes closure of a linear space $\{\cdots\}$ and $|k\rangle_n = (q^+(n))^k|0\rangle_n, q(n)|0\rangle_n = 0$. As a consequence of (5.1) the norm in $\mathscr{H}^{(N)}$ is positive definite provided ${}_n\langle 0|0\rangle_n = 1$:

$$\langle k|l\rangle_n = \delta_{kl}(\hbar)^k\prod_{m=1}^{k}c_m, \quad c_m = \frac{1-e^{2\eta m}}{1-e^{2\eta}}. \tag{5.3}$$

The parameter $\eta = \frac{1}{2}\ln(1-\epsilon\hbar)$ is more appropriate in the following formulas. In order that the Heisenberg equations of motion coincide with (4.9), we must add the quantum corrections to the classical expression of the Hamiltonian (4.12):

$$\begin{aligned} H = -\sum_n\{q^+(n)[q(n+1)-q(n-1)] \\ -[2\hbar/\ln(1-\epsilon\hbar)]\ln[1-\epsilon q^+(n)q(n)]\}. \end{aligned} \tag{5.4}$$

The quantum version of IST (QIST) also uses an auxiliary linear problem. In this case we can take the same L operator (1.2), $r(n) = \epsilon q^+(n)$, with its entries as operators in $\mathscr{H}_n$ (5.2). The main step of QIST is the determination of commutation relations of the quantum scattering data or, to be more precise, the operator-valued entries of the monodromy matrix

$$T_N(z) = L_N(z)L_{N-1}(z)\cdots L_{-N+1} = \begin{pmatrix} A_N(z) & B_N(z) \\ C_N(z) & D_N(z) \end{pmatrix}, \tag{5.5}$$

$$R(\varphi)[T_N(z)\otimes T_N(\zeta)] = [I\otimes T_N(\zeta)][T_N(z)\otimes I]R(\varphi), \tag{5.6}$$

where $R(\varphi)$ is a 4×4 c-number matrix or intertwining operator, I is the identity operator in C^2, $\exp\varphi = z/\zeta$. The R matrix can be calculated from the very same relation (5.6) but with $L_n(z)$, $L_n(\zeta)$ instead of $T_N(z)$, $T_N(\zeta)$:

$$\begin{aligned} R(\varphi) &= \begin{pmatrix} a & 0 & 0 & 0 \\ 0 & b^- & c & 0 \\ 0 & c & b^+ & 0 \\ 0 & 0 & 0 & a \end{pmatrix}, \\ a &= \sinh(\varphi-\eta), \quad b^\pm = e^{\pm\eta}\sinh\varphi, \\ c &= -\sinh\eta. \end{aligned} \tag{5.7}$$

For finite chain with periodic boundary conditions ($2N + k = k$, mod $2N$) the trace of the monodromy matrix $t_N(z) = A_N(z) + D_N(z)$ is the generating functional of the quantum integrals of the motion. In order to define eigenstates and eigenvalues of $t_N(z)$, we shall need the following commutation relations (5.6):

$$[t_N(z),t_N(\zeta)] = 0, \quad [C_N(z),C_N(\zeta)] = 0, \tag{5.8}$$

$$\begin{aligned} A_N(z)C_N(\zeta) &= [1/b^-(\varphi)]C_N(\zeta)A_N(z) \\ &\quad -[c(\varphi)/b^-(\varphi)]C_N(z)A_N(\zeta), \\ D_N(\zeta)C_N(z) &= [1/b^-(\varphi)]C_N(z)D_N(\zeta) \\ &\quad -[c(\varphi)/b^-(\varphi)]C_N(\zeta)D_N(z). \end{aligned} \tag{5.9}$$

Since, when applied to the vacuum $|0\rangle = \Pi_{n=-N+1}^{N}|0\rangle_n$, $L_n(z)$ becomes triangular, one easily finds for the action of $A_N(z)$, $B_N(z)$, $D_N(z)$ on $|0\rangle$

$$\begin{aligned} A_N(z)|0\rangle &= z^{2N}|0\rangle, \quad D_N(z)|0\rangle = z^{-2N}|0\rangle, \\ B_N(z)|0\rangle &= 0. \end{aligned} \tag{5.10}$$

Using (5.8)–(5.10) via the general scheme of the QIST,[17,18] one constructs the eigenstates of $t_N(z)$:

$$|z_1,...,z_n\rangle = \prod_{k=1}^{n}C_N(z_k)|0\rangle \tag{5.11}$$

provided the quasimomenta z_k satisfy the algebraic equations ($\exp\lambda = z$)

$$(z_k)^{4N} = \prod_{l \neq k} \frac{\sinh(\lambda_k - \lambda_l + \eta)}{\sinh(\lambda_k - \lambda_l - \eta)}, \quad k = 1,2,\ldots,n. \quad (5.12)$$

The corresponding eigenvalue is given by

$$V(z,\{z_k\}_1^n) = z^{2N} \prod_{k=1}^{n} \frac{e^{\eta} \sinh(\lambda - \lambda_k - \eta)}{\sinh(\lambda - \lambda_k)} + z^{-2N} \prod_{k=1}^{n} \frac{e^{\eta} \sinh(\lambda - \lambda_k + \eta)}{\sinh(\lambda - \lambda_k)}. \quad (5.13)$$

For the energy of the state (5.11) we have

$$E(\{z_k\}_1^n) = \sum_{k=1}^{n} \epsilon(z_k), \quad \epsilon(z) = 2\hbar - z^2 - z^{-2}. \quad (5.14)$$

There exist different phases in the limit $N \to \infty$. The phase with finite number of particles is the simplest one. The state space has the Fock type structure with vacuum $|0\rangle$ and creation operators $q^+(n)$, $n = 0, \pm 1, \pm 2, \cdots$, or

$$R^+(z) = \lim_{N\to\infty} C_N(z)/z^{2N} A_N(z), \quad |z| = 1. \quad (5.15)$$

The additional factor z^{2N} is a consequence of the transition matrix definition

$$T(z) = \lim_{N\to\infty} E^{-N}(z) T_N(z) E^{-N}(z) = \lim_{N\to\infty} \begin{vmatrix} z^{-2N} A_N(z) & B_N(z) \\ C_N(z) & z^{2N} D_N(z) \end{vmatrix} = \begin{vmatrix} A(z) & B(z) \\ C(z) & D(z) \end{vmatrix}, \quad (5.16)$$

where $E(z) = \langle 0|L_n(z)|0\rangle = \mathrm{diag}(z,z^{-1})$. It is possible to define operator-valued Jost solutions (in the weak sense) and their analytic properties and relations to the transition matrix $T(z)$. The inverse to $L_n(z)$ is $[\rho_n = 1 - \epsilon q^+(n) q(n)]$

$$L_n^{-1}(z) = \frac{e^{-\eta}}{\rho_n} V L_n(e^{-\eta}/z) V^{-1}, \quad V = \mathrm{diag}(e^{-\eta/2}, -e^{\eta/2}). \quad (5.17)$$

Using $L_n^t(z) = \sigma_1 L_n(1/z)\sigma_1$, we get

$$T_N^{-1}(z) = Q_N^{-1} W T_N^t(e^{\eta} z) W^{-1}, \quad W = V\sigma_1, \quad Q_N = \prod_{n=-N+1}^{N} e^{\eta} \rho_n. \quad (5.18)$$

The operator $R^+(z)$, $R(z) = \epsilon D^{-1}(z) B(z)$ is called quantum scattering data. They are generators of the Zamolodchikov–Faddeev algebra. By means of the formulas (5.17) and (5.18) one can obtain a quantum analog of (2.5), i.e., the quantum Riemann problem. The reconstruction of the local quantum operators $q(n)$ and $q^+(n)$ from the quantum scattering data would enable one to calculate the Green's functions of this model.

ACKNOWLEDGMENTS

The authors are deeply grateful to E. Kh. Khristov and A. G. Reyman for helpful discussions.

[1] V. E. Zakharov, S. V. Manakov, S. P. Novikov, and L. P. Pitaevskii, *Soliton Theory: The Inverse Problem Method* (in Russian) (Nauka, Moscow, 1980).
[2] M. Ablowitz, D. Kaup. A. Newell, and H. Seeger, Stud. Appl. Math. **53**, 249 (1974).
[3] M. Ablowitz, Stud. Appl. Math. **58**, 17 (1978).
[4] L. D. Faddeev. in *Solitons*, edited by R. K. Bullough and P. Candrey, Topics in Current Physics (Springer-Verlag, Berlin, 1980), Vol. 17, p. 339.
[5] D. J. Kaup, Math. Anal. Appl. **54**, 789 (1976).
[6] D. J. Kaup and A. C. Newell, Adv. Math. **31**, 67 (1979).
[7] V. S. Gerdjikov and E. Kh. Khristov, Mat. Zametki **28**, 501 (1980) (in Russian); Bulg. J. Phys. **7**, 28 (1980) (in Russian).
[8] V. S. Gerdjikov and E. Kh. Khristov, Bulg. J. Phys. **7**, 119 (1980) (in Russian).
[9] P. P. Kulish and A. G. Reiman, Zap. Nauchnich Semin. LOMI **77**, 134 (1978) (Leningrad, USSR, in Russian).
[10] S. V. Manakov, Zh. Eksp. Teor. Fiz. **67**, 543 (1974) [Sov. Phys. JETP **40**, 269 (1975)].
[11] M. Ablowitz and J. F. Ladik, J. Math. Phys. **16**, 598 (1978); **17**, 1011 (1976).
[12] F. Kako and N. Mugibayashi, Prog. Theor. Phys. **61**, 776 (1979).
[13] S. C. Chiu and J. F. Ladik, J. Math. Phys. **18**, 690 (1977).
[14] D. Levi and O. Ragnisco, Lett. Nuovo Cimento **22**, 691 (1978); M. Bruschi, D. Levi, and O. Ragnisco, J. Phys. A: Math. Gen. **13**, 2531 (1980).
[15] V. Ju. Novokshenov and I. T. Khabibulin. Dokl. Akad. Nauk SSSR **257**, 543 (1981).
[16] V. S. Gerdjikov, M. I. Ivanov, and P. P. Kulish,, JINR Preprint E2-80-882, Dubna, USSR, 1981.
[17] L. D. Faddeev, Sov. Sci. Rev. Math. Phys. C **1**, 107 (1980).
[18] P. P. Kulish and E. K. Sklyanin, in *Integrable Quantum Field Theories*, edited by J. Hietarinta and C. Montonen, Lecture Notes in Physics (Springer-Verlag, Berlin, 1982), Vol. 151, p. 61.
[19] P. P. Kulish, Lett. Math. Phys.. **5**, 191 (1981).
[20] I. T. Khabibulin, Dokl. Akad. Nauk SSSR **249**, 67 (1979).
[21] V. S. Gerdjikov, JINR Preprint E2-81-652, Dubna, USSR, 1981.
[22] I. M. Gel'fand and L. A. Dickey, Funct. Anal. Appl. **11** (2), 11 (1977) (in Russian).
[23] V. S. Gerdjikov and M. I. Ivanov, JINR Preprint, P5-82-412, Dubna, USSR 1982.

A new integral equation for summing Feynman graph series (general scalar Lagrangian case)

C. Gilain and D. Lévy
Service de Physique Théorique, Division de la Physique, Centre d'Etudes Nucleaires de Saclay, B.P. N° 2, 91190 Gif-Sur-Yvette, France

(Received 30 June 1981; accepted for publication 14 January 1983)

The Schwinger parameter formalism is used to derive a new integral equation verified by the "open" four-point amplitude built from any scalar Lagrangian. This integral equation is a generalization of the one already obtained and studied by the authors in the φ^3 ladder graph case. One of the main results obtained here is a new representation of the Feynman amplitudes: the so-called β-representation, which expresses the Bethe–Salpeter structure of a graph in the Schwinger parameter space. The integrand of the β-representation satisfies a recurrence relation which is used to sum the perturbation series, and which leads to an integral equation for its sum. The expression of this integral equation is also given in some particular cases (particular values of the invariants, particular classes of graphs, etc.). The Mellin transform of the open amplitude satisfies a similar integral equation which may be used to describe the Regge behavior.

PACS numbers: 02.30.Rz, 11.10.Mn, 11.10.Ef

INTRODUCTION

This work takes place in a set of studies whose aim is to obtain, in the framework of Lagrangian field theory, results on the infinite sum of the perturbation series, whatever the value of the coupling constant is. The common feature of this set of studies is that they are performed in the framework of the Schwinger parametrization of Feynman integrals.

Some years ago, powerful results were obtained on the complete asymptotic behavior of each term of the perturbation series (mainly in the Regge limit) for scalar Lagrangians, and on their sum.[1]

Another way, more recently explored, provided results on the four-point amplitude which are not restricted to asymptotic values of the invariants. It relies on the existence of a new integral equation (IE) that does not apply to the amplitude itself, as it is the case for the Bethe–Salpeter (BS) integral equation, but rather to a new quantity: the "open amplitude." The first step has consisted of deriving this new IE in the restricted case of φ^3 ladder subseries.[2] The present work is the generalization of this first step to the complete perturbation series built from any scalar Lagrangian.

The advantages of working with IE are well known: Under conditions of sufficient regularity of the inhomogeneous term and of the kernel, the solution of an IE can be computed. For example, when an IE satisfies the conditions of the Fredholm theorems, its solution is the ratio of two holomorphic functions, and its singularities are poles, given by the zeros of the Fredholm denominator, which depends only on the kernel.

As for the Bethe–Salpeter IE in momentum space, our IE makes use of the Bethe–Salpeter structure of the amplitude, that is to say, its decomposition into generalized ladders whose rungs are t-channel two-particle irreducible subgraphs (t-2PI subgraphs) [see Fig. 1(b)]. The Bethe–Salpeter IE reflects directly the factorization of the integrand when the Feynman amplitude is expressed as an integral over internal 4-momenta.

The Schwinger parametrization of the same amplitude destroys this factorization. For example, the quadratic form $D_G(\alpha)$ which appears in the integrand is a complicated function of all the Schwinger parameters of the graph G. However, the ladder structure of the graph was still reflected, in the φ^3 ladder case, by the open amplitude built in Ref. 2: Inside the set of all integration variables of the Schwinger parametrization [Eq. (1)], we have distinguished there a subset $\alpha_c = \{\alpha_{i_1}, \alpha_{i_2}, \cdots\}$, called the closing variables. The open amplitude $O_G(\alpha_c)$ is then defined by the same integration as the Feynman amplitude I_G itself, except that the closing variable integration is not performed. Of course, the Feynman amplitude of the graph G is the integral of $O_G(\alpha_c)$:

$$I_G = \int d\alpha_c \; O_G(\alpha_c).$$

We have shown that the open amplitude obeys a recurrence law on the number of rungs of the ladder. This recurrence law is the key result from which the existence and the properties of the IE verified by the infinite sum of the open amplitudes is deduced.

We show in the present paper that an analogous work can be done independently of the ladder restriction and for any scalar Lagrangian φ^n.

Although the Bethe–Salpeter IE and our IE, both, reflect the Bethe–Salpeter structure of the amplitude, they are qualitatively different: It is not possible to transform one of them into the other. They concern different amplitudes and different variables:

(i) Our IE is not satisfied by the amplitude, but by the open amplitude.

(ii) The integration variables in the Bethe–Salpeter IE are the external momenta, whereas our IE involves as integration variables the closing variables, i.e., a given subset of the Schwinger parameters.

A consequence of the qualitative difference between the two IE appears in the actual computation: In the φ^3 ladder case, our IE turns to be very appropriate; indeed it provides

 0022-2488/84/010035-13$02.50

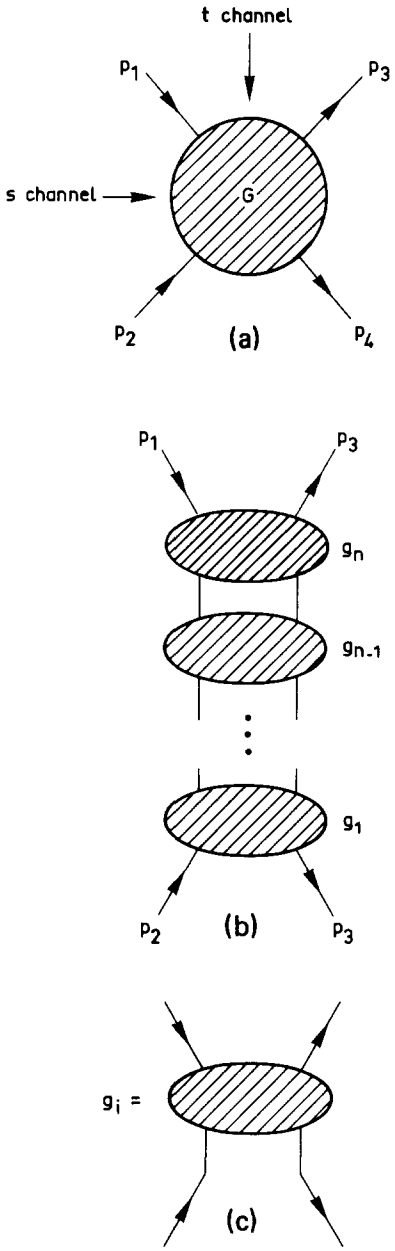

FIG. 1. The kinematics of the four point graph and its Bethe–Salpeter structure: $G = \cup_i g_i$. In the first part of this work, the subgraphs g_i have no special property of reducibility: There is, in general more than one such decomposition of the graph G. In the second part, g_i are restricted to be two particule irreducible in the t channel: There is an only Bethe–Salpeter decomposition of G.

not only the Regge singularities but gives directly the amplitude itself, whereas the Bethe–Salpeter IE has to be studied by two different methods to obtain the same results.[3] From the method initiated by Lee and Sawyer, indeed, the Regge singularity analysis is obtained from an analytic continuation of the partial waves, the problem of the summation of the partial wave expansion, which gives the amplitude, being left over. If one is interested in the amplitude, other methods must be used (such as the perturbation-theoretical integral representation,[3] for instance), and so the complete study of the properties of the amplitude through the Bethe–Salpeter equation is difficult and lengthy.

Though our integral equation is singular, we prove the existence and uniticity of its solution and make explicit its singularity structure. Our fundamental result is that, for each given value of the coupling constant λ, the solution can be written as a finite sum of solutions of Fredholm equations plus a function which is the sum of a convergent series in λ. Moreover, our IE allows simple approximate quantitative computation: For example, the trace approximation gives good results for the dominant trajectory.[4]

To achieve the generalization of our IE, we split the problem into two steps: First we neglect the UV divergences and focus our attention on the algebraic structure. This step is completely done for all scalar Lagrangians (Secs. I–IV). Then we have to face the ultraviolet divergence problem, namely, in our approach the compatibility of the Bergère–Zuber[5] renormalization procedure and of the structure of the recurrence relation [see Sec. I A, Eqs. (32)]. This is done here only for the φ^3 interaction.

The price to pay for the generality of our result is, of course, the formal character of the equation obtained. The kernel, which governs the properties of the solution, is given in terms of an infinite series. The logical following step of our program is the link between the properties of this series and those of the four-point amplitude.

We conclude this introduction with a more precise presentation of the content of this paper. We obtain two new results: the first one, presented in Sec. I, is a scalar integral representation for the Feynman amplitudes, which is an alternative to the Schwinger one. The Schwinger α-parametrization gives in fact the amplitude associated with a given graph as a multiple scalar integral involving as many scalar variables as internal lines in the graph. There appears in the integrand no factorization according to the "rungs" of the generalized ladder [see Fig. 1(b)]. Our aim is to make explicit on the Schwinger integrand the BS structure of a graph. This requires, as presented in subsection I A, a change in the choice of the invariants and consequently of the topological functions which are their coefficients in the quadratic from $D_G(\alpha)$. In subsection I B, important properties of quasifactorization and of recurrence of these topological functions of the graphs are given. In subsection I C, the structure of the quadratic form $D_G(\alpha)$ is made precise. In subsection I D, we are then led to establish our alternative parametrization for the Feynman amplitude: the β-parametrization. Let us consider the graph G of Fig. 1(b), which is a generalized ladder with n rungs g_i. The β-parametrization is an integral over $6 \times n$ scalar variables (the β variables) and its integrand is a product of two factors:

(1) The first one is completely factorized, and is a product of n functions of six variables, each one being attached to one rung g_i.

(2) The other one is a global factor, which depends only on n and is independent of the structure of each rung: It is the skeleton of BS structure of the graph.

The β variables represent appropriate combinations of the topological polynomials associated with each g_i. Their variation domains are always explicitly indicated by means of θ step functions.

We have then the adequate tools for proving the existence of the integral equation, which is the second new result

of this work. It is the aim of Sec. II. From the recurrence relation obeyed by the open amplitudes (defined in II A) we deduce the integral equation satisfied by their sum (II B and II C).

Then we discuss in this framework the Regge limit, that is to say, the structure of the IE in the Mellin space (Sec. III).

Some physically interesting particular situations are grouped in the fourth section: forward scattering, bound states equation,... . In this section one can find also the simplified expression of the IE for a special class of kernels (the ladder with generalized rungs; see Fig. 4), or for particular values of the variables β.

Finally, the renormalization problem is achieved for the φ^3 interaction Lagrangian in Sec. V. Some technical points are grouped in the Appendix.

I. BETHE–SALPETER STRUCTURE AND TOPOLOGICAL POLYNOMIALS

We consider here the scalar Lagrangian field theories. With any graph G is associated its Feynman amplitude, whose Schwinger integral representation is

$$I_G^{\epsilon}(P) = \lambda^{N(G)}(ie^{-i\epsilon})^{-\omega(G)/2}\int_0^{\infty}\prod_{a=1}^{l(G)} d\alpha_a \times\exp\left(-ie^{-i\epsilon}\sum_{a=1}^{l(G)}\alpha_a m^2\right) \times R\left(\frac{e^{ie^{-i\epsilon}D_G(\alpha)}}{P_G^2(\alpha)}\right). \tag{1}$$

In (1), $\omega(G)$ is the superficial degree of divergence of the graph G:

$$\omega(G) = 4L(G) - 2l(G),$$

where $L(G)$, $l(G)$, and $N(G)$ are, respectively, the number of independent loops, of internal lines, and of vertices of G. P is the set of external 4-momenta, and λ is the coupling constant of the theory. There is a scalar variable α_a attached to each internal line of the graph. The set $\{\alpha_1,\alpha_2,...,\alpha_{l(G)}\}$ will be noted α or α_G every time an ambiguity is possible. The operator R is the Bergère–Zuber[5] subtraction operator which ensures the ultraviolet (UV) convergence of the Feynman amplitude. In this work we will pay no attention to the UV convergence problems, but for the case of the interaction φ^3 which we treat exhaustively (see Sec. V).

In Minkowsky space the amplitude is the limit $\epsilon\to 0_+$ of I_G^{ϵ}. As we are mainly interested with the algebraic structure of the integrand, and not with the convergence conditions of the integral, we place our problem in Euclidean space, in which the amplitude is given from (1) with $\epsilon = \pi/2$:

$$I_G(P) = \int_0^{\infty} d\mu_G(\alpha_G) e^{D_G(\alpha)}, \tag{1'}$$

where

$$d\mu_G(\alpha_G) = \lambda^{N(G)}\prod_{a=1}^{l(G)} d\alpha_a \frac{\exp(-\Sigma_{a=1}^{l(G)}\alpha_a m^2)}{P_G^2(\alpha)}. \tag{2}$$

The function $D_G(\alpha)$ is a quadratic form built from the external 4-momenta. In 2 particles $\to$2 particles case which we are studying, it is equal to

$$D_G(\alpha) = s\frac{A_G^s(\alpha)}{P_G(\alpha)} + t\frac{A_G^t(\alpha)}{P_G(\alpha)} + u\frac{A_G^u(\alpha)}{P_G(\alpha)} + \sum_{i=1}^{4} p_i^2\frac{A_G^i(\alpha)}{P_G(\alpha)}. \tag{3}$$

s,t,u are the Mandelstam invariants built from the external momenta p_i [see Fig. 1(a)], and $P_G(\alpha)$, $A_G^s(\alpha)$, $A_G^t(\alpha)$, $A_G^u(\alpha)$, $A_G^i(\alpha)$ $(i = 1,...,4)$ are the topological polynomials, characteristic of the graph G. Their definition can be found in the Appendix A of Ref. 2. Let us only say that they are polynomials, homogeneous in the set α, and of degree $L(G)$ for P_G, $(L(G)+1)$ for the other ones.

The problem we solve here is the adaptation of this formalism in order to make use of the Bethe–Salpeter structure of the four-point amplitude: Any graph composed of at least n two-particle irreducible subgraphs in the t channel may be drawn as the generalized ladder of Fig. 1(b). As we consider the two vertical lines attached under each bubble as internal lines of the corresponding subgraph, the graph G is exactly the union of each subgraph g_i:

$$G = \{g_1,...,g_n\}.$$

In this first section, except for the existence of the two additional vertical lines, the graphs g_i can have absolutely any structure: They can be reducible or irreducible.

The problem stands of course in the fact that the integrand $e^{D_G(\alpha)}/P_G^2(\alpha)$ in (1') is not factorized in functions, each attached to each subgraph g_i. As we want to build G as a ladder of rungs g_i, we are faced with the necessity of performing loop integrals to link two subgraphs: In the following paragraph a change of external momenta is performed in order to make easier this integration.

A. Alternative expression for the quadratic form $D_G(\alpha)$

The first step consists in modifying the usual form of $D_G(\alpha)$. We choose as external momenta the three combinations,

$$\begin{aligned} q_1 &= \tfrac{1}{2}(p_1+p_3),\\ q_2 &= \tfrac{1}{2}(p_2+p_4),\\ q &= (p_1-p_3) = (p_4-p_2), \end{aligned} \tag{4}$$

and build their associated invariants,

$$\begin{aligned} s_{11} &= q_1^2 = \tfrac{1}{2}(p_1^2+p_3^2) - \tfrac{1}{4}t,\\ s_{12} &= 2q_1q_2 = \tfrac{1}{2}(s-u) = s + \frac{1}{2}\left(t - \sum_{i=1}^{4} p_i^2\right),\\ s_{22} &= q_2^2 = \tfrac{1}{2}(p_2^2+p_4^2) - \tfrac{1}{4}t,\\ s_1 &= 2qq_1 = p_1^2 - p_3^2,\\ s_2 &= 2qq_2 = p_4^2 - p_2^2,\\ s_t &= q^2 = t. \end{aligned} \tag{5}$$

The seven invariants s,t,u,p_i^2, $i = 1,2,3,4$, are not independent $(s+t+u = \Sigma p_i^2)$, so it is enough to define the six independent invariants s_j, $j\in K$, where K is the set of indices:

$$K = \{11,12,22,1,2,t\}.$$

Putting in (3) the inverse relations of (5), which gives the Mandelstam invariants in terms of the s_j variables, we find

$$D_G(\alpha) = \sum_{j\in K} s_j \beta^j_G(\alpha). \tag{6}$$

The $\beta^j_G(\alpha)$ are the combinations of topological polynomials associated with s_j:

$$\begin{aligned}
\beta^{11}_G(\alpha) &= [1/P(\alpha)][A^s_G(\alpha) + A^u_G(\alpha) + A^1_G(\alpha) + A^3_G(\alpha)],\\
\beta^{12}_G(\alpha) &= [1/P_G(\alpha)][A^s_G(\alpha) - A^u_G(\alpha)],\\
\beta^{22}_G(\alpha) &= [1/P_G(\alpha)][A^s_G(\alpha) + A^u_G(\alpha) + A^2_G(\alpha) + A^4_G(\alpha)],\\
\beta^1_G(\alpha) &= \tfrac{1}{2}[1/P_G(\alpha)][A^1_G(\alpha) - A^3_G(\alpha)],\\
\beta^2_G(\alpha) &= \tfrac{1}{2}[1/P_G(\alpha)][A^4_G(\alpha) - A^2_G(\alpha)],\\
\beta^t_G(\alpha) &= [1/P_G(\alpha)]\{A^t_G(\alpha)\\
&\quad + \tfrac{1}{4}[A^1_G(\alpha) + A^2_G(\alpha) + A^3_G(\alpha) + A^4_G(\alpha)]\}.
\end{aligned} \tag{7}$$

The set of the six functions β^j_G, $j\in K$, will be noted β_G.

Let us now give the variation domain of β_G, when the α parameters vary from zero to infinity. For the most general graph, the topological functions $A^j_G(\alpha)/P_G(\alpha)$, $j = s,t,u,1,2,3,4$ are independent and vary from zero to plus infinity. Then, using (7), one obtains the bounded domain:

$$|\beta^{12}_G| + 2|\beta^1_G| \leqslant \beta^{11}_G, \tag{8a}$$

$$|\beta^{12}_G| + 2|\beta^2_G| \leqslant \beta^{22}_G, \tag{8b}$$

$$|\beta^1_G| + |\beta^2_G| \leqslant 2\beta^t_G. \tag{8c}$$

In opposition with $A^j_G(\alpha)/P_G(\alpha)$, some of the β^j_G may become negative.

In fact, we will see in the following that the six β^j, $j\in K$, do not play an equivalent role: We have to group them into two sets:

$$\gamma = \{\beta^{12}, \beta^2, \beta^{22}\} \tag{9a}$$

and

$$\bar\gamma = \{\beta^{11}, \beta^1, \beta^t\}. \tag{9b}$$

Thus, the variation domain (8) may be built in two steps: the variation domain of $\bar\gamma$, γ being kept fixed and the domain for γ, whatever $\bar\gamma$ is. These variation domains play an important role in the following. To each of them are attached, respectively, the function Θ_1, Θ_2, Θ_3 with

$$\Theta_1(\beta) = \Theta_2\cdot\Theta_3, \tag{10a}$$

where

$$\Theta_2(\gamma,\bar\gamma) = \theta(\beta^{11} - |\beta^{12}| - 2|\beta^1|)\cdot\theta(2\beta^t - |\beta^1| - |\beta^2|), \tag{10b}$$

$$\Theta_3(\gamma) = \theta(\beta^{22} - |\beta^{12}| - 2|\beta^2|), \tag{10c}$$

with θ the usual step function.

B. Bethe–Salpeter structure of the β functions

The theorem we establish now concerns the Bethe–Salpeter structure of the topological polynomials. It is the result upon which the whole work relies.

Theorem 1: Let us consider a graph G which can be written as a generalized ladder with n rungs [Fig. 1(b)]:

$$G = \{g_1, g_2, \ldots, g_n\}.$$

Then there exists seven functions of $6\times n$ variables, S^j_n, $j\in K$, and S^0_n, verifying the three following properties:

—They are independent of the graph g_i, depending only on the number n of such subgraphs,

$$-\beta^j_G(\alpha_G) = S^j_n(\beta_{g_1}(\alpha_{g_1}), \beta_{g_2}(\alpha_{g_2}), \ldots, \beta_{g_n}(\alpha_{g_n})), \tag{11}$$

and

$$P_G(\alpha_G) = \left(\prod_{i=1}^n P_{g_i}(\alpha_{g_i})\right) S^0_n(\beta_{g_1}(\alpha_{g_1}), \ldots, \beta_{g_n}(\alpha_{g_n})). \tag{12}$$

—The functions S^j_n, $j\in K$, and S^0_n verify the following recurrence relations:

$$S^j_n(\beta_1,\ldots,\beta_n) = S^j_{n-1}(\beta_1,\ldots,\beta_{n-2}, S_2(\beta_{n-1},\beta_n)), \tag{13}$$

$$S^0_n(\beta_1,\ldots,\beta_n) = S^0_{n-1}(\beta_1,\ldots,\beta_{n-2}, S_2(\beta_{n-1},\beta_n)) S^0_2(\beta_{n-1},\beta_n). \tag{14}$$

The meaning of this theorem is the following: The β functions associated with the graph are themselves functions of the β functions associated with each subgraph g_i in a way which is independent of the graph G except for the number of subgraphs g_i.

It is this property which replaces the factorization property of the integrand in the momentum space.

Proof: The proof proceeds through two stages: first we show it directly for the case $n = 2$. Then the proof works by recurrence.

$n = 2$ case: Let us consider a graph G which is two-particule reducible in the t channel (see Fig. 2): $G = \{g_1, g_2\}$.

We write the amplitude I_G in terms of the convolution of the two amplitudes I_{g_1} and I_{g_2}:

$$I_G(q_1,q_2,q) = cst\int d^4q'\, I_{g_1}(-q',q_2,q)\cdot I_{g_2}(q_1,q',q), \tag{15}$$

where

$$q' = \tfrac{1}{2}(p'_4 + p'_2).$$

In the two members of Eq. (15), we use for I the expression (1′), where $D(\alpha)$ is given by (6). After having done the integration over q', we can identify on the two sides the denominators and the coefficients of the invariants. We remark that β_G depends on α_{g_1} and α_{g_2} only through β_{g_1} and β_{g_2}. We thus obtain explicitly the functions S_2:

$$\begin{aligned}
S^{11}_2(\beta_1,\beta_2) &= \beta^{11}_2 - (\beta^{12}_2)^2/(\beta^{11}_1 + \beta^{22}_2),\\
S^{12}_2(\beta_1,\beta_2) &= \beta^{12}_1\beta^{12}_2/(\beta^{11}_1 + \beta^{12}_2),\\
S^{22}_2(\beta_1,\beta_2) &= \beta^{22}_1 - (\beta^{12}_1)^2/(\beta^{11}_1 + \beta^{22}_2),\\
S^1_2(\beta_1,\beta_2) &= \beta^1_2 - \beta^{12}_2(\beta^2_2 - \beta^1_1)/(\beta^{11}_1 + \beta^{22}_2),\\
S^2_2(\beta_1,\beta_2) &= \beta^2_1 + \beta^{12}_1(\beta^2_2 - \beta^1_1)/(\beta^{11}_1 + \beta^{22}_2),\\
S^t_2(\beta_1,\beta_2) &= \beta^t_1 + \beta^t_2 - (\beta^2_2 - \beta^1_1)^2/(\beta^{11}_1 + \beta^{22}_2),
\end{aligned} \tag{16}$$

and finally

$$S^0_2(\beta_1,\beta_2) = \beta^{11}_1 + \beta^{22}_2. \tag{17}$$

n subgraph case: Let us turn now to the graph of Fig. 1(b). We build by recurrence the set of functions S^j_n, $j\in K$ [see (13)]. Inside the graph G we can group together the two last subgraphs g_{n-1} and g_n:

$$G = \{g_1, g_2, \ldots, g_{n-2}, g'_{n-1}\},$$

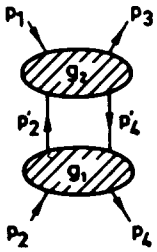

FIG. 2. The graph $G = \{ g_1, g_2 \}$.

with

$$g'_{n-1} = \{ g_{n-1}, g_n \}.$$

If we assume that the S^j_{n-1} functions are known, we have

$$\beta^j_G(\alpha_G) = S^j_{n-1}(\beta_{g_1}(\alpha_{g_1}),..., \beta_{g_{n-2}}(\alpha_{g_{n-2}}), \beta_{g'_{n-1}}(\alpha_{g'_{n-1}})).$$

then using Eq. (16) to compute $\beta_{g'_{n-1}}$, we obtain

$$\beta^j_G(\alpha_G) = S^j_{n-1}(\beta_{g_1}(\alpha_{g_1}),..., \beta_{g_{n-2}}(\alpha_{g_{n-2}}), S_2(\beta_{g_{n-1}}(\alpha_{g_{n-1}}), \beta_{g_n}(\alpha_{g_n}))). \tag{18}$$

The comparison of (18) with (11) proves the existence of S^j_n ($j \in K$) and gives us their recurrence law. With the same procedure, we deduce the recurrence law (14).

This achieves the proof of Theorem 1.

C. Bethe–Salpeter structure of the quadratic form

In this subsection, the dependence of the quadratic form in function of the variables $\bar{\gamma}_n$ is studied. In the recurrence relations (13) and (14), the six variables β_n do not play an equivalent role. The dependence in function of three of them ($\bar{\gamma}_n$) is linear and does not depend on all the $6 \times n - 3$ other variables. It will be seen further that this property allows to obtain an IE with only three integration variables and not six. To lighten the notations, we write $\beta_{(n)} = \{ \beta_1,..., \beta_n \}$. The functions $S^j_n, j \in K$, are homogeneous functions of degree one in the set of the $6 \times n$ variables $\beta_{(n)}$, and S^0_n is homogeneous of degree $(n-1)$ in the same set. We recall that we have defined the two subsets [see (9a) and (9b)]: $\gamma = \{ \beta^{12}, \beta^{22}, \beta^2 \}$ and $\bar{\gamma} = \{ \beta^{11}, \beta^1, \beta^t \}$; we define also the two subsets of indices:

$$K' = \{12,22,2\} \quad \text{and} \quad \bar{K}' = \{11,1,t\}.$$

From (16) and (13), one can show by recurrence that it is possible to define a set of functions $\hat{S}^j_n$ such that

$$\begin{aligned} S^j_n(\beta_{(n)}) &= \hat{S}^j_n(\beta_{(n-1)}, \gamma_n), \quad \text{for } j \in K' \text{ and } j = 0, \\ S^j_n(\beta_{(n)}) &= \hat{S}^j_n(\beta_{(n-1)}, \gamma_n) + \beta^j_n, \quad \text{for } j \in \bar{K}'. \end{aligned} \tag{19}$$

As a direct consequence of Theorem 1, we are led to define a function D_n:

$$D_n(\beta_1,..., \beta_n) \equiv \sum_{j \in K} s_j S^j_n(\beta_1,..., \beta_n). \tag{20}$$

The quadratic form $D_G(\alpha_G)$ has a simple expression in function of D_n:

$$D_G(\alpha_G) = D_n(\beta_{g_1}(\alpha_{g_1}), \beta_{g_2}(\alpha_{g_2}),..., \beta_{g_n}(\alpha_{g_n})). \tag{21}$$

The useful properties of D_n are given in the following theorem.

Theorem 2: The dependence of the D_n function upon the three variables $\bar{\gamma}_n$ of the last graph g_n is explicit and linear:

$$D_n(\beta_{(n)}) = \hat{D}_n(\beta_{(n-1)}, \gamma_n) + \sum_{j \in \bar{K}'} s_j \beta^j_n, \tag{22}$$

where the $\hat{D}_n$ function depends, as far as the last subgraph is concerned, only on the set γ_n. The $\hat{D}_n$ function verifies the recurrence law:

$$\begin{aligned} &\hat{D}_n(\beta_{(n-1)}, \gamma_n) \\ &\quad = \hat{D}_{n-1}(\beta_{(n-2)}, \hat{S}_2(\beta_{n-1}, \gamma_n)) + d(\bar{\gamma}_{n-1}, \gamma_n), \end{aligned} \tag{23a}$$

where

$$\begin{aligned} d(\bar{\gamma}_1, \bar{\gamma}_2) = &- s_{11} \frac{(\beta^{12}_2)^2}{\beta^{11}_1 + \beta^{22}_2} - s_1 \frac{\beta^{12}_2 (\beta^2_2 - \beta^1_1)}{\beta^{11}_1 + \beta^{22}_2} \\ &+ s_t \left(\beta^t_1 - \frac{(\beta^2_2 - \beta^1_1)^2}{\beta^{11}_1 + \beta^{22}_2} \right) \end{aligned} \tag{23b}$$

and where $\hat{S}_2$ represents the set of functions $\{ \hat{S}^j_2, j \in K' \}$.

Proof: The relation (22) follows immediately from Eqs. (19) and (20). The function D_n and the term $\Sigma_{j \in K'} s_j \beta^j_n$, follow the same recurrence law (13) as S^j_n. Thus, using (22), one can obtain the recurrence law (23) for $\hat{D}_n$.

Let us remark that the function d, and the term $(\Sigma_{j \in \bar{K}'} s_j \beta^j_n)$ in (22) correspond exactly to the violation of the * law in the framework of our work on the φ^3 ladder.[2]

D. β-parametrization of the Feynman amplitudes

We are now able to proceed any further and to propose an alternative form for the Schwinger parametrization, form which reflects the Bethe Salpeter structure of the amplitude:

Theorem 3: The amplitude I_G attached to the graph of Fig. 1(b) may be written as

$$I_G = \int \prod_{i=1}^{n} (d\beta_i j_{g_i}(\beta_i)) \frac{e^{D_n(\beta_1,..., \beta_n)}}{[S^0_n(\beta_1,..., \beta_n)]^2}, \tag{24}$$

where

$$j_g(\beta) = \Theta_1(\beta) \int_0^\infty d\mu_g(\alpha_g) \prod_{j \in \bar{K}} \delta(\beta^j - \beta^j_g(\alpha_g)). \tag{25}$$

Proof: Theorem 3 is easily proved if, inside expression (1') where $D_G(\alpha)$ is given by Eq. (21), we insert

$$1 = \int \prod_{i=1}^{n} \prod_{j \in \bar{K}} \delta(\beta^j_i - \beta^j_{g_i}(\alpha_{g_i})) \, d\beta^j_i. \tag{26}$$

Let us make three remarks:

(1) We purposely make explicit the integration region for the β via the factor Θ_1 [see Eq. (10)].

(2) For a given graph G, the decomposition $G = \{ g_1,..., g_n \}$ is not unique, as far as the irreducibility of the subgraphs g_i is not required. In particular, to any graph is associated its β-parametrization with $n = 1$:

$$I_G = \int d\beta j_G(\beta) e^{D_1(\beta)}.$$

(3) The strength of the expression (24) is that the integrand appears as the product of two qualitatively different factors:

(a) $e^{D_n(\beta_{(n)})}/[S_n^0(\beta_{(n)})]^2$ is independent of the characteristics of the graph G but the number n of subgraphs g_i.

(b) The n functions $j_{g_i}(\beta_i)$ depend on the subgraphs g_i.

This factorized structure is the main property which is used to build the integral equation derived in the next section.

II. INTEGRAL EQUATION

The Bethe–Salpeter integral equation is written for the amplitude. It is not the case here. Our work relies upon the properties of the partially integrated integrand. The first subsection is devoted to define this "open amplitude." Then a first form of the integral is given. The third subsection gives the final form of this equation.

From now on and up to the end of the work we consider for each graph its unique decomposition in the generalized ladder [see Fig. 1(b)] of t-2PI subgraphs: Here the notation g_i will always refer to such a two-particle irreducible subgraph.

A. The recurrence relation obeyed by the open amplitude

The open amplitude $O_{G_{n-1}}(\gamma_n)$ is defined by the relation

$$O_{G_{n-1}}(\gamma_n)=\int\prod_{i=1}^{n-1}[d\beta_i\, j_{g_i}(\beta_i)]\,\frac{e^{\hat{D}_n(\gamma_{(n)})}}{[S_n^0(\gamma_{(n)})]^2}, \tag{27}$$

where $\gamma_{(n)}$ is a condensed notation:

$$\gamma_{(n)}=\{\beta_{(n-1)},\gamma_n\}. \tag{28}$$

Equation (27) is nothing but Eq. (24) where we let remain the last six integrations $d\beta_n$:

$$I_{G_n}=\int d\beta_n\, j_{g_n}(\beta_n)\exp\Big(\sum_{j\in\bar{K}'}s_j\,\beta_n^j\Big)O_{G_{n-1}}(\gamma_n). \tag{29}$$

The open amplitude is only dependent on the $(n-1)$ subgraphs $G_{n-1}=\{g_1,\ldots,g_{n-1}\}$, and not on the last subgraph g_n. From a given open amplitude $O_{G_{n-1}}$, it is possible, using (29), to reconstruct all the amplitudes of the family of graphs G_n which have the sames $(n-1)$ first subgraphs and a different nth subgraph g_n. Such graphs, which are generalized ladder with n rungs, but with only the $(n-1)$ first subgraphs G_{n-1} specified, will be called n-open graphs (see Fig. 3).

The integration (29) can be simplified: Inside the set of the six variables β_n, the three integrations $d\bar{\gamma}_n$ can be performed:

$$I_{G_n}=\int d\gamma_n\,\bar{j}_{g_n}(\gamma_n)O_{G_{n-1}}(\gamma_n) \tag{30}$$

with

$$\bar{j}_g(\gamma)=\int d\bar{\gamma}\, j_g(\beta)\exp\Big(\sum_{j\in\bar{K}'}s_j\,\beta^j\Big).$$

Inserting definition (25) for $j_g(\beta)$, we find

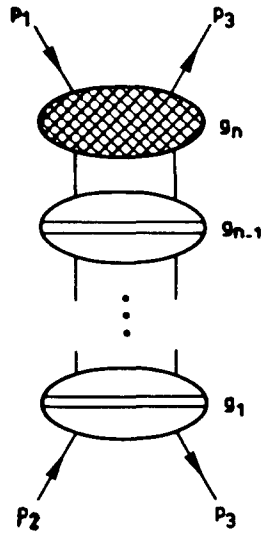

FIG. 3. The n-open graph $G_{n-1}=\{g_1,\ldots,g_{n-1}\}$. An n-open graph, and the open amplitude which is associated with it, depend on the $(n-1)$ subgraphs g_i, $i=1,\ldots,n-1$, but not on the nth subgraph. The nth bubble is a skeleton which may be dressed by any graph g_n.

$$\bar{j}_g(\gamma)=\Theta_3(\gamma)\int d\mu_g(\alpha_g)\exp\Big(\sum_{j\in\bar{K}'}s_j\,\beta_g^j(\alpha_g)\Big)$$
$$\times\prod_{j\in\bar{K}'}\delta(\beta^j-\beta_g^j(\alpha_g)). \tag{31}$$

The way to build the recurrence relation on the number of subgraphs of the open amplitude is straightforward: In the expression $O_{G_n}(\gamma_{n+1})$, we make use of the recurrence relations (14) and (23).

We recognize the open amplitude $O_{G_{n-1}}(\hat{S}_2(\beta_n,\gamma_{n+1}))$ in the integrand and so

$$O_{G_n}(\gamma_{n+1})=\int d\beta_n\, j_{g_n}(\beta_n)\,\frac{e^{d(\bar{\gamma}_n,\gamma_{n+1})}}{[\hat{S}_2^0(\beta_n,\gamma_{n+1})]^2}$$
$$\times O_{G_{n-1}}(\hat{S}_2(\beta_n,\gamma_{n+1})). \tag{32}$$

In the relation (32), only the variables with an index equal to n or $n+1$ appear. Thus the notations can be simplified: instead of $\beta_n=\{\gamma_n,\bar{\gamma}_n\}$ and $\beta_{n+1}=\{\gamma_{n+1},\bar{\gamma}_{n+1}\}$, we will use in the remainder $\beta'=\{\gamma',\bar{\gamma}'\}$ and $\beta=\{\gamma,\bar{\gamma}\}$.

Let us remark that the open amplitude has been defined in perfect analogy with the φ^3 ladder case.[2] We recall that, in this latter work, the closing variables (see the Introduction) were the three Schwinger parameters attached to the last rung and the last vertical lines of the ladder, whose correspondent here is exactly the last subgraph g_n [see Figs. 1(b) and (c)]. One can get convinced that the elements of γ_n concern the same topological polynomials of g_n that the closing variables of the ladder.

B. Summing the series over all graphs

As we already said, we may draw all graphs generated by any scalar Lagrangian as a generalized ladder (see Fig. 1(b)], where each subgraph g_i is t-2PI. Now, the crucial point when one wants to face the whole perturbation is to organize the infinite sum. The results already obtained [the factorization of Eq. (24) on one hand, the definition of the open amplitude on the other] lead us to the following four steps:

(i) For any n-open graph, we define its open amplitude $O_{G_{n-1}}$.

(ii) We group together all the n-open graphs and define the quantity

$$O_n(\gamma) = \sum_{G_{n-1}} O_{G_{n-1}}(\gamma). \tag{33}$$

From Eqs. (32) and (33), we see that $O_n(\gamma)$ verifies the recurrence relation

$$O_n(\gamma) = \int d\beta'\, k(\gamma,\beta')O_{n-1}(\hat{S}_2(\beta',\gamma)), \tag{34}$$

with

$$k(\gamma,\beta') = \sum_g k_g(\gamma,\beta'), \tag{35a}$$

where Σ_g is the sum over all the t-2PI graphs and where

$$k_g(\gamma,\beta') = j_g(\beta')\frac{e^{d(\bar{\gamma}',\gamma)}}{[\hat{S}_2^0(\beta',\gamma)]^2}. \tag{35b}$$

(iii) The following step consists in summing over each such set of graphs; we define

$$O(\gamma) = \sum_{n=1}^{\infty} O_n(\gamma). \tag{36}$$

then $O(\gamma)$ verifies the integral equation

$$O(\gamma) = O_1(\gamma) + \int d\beta'\, k(\gamma,\beta')O(\hat{S}_2(\beta',\gamma)). \tag{37}$$

with

$$O_1(\gamma) = e^{\hat{D}_1(\gamma)} = e^{s_{12}\beta^{12} + s_{22}\beta^{22} + s_2\beta^2}. \tag{38}$$

Equation (37) is essentially the integral equation we are looking for.

(iv) The last step consists of performing the integration on the variables γ in order to get the four-point amplitude I:

$$I = \int d\gamma\, \bar{j}(\gamma)O(\gamma), \tag{39}$$

where

$$\bar{j}(\gamma) = \sum_g \bar{j}_g(\gamma) \tag{40}$$

with $\bar{j}_g$ given by (31).

C. Final form for the integral equation

We will now proceed a little further in order to get the integral equation verified by $O(\gamma)$ in a more classical form, and see whether it falls under the scope of classical theorems.

We define the change of variables

$$\gamma' \to \gamma^* \tag{41}$$

such that

$$\begin{aligned}
&\beta^{12'} \to \beta^{12*} = \hat{S}_2^{12}(\beta',\gamma) = \beta^{12'}\beta^{12}/(\beta^{11'} + \beta^{22}),\\
&\beta^{22'} \to \beta^{22*} = \hat{S}_2^{22}(\beta',\gamma) = \beta^{22'} - (\beta^{22'})^2/(\beta^{11'} + \beta^{22}),\\
&\beta^{2'} \to \beta^{2*} = \hat{S}_2^{2}(\beta',\gamma)\\
&\qquad = \beta^{2'} + \beta^{12'}(\beta^2 - \beta^{1'})/(\beta^{11'} + \beta^{22}).
\end{aligned} \tag{42}$$

This change of variables does not concern the variables $\bar{\gamma}'$.

We define

$$u = \beta^{12*}/\beta^{12} = \beta^{12'}/(\beta^{11'} + \beta^{22}). \tag{43}$$

One can immediately see that the variation domain of γ^* is at most as large as the domain defined by $\Theta_3(\gamma^*)$, as γ^* is nothing but the γ variables of the graph of Fig. 2 with $\{g_1, g_2\} = \{g', g\}$.

The computation of the actual variation domain of γ^* is given in Appendix A. It is given by the following function:

$$\Theta_4(\gamma,\gamma^*) = \theta(1 - |u|)\theta(\beta^{22*} - |u|\beta^{22} - 2|\beta^{2*} - u\beta^2|). \tag{44}$$

The Jacobian of the transformation (41) is

$$J(\gamma' \to \gamma^*) = (\beta^{11'} + \beta^{22})/\beta^{12}. \tag{45}$$

Among the six integrations of the integral (37), three can be done explicitly and a new kernel K can be defined by

$$K(\gamma,\gamma^*) = \sum_g K_g(\gamma,\gamma^*) \tag{46}$$

with

$$K_g(\gamma,\gamma^*) = \int d\bar{\gamma}'\, k_g(\gamma,\beta')J(\gamma' \to \gamma^*).$$

Using (35b) and the expression (25) of j_g, one obtains

$$\begin{aligned}
K_g(\gamma,\gamma^*) = \Theta_4(\gamma,\gamma^*) &\int d\nu_g(\alpha_g)\exp[d(\delta\bar{\gamma}(\alpha_g),\gamma)]\\
&\times \prod_{j\in K'} \delta(\beta^{j*} - \hat{S}_2^j(\beta_g(\alpha_g),\gamma))
\end{aligned} \tag{47}$$

with

$$d\nu_g(\alpha_g) = \frac{d\mu_g(\alpha_g)}{[\hat{S}_2^0(\beta(\alpha_g),\gamma)]^2} = \frac{d\mu_g(\alpha_g)}{[\delta^{11}(\alpha_g) + \delta^{22}]^2}.$$

We finally have

$$O(\gamma) = O_1(\gamma) + \int d\gamma^*\, K(\gamma,\gamma^*)O(\gamma^*), \tag{48}$$

with $O_1(\gamma)$ given by (38) and $K(\gamma,\gamma^*)$ by Eqs. (46) and (47). Of course, we obtain the amplitude 1 from Eq. (39).

Let us make three last remarks about the IE:

—The dependence of $O(\gamma)$ as function of $s_j, j\in K$, has two sources: $O_1(\gamma)$ depends upon s_{12}, s_{22}, and s_2, and the kernel K depends on the three other invariants s_{11}, s_1, and s_t.

—Whereas the number of integration variables was six in the IE (37), it is only three in (48). This difference reflects exactly the difference between the recurrence relation verified by S_n and D_n and which concerns six variables [see (13)], and the one verified by $\hat{D}_n$, where only the three variables γ_n are concerned.

—The inhomogeneous term $O_1(\gamma)$ is a simple explicit function [Eq. (38)] which is independent of the Lagrangian.

III. MELLIN TRANSFORM AND REGGE POLES

A. The integral equation verified by the open amplitude of the Mellin transform

The reasons for working with the Mellin transform of the amplitudes are of two different types:

—First, there are technical reasons which are linked to the Wick rotation problem and to the Landau singularities. These points have been discussed in Ref. 2 for the φ^3 ladder case, and we shall not come back to it in the present paper.

—On the other hand, it is well known that the Mellin

transform is very well adapted for the study of the amplitude at high energy, where the Regge model is relevant. The singularities of the Mellin transform are linked to the Regge singularities in the angular momentum space.

In term of the invariants $s_j, j\in K$, the Regge limit is defined by

$$s_{12}\to\infty, \quad s_j = cst \quad \text{for } j\neq 12;$$

so we are going to perform the Mellin transform of the amplitude for the variable s_{12}. The Mellin transform $\bar f(x)$ of a function $f(s)$, which is integrable and regular when s goes to zero, is defined by

$$e^{-i\pi x}\Gamma(-x)\bar f(x) = \int_0^\infty ds\, s^{-x-1} f(s)$$

$$\text{for } -1 < \mathrm{Re}(x) < 0. \tag{49}$$

For the values of x where the integral (49) does not exist, $\bar f(x)$ can be defined by analytic continuation. If one uses the β-representation (24) of I_G, it is possible to perform the integration (49) over the variable s_{12} explicitly, and one obtains

$$\bar I_G(x) = \int \prod_{i=1}^{n} [d\beta_i\, j_{g_i}(\beta_i)] \times \frac{[S_n^{12}(\beta_1,\ldots,\beta_n)]^x}{[S_n^0(\beta_1,\ldots,\beta_n)]^2} \times \exp\left(\sum_{\substack{j\in K\\ j\neq 12}} s_j S_n^j(\beta_1,\ldots,\beta_n)\right).$$

Using the relations (13) and (16), it can be shown that

$$S_n^{12}(\beta_1,\ldots,\beta_n) = \frac{\Pi_{i=1}^n \beta_i^{12}}{S_n^0(\beta_1,\ldots,\beta_n)};$$

the Mellin transform $\bar I_G(x)$ becomes

$$\bar I_G(x) = \int \prod_{i=1}^{n} [d\beta_i\, j_{g_i}(\beta_i)(\beta_i^{12})^x] \times \frac{1}{[S_n^0(\beta_1,\ldots,\beta_n)]^{x+2}} \exp\left(\sum_{\substack{j\in K\\ j\neq 12}} s_j S_n^j(\beta_1,\ldots,\beta_n)\right). \tag{50}$$

The factors $(\beta_i^{12})^x$ and $[S_n^0(\beta_1,\ldots,\beta_n)]^{-x}$ introduce no new singularity in the integral when $x > -1$, and $\bar I_G(x)$ is defined when I_G is defined.

The open amplitude of the Mellin transform is a function of three variables $\gamma = (\gamma^{12},\gamma^{22},\gamma^2)$ defined by suppressing in (50) the last integration $d\beta_n$ and the factor $j_{g_n}(\beta_n)(\beta_n^{12})^x \exp(\Sigma_{j\in\bar K} s_j \beta_n^j)$, which depends only on the variables β_n:

$$\bar O_{G_{n-1}}(\gamma_n) = \int \prod_{i=1}^{n} [d\beta_i\, j_{g_i}(\beta_i)(\beta_i^{12})^x] \times \frac{1}{[S_n^0(\beta_1,\ldots,\beta_n)]^{x+2}} \times \exp\left(\sum_{\substack{i\in K\\ i\neq 12}} s_i \bar S_n^i(\beta_1,\ldots,\beta_n)\right).$$

The important property of the function $\bar O_{G_n}(\gamma)$ is the fact that it follows nearly the same recurrence relation than $O_{G_n}(\gamma)$. The only difference comes from a factor

$$[\beta_n^{12}/S_2^0(\beta_n,\gamma)]^x = (\beta^{12*}/\beta^{12})^x$$

which appears in the kernel

$$\bar O_{G_n}(\gamma) = \int d\beta_n\, j_{g_n}(\beta_n)\left(\frac{\beta^{12*}}{\beta^{12}}\right)^x \times \frac{\exp[d(\bar\gamma_n,\gamma)]}{[S_n^0(\beta_1,\ldots,\beta_n)]^{x+2}} \bar O_{G_{n-1}}(\hat S_2(\beta_n,\gamma)).$$

Then, in the same manner as for the function $O(\gamma)$ [see Eq. (48)], one can show that the sum $\bar O$ of all the open amplitudes of the Mellin transforms verifies an IE:

$$\bar O(\gamma) = \bar O_1(\gamma) + \int d\gamma^*\, \bar K(\gamma,\gamma^*) O(\gamma^*), \tag{51}$$

with

$$\bar K(\gamma,\gamma^*) = (\beta^{12*}/\beta^{12})^x K(\gamma,\gamma^*). \tag{52}$$

Due to the factor $(S_n^{12})^x$, or $(\beta_i^{12})^x$, all the expressions derived here would be well defined only if S_n^{12} or β_i^{12} would never become negative. It is not true in general [see Eqs. (7) and (11)].

So it is necessary to replace $(S_n^{12})^x$ by

$$(S_n^{12})^x \to \theta(S_n^{12})(S_n^{12})^x + \theta(-S_n^{12}) e^{i\pi x}(-S_n^{12})^x,$$

and similarly for $(\beta_i^{12})^x$. It is known that the step functions θ are the origin of the Mandelstam cut.

B. Particular value of γ: $\beta^{12} = 0$

Usually, when an IE is written for a particular values of a variable, the number of integration variables does not vary. Here, if we put $\beta^{12} = 0$, then

$$S_2^{12}(\beta',\beta) = 0,$$

the interval of integration for the variable β^{12*} disappears and the IE becomes a two-variable IE. Actually it is not possible to put $\beta^{12} = 0$ directly in the IE of Eq. (51) because the Jacobian $J(\gamma'\to\gamma^*)$ becomes infinite and one must come back to Eq. (37). The change of variables $\gamma'\to\gamma^*$, being not allowed when $\beta^{12} = 0$, we replace it by the change $\gamma'\to(\beta^{22*},\beta^{2*},u)$ with

$$u = \hat S_2^2(\beta',\gamma)/\beta^{12} = \beta^{12\prime}/(\beta^{11\prime} + \beta^{22}). \tag{53}$$

For any value of β^{12}, the IE can be written as

$$\bar O(\gamma) = \bar O_1(\gamma) + \int d\beta^{22*}\, d\beta^{2*}\, du \times \bar L(\gamma,\beta^{22*},\beta^{2*},u)\bar O(\beta^{22*},\beta^{2*},u\beta^{12}) \tag{54}$$

with

$$\bar L(\gamma,\beta^{22*},\beta^{2*},u) = \Theta_4(\beta^{22},\beta^2,\beta^{22*},\beta^{2*},u)u^x \int dv_g \exp(d) \times \prod_{j=22,2} \delta(\beta^{j*} - \hat S_2^j)\delta\left(u - \frac{\beta^{12}(\alpha)}{\hat S_2^0}\right), \tag{55}$$

where d is defined by (23b).

It is now possible to put $\beta^{12} = 0$ in the previous equation, and we find

$$\bar{O}^0(\beta^{22},\beta^2)=\bar{O}_1^0(\beta^{22},\beta^2)+\int d\beta^{22*}\,d\beta^{2*}\times\bar{K}(\beta^{22},\beta^2;\beta^{22*},\beta^{2*})\bar{O}^0(\beta^{22*},\beta^{2*}) \tag{56}$$

with $\bar{K}^0=\Sigma\bar{K}_g^0$ and

$$\bar{K}_g^0(\beta^{22},\beta^2,\beta^{22*},\beta^{2*})=\int d\nu_g(\alpha_g)\left(\frac{\beta_g^{12}(\alpha_g)}{\hat{S}_2^0(\beta(\alpha),\gamma)}\right)^x\times\exp\left[s_t\left(\beta^t(\alpha)-\frac{(\beta^2-\beta^1(\alpha))^2}{\beta^{11}(\alpha)+\beta^{22}}\right)\right]\times\prod_{j=22,2}\delta(\beta^{j*}-\hat{S}_2^j(\beta(\alpha),\beta^{22},\beta^2)) \tag{57}$$

and $\bar{O}^0(\beta^{22},\beta^2)=\bar{O}(\beta^{12}=0,\beta^{22},\beta^2)$.

When $\beta^{12}=0$, two invariants s_{11} and s_1 disappear in the expression of the IE. The solution $\bar{O}^0$ depends on the three remaining invariants s_{22}, s_2, and s_t and on the Mellin variable x. The kernel itself depends only on $s_t=t$ and x.

C. Expansion of the IE. Leading Regge poles

In Ref. 2, this reduction of the number of integration variables, when $\beta^{12}=0$, was the basis of a method of computing the amplitudes and its singularities by means of an expansion, the first term of which is precisely the function $\bar{O}_0(\beta^{22},\beta^2)$. Each term of this expansion was the solution of a Fredholm type IE, and so its singularities were given by the annulation of the determinant of the kernel. This expansion classifies the singularities, which give the Regge singularities of the amplitude, in a simple manner: Only the first term of the expansion contributes to the leading Regge pole, only the two first terms contribute to the subleading poles, and so on... . In the general case we study here, it is again possible to perform such an expansion which has the same formal structure. Of course, the nature of the kernel depends on the Lagrangian, and on the particular graphs one actually keeps in the kernel. For the complete perturbation it will be difficult to verify if we are or not in the Fredholm case.

Let us expand the function $\bar{O}(\beta^{12},\beta^{22},\beta^2)$ as a series of β^{12}:

$$\bar{O}(\beta^{12},\beta^{22},\beta^2)=\sum_n\bar{O}^n(\beta^{22},\beta^2)(\beta^{12})^n/n! \tag{58}$$

Using Eq. (51), it can be easily verify that each function $\bar{O}^n$ is the solution of an IE with a kernel $\bar{K}^n=\Sigma_g\,\bar{K}_g^n$, where $\bar{K}_g^n$ is equal to $\bar{K}_g^0$, with x replaced by $x+n$.

If we note explicitly the dependence of these kernels, they verify

$$\bar{K}^n(x)=\bar{K}^0(x+n). \tag{59}$$

If the kernel $\bar{K}_n^x$ are of the Fredholm type (bounded, squared-integrable, kernel of a compact operator,...), the relation (59) shows that all the Regge poles are given by the first kernel $\bar{K}^0(x)$. The poles coming from the other kernels $\bar{K}^n(x)$ are obtained by a simple translation: $x\to x-n$. In the φ^3 ladder case, the kernel $\bar{K}^0(x)$ are not of the Fredholm type. The expansion (58) must be slightly modified in order to obtain Fredholm type kernels, and the degenerescence of the daughter spectrum is lost (the exact degenerescence is true only in the limit $\lambda\to 0$).

As the kernels $\bar{K}^n(x)$ depend only on $s_t=t$ and x and, of course, on the coupling constant λ, the Regge poles, when they exist, depend only on t and λ. We recover here the well-known property that the Regge poles are independent of external squared four momenta p_i^2.

IV. PARTICULAR CASES

A. Particular values of the invariants

When some of the invariants are equal to zero, the structure of the integral equations changes: The number of integration variables is reduced from three to two, and even only one in one case.

1. Forward elastic scattering

The elastic scattering in the forward direction is defined by

$$p_1^2=p_3^2,\quad p_2^2=p_4^2,\quad t=0,$$

in term of the Mandelstam invariants or by

$$s_1=s_2=s_t=0$$

in term of the s_j variables [Eq. (5)].

The kernel K of Eq. (48) depends on β_2 and β_2^* only through the combination $\beta^{2*}-u\beta^2$. In particular, one of the three δ functions contains this combination. Thus, if one integrates the kernel with a function $f(\beta^{12},\beta^{22})$ which is independent of β^2, the result is also independent of β^2:

$$\int d\gamma^*\,K(\gamma,\gamma^*)f(\beta^{12*},\beta^{22*})=\int d\beta^{12*}d\beta^{22*}\,k(\beta^{12},\beta^{22};\beta^{12*},\beta^{22*})f(\beta^{12*},\beta^{22*})$$

with

$$k=\sum_g k_g$$

and

$$k_g(\beta^{12},\beta^{22};\beta^{12*},\beta^{22*})=\theta(1-|u|)\theta(\beta^{22*}-|u|\beta^{22}-2|\beta^2(\alpha_g)-u\beta^1(\alpha_g)|)\times\int d\nu_g\exp\left(-s_{11}\frac{(\beta^{12})^2}{\beta^{12}(\alpha_g)+\beta^{22}}\right)\times\prod_{j=12,22}\delta(\beta^{j*}-S_2^j(\beta(\alpha_g);\beta^{12},\beta^{22})). \tag{60}$$

So, since the first term O_1 does not depend on β^2 when s_2 is equal to zero, $O_2,O_3,\ldots,O_n$, and thus their sum O does not depend on β^2. This last function is a function of only two variables, $O=O(\beta^{12},\beta^{22})$ and it verifies an IE, the kernel of which is $k(\beta^{12},\beta^{22};\beta^{12*},\beta^{22*})$.

The reduction from three to two of the number of integration variables is a consequence of the well-known result[6] that in the equal mass case and at $t=0$ the BS IE have a supplementary symmetry.

2. Threshold in the t channel

The annulation of the three invariants s_{11}, s_{12}, and s_1 corresponds to the threshold in the t channel:

$$p_1 = -p_3 \quad \text{or} \quad p_1^2 = p_3^2 = \tfrac{1}{4}t \quad \text{and} \quad s = u.$$

When $s_{11} = s_{12} = s_1 = 0$, it can be shown, in the same manner as in the previous subsection, that the amplitude verifies an IE of only two variables β^{22} and β^2. The first term is

$$O_1 = O_1(\beta^{22}, \beta^2) = \exp(s_{22}\beta^{22} + s_2\beta^2),$$

and the kernel corresponding to the graph g becomes

$$k_g(\beta^{22}, \beta^2; \beta^{22*}, \beta^{2*}) = \int d\nu_g(\alpha_g) \exp\left[s_t\left(\beta^t(\alpha_g) - \frac{\beta^2 - \beta^1(\alpha_g)}{\beta^{11}(\alpha_g) + \beta^{22}}\right)\right] \times \prod_{j=22,2} \delta(\beta^{j*} - \hat{S}_2^j(\beta(\alpha_g); \beta^{22}, \beta^2)) \times \theta(\beta^{22*} - |u(\alpha)|\beta^{22} - 2|\beta^{2*} - u(\alpha)\beta^2|) \tag{61}$$

with

$$u(\alpha) = \beta^{12}(\alpha)/[\beta^{11}(\alpha) + \beta^{22}].$$

3. Elastic scattering with some external momentum equal to zero

Here we consider the case where

$$p_1 = p_3 = 0$$

or

$$p_1^2 = p_3^2 = t = 0$$

and

$$s = u = p_2^2 = p_4^2.$$

This case contains, as an even more particular case, the scattering when all the momentum and all the invariants are equal to zero:

$$p_i = 0, \quad i = 1,2,3,4.$$

The simplifications of the two previous paragraphs can be done together, and one obtains an IE of only one variable β^{22}, with a kernel which is given by $k = \Sigma_g k_g$ and

$$k_g(\beta^{22}, \beta^{2*}) = \int d\nu_g \, \delta(\beta^{22*} - \hat{S}_2^{22}(\beta(\alpha_g); \beta^{22})) \times \theta(\beta^{22*} - |u(\alpha)|\beta^{22} - 2|\beta^2(\alpha) - u(\alpha)\beta^1(\alpha)|). \tag{62}$$

4. Bound states

It is possible to give another interesting interpretation of the cases studied in subsections 2 and 3. Using, for example, the Schwinger representation of the graphs, one can see that the vertex function of a bound state (the "relativistic wavefunction"), of squared mass t and which contains two particles of momentum p_2 and p_4, has exactly the same structure in the invariant $s_t = t$, $s_2 = \frac{1}{2}(p_4^2 - p_2^2)$, and $s_{22} = \frac{1}{2}(p_2^2 + p_4^2) - \frac{1}{4}t$ as the $2 \to 2$ amplitudes in the t channel when $p_1 + p_3 = 0$. Thus one can define on "open" relativistic wavefunction $\varphi(\beta^{22}, \beta^2)$ which verifies a homogeneous IE:

$$\varphi(\beta^{22}, \beta^2) = \int d\beta^{22*} d\beta^{2*} \, k(\beta^{22}, \beta^2; \beta^{22*}, \beta^{2*}) \varphi(\beta^{22*}, \beta^{2*}). \tag{63}$$

The wavefunction itself can be computed by integrating the open wavefunction φ over β^{22} and β^2.

Similarly, if now one considers a bound state of mass equal to zero ($t = 0$), its "open" relativistic wavefunction $\varphi(\beta)$ verifies a homogeneous IE of one variable:

$$\varphi(\beta^{22}) = \int d\beta^{22*} \, k(\beta^{22}; \beta^{22*}) \varphi(\beta^{22*}). \tag{64}$$

B. Particular value of γ: $\beta^{12} = 0$

It has been seen in the previous section that the IE verified by the Mellin transform became simpler when β^{12} was equal to zero. It is also the case for the IE (48). The amplitudes $O^0(\beta^{22}, \beta^2) = O(\beta^{12} = 0, \beta^{22}, \beta^2)$ verifies an IE with a kernel $K^0 = \Sigma_g K_g^0$ defined by

$$K_g^0(\beta^{22}, \beta^2; \beta^{22*}, \beta^{2*}) = \int d\nu_g(\alpha) \exp\left[s_t\left(\beta^t(\alpha) - \frac{(\beta^2 - \beta^1(\alpha))^2}{\beta^{11}(\alpha) + \beta^{22}}\right)\right] \times \prod_{j=22,2} \delta(\beta^{2*} - \hat{S}_2^j(\gamma(\alpha), \beta^{22}, \beta^2)). \tag{65}$$

When $\beta^{12} = 0$, three invariants, s_{12}, s_{11}, and s_1, disappear in the expression of the IE. The solution depends only on the three remaining invariants s_{22}, s_2, and s_t (the kernel depends on s_t and the first term on s_{22} and s_2).

C. Particular class of graphs

In this section, we consider a particular class of graphs: the generalized rung ladder graph (GRLG), where the rungs are made with subgraphs which are linked to each upright by only one vertex (see Fig. 4, where different examples of such generalized rungs are given). To this class belongs the ladder graph of φ^3, which has been already widely studied in our previous paper.[2] When one considers the GRLG, four among the seven topological polynomials [Eq. (3)] of the rungs are equal to zero,

$$A^t = A^u = A^1 = A^3 = 0, \tag{66}$$

and the formalism which has been worked up in the first sections become simpler: All the kernels [Eqs. (25), (31), (35), and (46)] depend on the graph by the same function $\tilde{j}_g$ of only one variable, and, except for this function, they are explicit functions. In the φ^3 ladder graph case, the function $\tilde{j}_g$ is a constant, and, of course, we find the IE of our previous paper again.

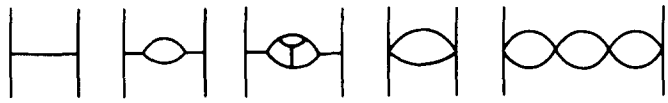

FIG. 4. Some example of generalized rungs. The three first graphs come from a φ^3 Lagrangian; the two last ones, from a φ^4 Lagrangian.

In this subsection we note g the generalized rung itself and $\bar{g}$ the same graph with two vertical lines added (see Fig. 5). If α and α' are the Schwinger parameters attached to these two lines, the measure $d\mu_{\bar{g}}$ becomes

$$d\mu_{\bar{g}}(\alpha_{\bar{g}}) = d\alpha\, d\alpha' \exp[-(\alpha+\alpha')m^2]\, d\mu_g(\alpha_g).$$

In addition to the relations (66), the particular structure of the graphs lead to simple expressions for the other topological polynomials of $\bar{g}$:

$$P_{\bar{g}} = P_g, \quad A^{\,2}_{\bar{g}} = \alpha A^{\,2}_g,$$
$$A^{\,4}_{\bar{g}} = \alpha' A^{\,4}_g, \quad A^{\,s}_{\bar{g}} = A^{\,s}_g.$$

If one computes the $\beta^{\,j}_{\bar{g}}$ functions, one obtains

$$\begin{aligned}
&\beta^{\,1}_{\bar{g}}(\alpha_{\bar{g}}) = 0,\\
&\beta^{\,11}_{\bar{g}}(\alpha_{\bar{g}}) = \beta^{\,12}_{\bar{g}}(\alpha_{\bar{g}}) = \beta^{\,12}_{g}(\alpha_g),\\
&\beta^{\,22}_{\bar{g}}(\alpha_{\bar{g}}) = \alpha+\alpha'+\beta^{\,12}_{g}(\alpha_g),\\
&\beta^{\,2}_{\bar{g}}(\alpha_{\bar{g}}) = \tfrac{1}{2}(\alpha'-\alpha), \quad \beta^{\,t}_{\bar{g}}(\alpha_{\bar{g}}) = \tfrac{1}{4}(\alpha+\alpha').
\end{aligned} \tag{67}$$

We are not going to transform all the results of the previous sections, but only the main ones.

Taking into account the relations (67), we can give the new expression of the kernel $j_{\bar{g}}(\beta)$ [see Eq. (25)]

$$j_{\bar{g}}(\beta) = \delta(\beta^{\,1})\delta(\beta^{\,11}-\beta^{\,12})\delta(\beta^{\,t}-\tfrac{1}{4}(\beta^{\,22}-\beta^{\,12})) \times \exp(-\beta^{\,22}m^2)\tilde{j}_g(\beta^{\,12}),$$

where the new function $\tilde{j}_g$ is defined by

$$\tilde{j}_g(\beta^{\,12}) = \theta(\beta^{\,12})\exp(\beta^{\,12}m^2)\int d\mu_g\,\delta(\beta^{\,22}-\beta^{\,12}(\alpha_g)). \tag{68}$$

In the particular case of the φ^3 ladder graphs, the function $\tilde{j}_g$ is a constant:

$$\tilde{j}_g(\beta^{\,12}) = \lambda^{\,2}.$$

The β-representation [Eq. (24)] can be written

$$I_G = \int \prod_{i=1}^{n} [d\beta^{\,12}_i \tilde{j}_{g_i}(\beta^{\,12}_i)]\, Q_n(\beta^{\,12}_{(n)}), \tag{69}$$

where Q_n is an explicit function of n variables, $\beta^{\,12}_{(n)} = (\beta^{\,12}_1, \beta^{\,12}_2, \ldots, \beta^{\,12}_n)$, independent of the graph and equal to

$$Q_n(\beta^{\,12}_{(n)}) = \int \prod_{i=1}^{n} [d\beta^{\,22}_i\, d\beta^{\,2}_i \exp(-\beta^{\,22}_i m^2)] \times \frac{\exp[D_n(\beta_1,\ldots,\beta_n)]}{[S^0_n(\beta_1,\ldots,\beta_n)]^2}\Bigg|_{\substack{\beta^{\,1}_i=0;\ \beta^{\,11}_i=\beta^{\,12}_i;\\ \beta^{\,t}_i=(\beta^{\,22}_i-\beta^{\,12}_i)/4;\\ i=1,\ldots,n.}}$$

For example, the β-representation of dimension 1 is

$$I_G = \int d\beta^{\,12}\, \tilde{j}_G(\beta^{\,12}) Q_1(\beta^{\,12}),$$

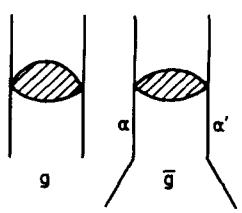

FIG. 5. Definition of g and $\bar{g}$. In subsection IV C, g is the rung itself and $\bar{g}$ is the rung to which two vertical lines have been added. α and α' are the Schwinger parameters attached to these two additional lines.

with

$$Q_1(\beta^{\,12}) = \frac{\exp[-\beta^{\,12}(m^2-s_{11}-s_{12}-s_{22})]}{(m^2-s_{22}-\frac{1}{4}s_t)^2-(\frac{1}{2}s_2)^2} = \frac{\exp[-(m^2-s)\beta^{\,12}]}{(m^2-p_2^2)(m^2-p_4^2)}.$$

The denominator of Q_1 represents the propagators of the two extra lines which have been added to the graph.

Let us now come to the integral equation itself [Eq. (48)]. The product of the three δ functions in Eq. (47) can be written as

$$[\beta^{\,22}/\beta^{\,12}(1-u)^2]\delta(\beta^{\,12}_g(\alpha_g) - [u/(1-u)]\beta^{\,22}) \times \delta(\alpha+\alpha'-\beta^{\,22*}+u\beta^{\,22})\delta(\tfrac{1}{2}(\alpha'-\alpha)-\beta^{\,2*}+u\beta^{\,2}). \tag{70}$$

In the definition of K_g [Eq. (47)], the integrations over the two variables α and α' can be done, using the two last δ functions. The remaining integrations of the first δ function give the $\tilde{j}_g$ function with an argument equal to $[u/(1-u)]\beta^{\,22}$. As $\beta^{\,12}$ and thus u are always positive variables, the Θ_4 function becomes simpler:

$$\Theta_4(\gamma,\gamma^*) = \theta(U-u)$$

with

$$U = \inf\left(1, \frac{\beta^{\,22*}-\beta^{\,2*}}{\beta^{\,22}-\beta^{\,2}}, \frac{\beta^{\,22*}+\beta^{\,2*}}{\beta^{\,22}+\beta^{\,2}}\right).$$

Finally we obtain

$$K_g(\gamma,\gamma^*) = \frac{1}{\beta^{\,12}\beta^{\,22}}\exp\left(-\beta^{\,22*}m^2 - \frac{u^2}{1-u}\beta^{\,22}m^2\right) \times \exp(d)\,\tilde{j}_g\left(\frac{u}{1-u}\beta^{\,22}\right) \times \theta(U(\beta^{\,22},\beta^{\,2},\beta^{\,22*},\beta^{\,2*})-u)$$

with

$$d = -[s_{11}(\beta^{\,12})^2 + s_1\beta^{\,12}\beta^{\,2} + s_t(\beta^{\,2})^2](1-u)/\beta^{\,22} + s_t(\beta^{\,22*}-u\beta^{\,22})/4.$$

In order to verify that, in the φ^3 ladder case ($\tilde{j}_g = \lambda^{\,2}$), this kernel is actually identical to the one of Ref. 2, two changes must be done. First we must perform the change of variables $\beta^{\,12}, \beta^{\,2}, \beta^{\,22} \to \alpha, \alpha', \beta$, defined by the relations

$$\begin{aligned}
&\alpha = \tfrac{1}{2}(\beta^{\,22}-\beta^{\,12}) - \beta^{\,2},\\
&\alpha' = \tfrac{1}{2}(\beta^{\,22}-\beta^{\,12}) + \beta^{\,2},\\
&\beta = \beta^{\,12}.
\end{aligned}$$

The other change comes from the different normalization of the amplitudes. Here the pole term is $O_1(\gamma)$ [see Eq. (38)] when in Ref. 2 it would be defined by

$$F_1(\alpha,\alpha',\beta) = \exp(s\beta + p_2^2\alpha + p_4^2\alpha').$$

V. RENORMALIZATION: φ^3 INTERACTION LAGRANGIAN CASE

As soon as some graphs of the theory are divergent, we have to take into account the renormalization operator R. We do this here only for the most simple case, namely the φ^3 Lagrangian case.

The general definition of the renormalization operator can be found in Refs. 1(d) or 5.

When we restrict ourself to φ^3 interaction, there is only one connected divergent subgraph in the theory (see Fig. 6).[1(d)] Such subgraphs are logarithmically divergent, and they are always disjoint. The renormalization operator for a given graph G reduces to

$$R^G = \prod_{g^d} (1 - \tau_{g^d}^{-4}) \tag{71}$$

where g^d are all the connected divergent subgraphs described by Fig. 6 of G and $\tau_{g^d}^{-4}$ is the generalized subtraction Taylor operator. If the graph G is partitioned into a set of subgraphs $g_i : G = (g_1, g_2, \ldots, g_n)$, for example, if one considers the Bethe–Salpeter structure of G [see Fig. 1(b)], then R^G appears as a product of renormalization operators, each acting on a given subgraph g_i:

$$R^G = \prod_{i=1}^{n} R^{g_i}. \tag{72}$$

It is this property which makes easy the demonstration of the compatibility of the renormalization and of the β-representation of the Feynman integral. Before going on, we give the expression of R^{g^d} for a simple divergent graph. The operator R^{g^d} is an operator which acts on a function $f(\alpha,\alpha')$ which depends on the two Schwinger parameters of the graph g^d (see Fig. 6). Putting $\lambda = 4$ in Eqs. (I.9) and (I.10) of Ref. 1(d) and using the integral representation of the Taylor remainder [see, for example, Eq. (III.15) of Ref. 2], R^{g^d} can be written as

$$\begin{aligned} R^{g^d} f(\alpha,\alpha') &= \int_0^1 \frac{dg^*(u)}{du}\, du \\ &= f(\alpha,\alpha') - \lim_{u\to 0} [u^4 f(\alpha u^2, \alpha' u^2)], \end{aligned} \tag{73}$$

where $g(u)$ is defined by

$$g(u) = u^4 f(\alpha u^2, \alpha' u^2).$$

The generalization to the case of several divergent subgraphs is straightforward, but we are not going to write it because the only property we need is actually the factorization property of Eq. (72).

In the Euclidian space, the amplitude I_G of a graph G is [see Eq. (1)]

$$I_G = \int d\lambda_G(\alpha_G)\, R^G \left(\frac{e^{D_G(\alpha_G)}}{[P_G(\alpha_G)]^2} \right),$$

with

$$d\lambda_G(\alpha_G) = P_G^2(\alpha_G)\, d\mu(\alpha_G).$$

The functions D_G and P_G verify the structure property of Eqs. (12) and (21). Then, using (26), one obtains

$$\frac{e^{D_G(\alpha_G)}}{[P_G(\alpha_G)]^2} = \frac{1}{[\Pi_{i=1}^n P_{g_i}(\alpha_{g_i})]^2} \int \prod_{i=1}^{n} \prod_{j\in K} d\beta_i^j \times \delta(\beta_i^j - \beta_{g_i}^j(\alpha_{g_i})) \frac{e^{D_n(\beta_{(n)})}}{[P_n(\beta_{(n)})]^2}.$$

If one commutes the integration on the variable β_i^j and the renormalization operator R^G, then one finds that the β-representation (24) remains unchanged, except that $j_g(\beta)$ needs to be renormalized and becomes

$$j_g^R(\beta) = \int d\lambda_g(\alpha_g)\, R^g \left(\frac{\Pi_{j\in K}\, \delta(\beta^j - \beta_g^j(\alpha_g))}{[P_g(\alpha_g)]^2} \right). \tag{74}$$

One sees now a supplementary advantage of the β-representation: the different singularities of the Schwinger representation are disconnected:

—The UV divergences appear only in the α_g integration which are contained in the expression of j_g.

—The Landau singularities can come only from the β integration because only the function $D_n(\beta_{(n)})$ depends on the Lorentz invariants s_i.

VI. CONCLUSION

In the present paper, it has been shown that the Schwinger parameter formalism, could be modified in such a way that the Bethe-Salpeter structure of the amplitude becomes explicit. This is done through the introduction of a new scalar representation of the Feynman amplitudes, the β-representation [Eq. (24)]. The fundamental feature of this β-representation is the quasifactorization property of Theorem 3. Reflecting the generalized ladder structure of the graphs, the β-representation naturally exhibits a recurrence law in the number of "rungs" [Eq. (32)]. We are then able to build the infinite sum of the "open amplitudes" as the solution of a three-variable integral equation [Eq. (48)]. The last step to obtain the four-point amplitude is to perform the closing integration [Eq. (39)].

We conclude and indicate the next steps that this program should follow. The treatment of the renormalization is, of course, one of them. It has been shown in the framework of asymptotic behavior studies[1(e)] that in the case of a strictly renormalizable theory (such as φ^3 in dimension six or φ^4 in dimension four) the renormalization procedure can be split into two steps: On the one hand, the divergent subgraphs occuring inside the t-2PI subgraphs have to be subtracted: a behavior predicted by the renormalization group is thus generated for the infinite sum of graphs building each "rung" of the generalized ladder; on the other hand, UV divergences arising from the ladder structure itself have to be treated. Obviously we have to look for such a two-step treatment within our framework. Already, for the φ^3 Lagrangian, we have shown (Sec. V) that the R operator respects the factorization property of Theorem 3 (see Sec. I D).

The next point of our program after renormalization has to do with the fact that the actual properties of the solution of our integral equation, of course, depends on the analytic structure of the kernel [the inhomogeneous term is explicit; see Eq. (38)]. This structure is not known in general for the complete perturbative expansion of the kernel.

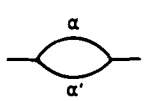

FIG. 6. The only connect divergent subgraph of φ^3.

However, our approach allows to reach many exact results even in cases where infinite subseries of the perturbation series are kept: The structure of the kernel is actually entirely explicit whenever it is restricted to a finite sum. It is then possible to classify the cases where global theorems (such as Fredholm theorems) may be used: quantitative work, such as in the φ^3 ladder case,[2] can be done.

In this paper we have paid attention essentially to the four-point amplitude. As outlined in Sec. IV, it is possible to exhibit an analogous integral equation for the three-point amplitude (vertex). This can also be obtained for the propagator.

Let us end this conclusion by a remark concerning the contested interest of the study of the φ^3 ladder subseries presented in Ref. 2. The results we have obtained in the present paper, taking into account the whole perturbation series, indeed show that essential properties of the perturbation series are already present in the ladder. For example, as in the ladder case, we find a three-variable integral equation and this equation happens to be simpler under the same circumstances (reduction of the number of variables in various particular cases). Also, the β^{12} expansion, the analog of the γ expansion for the ladder case, allows us to classify the singularities in the Mellin space. We even obtain a complete analogy between the ladder and the "generalized rung ladder" (see Sec. IV C and Fig. 4).

As a last statement, we want to stress the importance of the kind of factorization property of a Feynman amplitude into a "skeleton," which exhibits its BS structure and contains its external momentum dependence, and a "dressing," which carries the whole information concerning the dynamics attached to the interaction Lagrangian.

APPENDIX: VARIATION DOMAINS

In the integral (37), the integration domain of the variable β' is determined by the factor $\Theta_1(\beta')$, which is present in the kernel (see Eqs. (35) and (25)]. If one performs the change of variables $\gamma' \to \gamma^*$, the new integration domains of the integration variables $(\gamma^*, \bar{\gamma}')$ must be determined. As the change of variables depends on γ [see Eq. (42)], the new domain also depends on γ. We are going to describe this domain in two steps: the variation domain of $\bar{\gamma}'$ when γ and γ^* are fixed; the variation domain of γ^* and γ is fixed.

A. Variation domain of $\bar{\gamma}'$ with γ and γ^* fixed

Using Eq. (42), one calculates γ' as a function of γ, γ^* and $\bar{\gamma}'$:

$$\begin{aligned} \beta^{12'} &= u(\beta^{11'} + \beta^{22}), \\ \beta^{22'} &= \beta^{22*} + u^2(\beta^{11'} + \beta^{22}), \\ \beta^{2'} &= \beta^{2*} - u(\beta^2 - \beta^{1'}) \end{aligned} \tag{A1}$$

with

$$u = \beta^{12*}/\beta^{12}.$$

Then one writes that $\beta' = (\gamma', \bar{\gamma}')$ verifies the three conditions (8):

$$(8a) \Rightarrow -2|\beta^{1'}| + (1 - |u|)\beta^{11'} - |u|\beta^{22} \leqslant 0, \tag{A2a}$$

$$\begin{aligned} (8b) \Rightarrow & 2|\beta^{1'} - \beta^2 + \beta^{2*}/u| + (1 - |u|)\beta^{11'} \\ & - \beta^{22*}/|u| + \beta^{22}(1 - |u|) \leqslant 0 \end{aligned} \tag{A2b}$$

$$(8c) \Rightarrow |\beta^{1'}| + |\beta^{2*} - u(\beta^2 - \beta^{1'})| - 2\beta'' \leqslant 0. \tag{A2c}$$

These three inequalities define the variation domain of $\bar{\gamma}' = (\beta^{11'}, \beta^{2'}, \beta'')$.

B. Variation domain of γ^* with γ fixed

This domain is defined by the condition that the previous domain for $\bar{\gamma}'$ is not empty. A necessary condition for the inequality (A2a) to be verified is

$$|u| < 1. \tag{A3}$$

It can be easily shown that the compatibility of the relations (A2a) and (A2b) needs the fact that γ^* verifies the inequality

$$2|\beta^{2*} - u\beta^2| \leqslant \beta^{22*} - |u|\beta^{22}. \tag{A4}$$

The two relations (A3) and (A4) determine the variation domain of γ^*:

$$\Theta_4(\gamma, \gamma^*) = \theta(1 - |u|)\theta(\beta^{22*} - |u|\beta^{22} - 2|\beta^{2*} - u\beta^2|). \tag{A5}$$

[1](a) O. I. Zav'Yalov, Zh. Eksp. Teor. Fiz. **47**, 1099 (1964) [Sov. Phys. JETP **20**, 736 (1965)]; (b) O. I. Zav'Yalov and B. M. Stepanov, Yad. Fiz. **1**, 922 (1965) [Sov. J. Nucl. Phys. **1**, 658 (1965)]; (c) M. C. Bergère and Y. M. P. Lam, Comm. Math. Phys. **39**, 1 (1974), and Freie Universitat Berlin Preprint HEP May 74/9, 1974, unpublished; (d) M. C. Bergère and C. Gilain, J. Math. Phys. **19**, 1495 (1978); (e) M. C. Bergère and C. de Calan, Phys. Rev. D **20**, 2047 (1979).
[2]C. Gilain and D. Levy, J. Math. Phys. **22**, 1787 (1981).
[3]N. Nakanishi, Suppl. Prog. Theor. Phys. **43**, 1 (1969).
[4]C. Gilain, thesis, Université de Paris-Sud-Centre d'Orsay, 1981.
[5]M. C. Bergère and J. B. Zuber, Comm. Math. Phys. **35**, 113 (1974).
[6]G. Domokos and P. Suranyi, Nucl. Phys. **54**, 529 (1964).

Hamiltonian operators with maximal eigenvalues

Evans M. Harrell II[a)]
School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332

(Received 2 August 1983; accepted for publication 16 September 1983)

The potentials $V(x)$ with a given L^p norm that maximize the lowest eigenvalue of $-\Delta + V$ are characterized.

PACS numbers: 02.30.Tb, 03.65. – w

I. INTRODUCTION

How large can a given eigenvalue of a differential operator be? This question has implications for many topics in mathematical physics, especially when the operator has the form

$$H = -\Delta + V(x),$$

where V is a real-valued multiplication operator. Self-adjoint realizations of H are the fundamental mathematical objects of quantum mechanics. The eigenvalues are the energy levels of quantum-mechanical particles, and V is the potential energy. Here the variable will range over a finite open domain D in $\mathbb{R}^m$ with a smooth boundary (an assumption much stronger than necessary), and V will be a nonnegative function in $L^1(D)$. Nonnegativity of V is assumed only to avoid confronting questions of self-adjointness. Except for that it would follow automatically that a potential maximizing an eigenvalue would be nonnegative. H can be defined as a self-adjoint operator on L^2 by either of the following methods.

(a) Let $-\Delta$ be the usual self-adjoint Laplacian with Dirichlet boundary conditions on the boundary of D, and define $-\Delta + V(x)$ via the sum of the associated quadratic forms.[1] Alternatively, equip $-\Delta$ with Neumann or mixed boundary conditions.

(b) Extend H to the infinite domain $\mathbb{R}^m$ by forcing the potential outside D to equal an appropriate fixed function. With the assumptions to be imposed on V, it suffices to have the exterior potential be bounded below, locally integrable, and greater than a positive constant outside some compact set (the constant need only be large enough to ensure the existence of an eigenvalue).

Let $Q(p, c)$ denote the set of potentials V defined on D such that $\|V\|_p \leqslant c$. Let $E(V)$ denote the ground-state (lowest) eigenvalue of H. The question asked above can now be made specific: What is the supremum of $E(V)$ over the set $Q(p, c)$ and for what V is it attained, if any? The answer turns out to be that there is a maximizing potential, and that it is of a very special form, ordinarily the maximal eigenvalue times a characteristic function,

$$V_*(x) = E_{\max}\, \chi_S(x).$$

Indeed, the techniques of this paper also allow one to characterize the function $V(x)$ that maximizes the bottom of the spectrum of a rather general semibounded operator of the form $T + V$, where T represents a closed, semibounded operator on $L^2(D)$ with a few simple properties. Specifically, the domain of self-adjointness of $T + V$ should be the same for all V in $Q(p, c)$ and T should be local in the sense that if f is constant (a.e.) on an open subset U of D, then $Tf = 0$ a.e. on U. For example, T could be a positive higher-order differential operator with no zeroth order term. The maximizing potential function V is still ordinarily of the form $E_{\max}\, \chi_S(x)$, subject to qualifications analogous to the ones spelled out below for the case $T = -\Delta$.

This problem was raised most recently in a list of open problems in mathematical physics at a meeting of the American Mathematical Society.[2] Prominent among the reasons for interest in it are its implications for inverse spectral theory, where for practical as well as theoretical reasons it is important to know what properties of a potential are determined by incomplete spectral information. The result mentioned above would be read by an inverse-spectral theorist the other way around, as stating that if the lowest eigenvalue is larger than a certain amount, then the L^p norms of V are larger than something, and that if a potential has L^p norm equal to c and maximizes the eigenvalue, then it has a particular very simple form. From the latter point of view the statement is reminiscent of Levitan and Gasymov's striking version of Ambarzumian's theorem, viz., for $V \in L^1[0,1]$ and Neumann boundary conditions imposed at 0 and 1, if $E_0 = 0$,

$$E_n - n^2 \to 0,$$

where E_n is the nth eigenvalue, then necessarily $V(x) = 0$ a.e.[3]

II. MAXIMIZING POTENTIALS

Let H be as above, and suppose that V belongs to $Q(p, c)$ for some fixed p, c, and D. In the case $p = \infty$ it is obvious that the lowest or any other eigenvalue is maximized by $V = c$, so $p = \infty$ will not be considered further. It will first be established that there exists a V in $Q(p, c)$ that maximizes the lowest eigenvalue, at least for certain p.

Proposition 1: There is a bound on the lowest eigenvalue depending only on p, c, and D. Consequently there exists a maximizing sequence $V_n \in Q(p, c)$ such that

$$\lim_{n \to \infty} E(V_n) = E_{\max} \equiv \sup_{Q(p,c)} E(V).$$

Proof: The normalized ground-state eigenfunction f_0 of $-\Delta$ is bounded and hence in the quadratic-form domain of H. Therefore an upper bound for $E(V)$ is given by the Rayleigh–Ritz inequality as

a) Partially supported by NSF grant MCS 7926408.

 0022-2488/84/010048-04$02.50

$$E(V) \leqslant E(0) + (f_0, Vf_0)$$
$$\leqslant E(0) + \|f_0\|_\infty^2 \|V\|_1$$
$$\leqslant E(0) + \|f_0\|_\infty^2 \|V\|_p [Vol(D)]^{1/q}, \quad 1/p + 1/q = 1,$$

which depends only on c, p, and D. [$E(0)$ is just the lowest eigenvalue of $-\Delta$.] ∎

Remark: Any sufficiently smooth function f_0 in the quadratic-form domain of H will furnish an upper bound. The normalized ground-state eigenfunction gives a good estimate to compare with the exact answer for simple special cases.

Proposition 2: For all $N > 0$ there exists a $V \in Q(p, c) \cap Q(\infty, N)$ that maximizes $E(V)$ within that class. If $p > \max(2, m/2)$, then there exists a maximizing potential V within $Q(p, c)$.

Proof: By interpolation $Q(p, c) \cap Q(\infty, N)$ lies within $Q(r, c')$ for all $r \geqslant p$ and some c' depending on r. Choose $r > 2$ and $> m/2$; this ensures that the eigenvalue depends continuously on V in the $\| \|_r$ norm.[1,4] The maximizing sequence V_k within $Q(p, c) \cap Q(\infty, N)$ has a subsequence that converges weakly in L^r to some limit V_*. By a theorem of Mazur[5] there is a sequence of convex combinations of V_k that converges strongly to V_*. Since $Q(p, c) \cap Q(\infty, N)$ is convex, the new sequence remains within that class. By the Rayleigh–Ritz inequality, the replacement of V_k by convex combinations can only increase $E(V)$, i.e., if

$$\sum_i a_i = 1, \quad a_i \geqslant 0,$$

and f now denotes the normalized ground-state eigenfunction of

$$-\Delta + \sum_i a_i V_i,$$

then

$$E\left(\sum_i a_i V_i\right) = \left(f, \left(-\Delta + \sum_i a_i V_i\right) f\right)$$
$$= \sum_i a_i (f, (-\Delta + V_i) f)$$
$$\geqslant \sum_i a_i E(V_i).$$

It follows that $E(V_*) = E_{\max}$. Observe that the relevance of Mazur's theorem is more convex combination than the nature of the convergence. The latter takes place in a somewhat arbitrary L^r. Of course, if $p > \max(2, m/2)$, then the truncation to $Q(\infty, N)$ in this proof is unnecessary. ∎

Definition: The potential function V is a local eigenvalue extremizer for the set $Q(p, c)$ iff

(a) $\|V\|_p = c$;

(b) H (or its restriction to a given connected subset of D) has a nondegenerate eigenvalue Λ;

(c) for every bounded multiplicative function $W(x)$ such that

$$\left.\frac{d\|V_t\|_p}{dt}\right|_{t=0} = 0,$$

where $V_t = V + tW$, the eigenvalue $\Lambda(V_t)$ such that $\Lambda(V_0) = \Lambda$ satisfies

$$\left.\frac{d\Lambda(V_t)}{dt}\right|_{t=0} = 0.$$

Remarks: (a) Perturbation theory guarantees the existence and differentiability of $\Lambda(V_t)$ for sufficiently small real values of t.[4]

(b) This is a necessary condition for V to maximize the lowest eigenvalue, which is known to be nondegenerate (after restriction to a connected component of D, if necessary); if it were false, then W could be given some higher-order dependence on t so that $V + tW \in Q(p, c)$, but dE/dt would still differ from 0.

Proposition 3: Any local eigenvalue maximizer in $Q(p, c)$ is equivalent almost everywhere to a function satisfying the nonlinear partial differential equation

$$\Delta V^{(p-1)/2} = (V - \Lambda) V^{(p-1)/2} \tag{1}$$

on the interior of its support.

Remark: This curious equation has the obvious solution $V = \Lambda$ on $S = \text{int supp}(V)$, which is the only solution when $p = 1$. It would be surprising if other conceivable solutions were relevant, but they might arise if either the shape of D or the boundary conditions were peculiar enough. While (1) is trivially satisfied away from S, it is *not* satisfied on the boundary of S, and so does not hold throughout D in the usual distributional sense.

Proof: Let y and z be points in S at the centers of small balls of radius d, denoted Y and Z. Let

$$W(x) = \chi_Y(x) - k\chi_Z(x),$$

where k is chosen to satisfy the condition in (c) of the definition. Since for almost every y and z the averages of V^p over Y and Z approach $V^p(y)$ and $V^p(z)$ as $d \to 0$,[6] from the definition of the L^p norm, k can be taken arbitrarily close to the value

$$(V(y)/V(z))^{p-1}$$

for almost every y and z (write the integrand for $\|V_t\|_p^p$ to first order in t). Let $\psi(x)$ be the normalized eigenfunction for $\Lambda(V)$. By the Feynman–Hellmann theorem,[4]

$$\left.\frac{d\Lambda(V_t)}{dt}\right|_{t=0} = \int \chi_X \psi^2(x)dx - \int \chi_Y k\psi^2(x)dx.$$

For V to be a local eigenvalue extremizer it is necessary for the derivative to be 0 regardless of y, z, and d. By letting d tend to 0, it follows that for almost every y and z in S,

$$\psi^2(y) = (V(y)/V(z))^{p-1}\psi^2(z),$$

or, in other words, that

$$\psi(x) = CV^{(p-1)/2}(x) \text{ almost everywhere on } S \tag{2}$$

for some constant C. Since $\Delta\psi = (V - \Lambda)\psi$ (sense of distributions), Eq. (2) implies Eq. (1). ∎

Actually, Eq. (2) holds almost everywhere on supp(V) (the distinction is the possible existence of nowhere dense sets of positive measure), since the balls can be replaced with appropriate sets that "shrink nicely."[6]

Proposition 4: Let $V \geqslant 0$, $V \in L^p(D)$, $p \geqslant 1$, D as above and moreover assumed connected. Define $V_T = \min(V, T)$.

Let $E_0(H)$ and $\phi(H)$ denote the ground-state eigenvalue and eigenfunction of an operator H. Then $E_0(-\Delta + V_T)$ tends monotonically to $E_0(-\Delta + V)$ and $\phi(-\Delta + V_T)$ tends to $\phi(-\Delta + V)$ in L^2.

Remark: Connectedness just ensures nondegeneracy of the ground state.

Proof: For simplicity of notation, let $f = \phi(-\Delta + V)$ and $E = E_0(-\Delta + V)$. Monotonicity of the eigenvalue is an immediate and well-known consequence of the min-max principle, or the Rayleigh–Ritz inequality. From straightforward corollaries of the spectral theorem it suffices to show that

$$\|(-\Delta + V_T - E)f\|_2 \to 0.$$

Actually, this just ensures that some point of the spectrum of $-\Delta + V_T$ tends to E and the associated eigenfunction converges. But since the ground-state eigenfunctions are characterized by positivity, that point has to be the ground state. Also, set $p = 1$, which includes all the other cases.

Since $V_T(x)f(x)$ increases monotonically to $V(x)f(x)$, the distribution $(-\Delta + V_T)f$, which is only in L^1 *a priori*,[1] increases to $(-\Delta + V)f = Ef \in L^2$. Therefore

$$\|(-\Delta + V_T - E)f\|_2^2 = \int_D ((-\Delta + V_T - E)f)^2 d^m x$$

is finite, and hence tends to zero by the monotone convergence theorem. ∎

Theorem 1: For $p = 1$ or $p > 2$, $m/2$, there is a potential in $Q(p,c)$ that maximizes the lowest eigenvalue, and it satisfies (1) with $\Lambda = E_{\max}$ on S. In particular, when $p = 1$, $V_* = E_{\max}$ and ψ equals its maximum almost everywhere on S.

Proof: The foregoing propositions cover all p other than $p = 1$. If $p = 1$, then consider the set $Q(1, c) \cap Q(\infty, N)$ in place of $Q(1, c)$, where N is larger than the upper bound on $E(V)$ from Proposition 1. The proof of Proposition 3 goes through unchanged, so that on $\text{supp}(V)$, $\psi(x) = C$ (a fixed constant) and $V = E(V)$ almost everywhere, independently of N as $N \to \infty$. But truncation of V at high values affects the ground-state eigenvalue continuously by Proposition 4. Hence there cannot be an unbounded $V \in Q(1, c)$ with a higher eigenvalue than the maximum on $Q(1, p) \cap Q(\infty, N)$. ∎

Theorem 2: If $p = 1$, or if $p \neq 1$, but it is known that V_* exists and is constant on its support, then V_* is unique a.e.

Proof: Suppose that there were two distinct sets S. Then, as in the proof of Proposition 2, the eigenvalue corresponding to the average of the two maximizing potentials would be no less than $E_{\max}$, since the average is a convex combination. This is a contradiction, since the averaged potential would equal $E_{\max}/2$ on a set of positive measure. ∎

What makes the proof of Theorem 1 work is that all the maximizing potentials within $Q(1, c) \cap Q(\infty, N)$ satisfy a pointwise bound independent of N. If the same were known for all p, then the restriction to values for which V is relatively bounded could be dispensed with. It would suffice, for example, to know that the only solution of (1) of interest is the obvious one. In principle, these arguments leave open the possibility that different solutions are relevant for different N, and do not have a uniform bound.

III. EXAMPLES

The one-dimensional case of an interval is rather easy to analyze in detail, since there are no geometrical complications and since all eigenvalues are automatically nondegenerate. By a change of variable it suffices to consider only the interval $[0,1]$. The case of a sphere is similar.

Scholium: Let H be the one-dimensional operator $-d^2/dx^2 + V(x)$ on $L^2[0,1]$, with Dirichlet boundary conditions, and denote the nth eigenvalue E_n, $n = 0,1,2,\ldots$. Let V range over $Q(1, c)$. The eigenvalue E_n is maximized by potentials of the form

$$V_n(x) = E_{n,\max}\, \chi_{S_n}(x),$$

uniquely determined only for $n = 0$. If $n > 0$, then there are uncountably many distinct choices of S_n, which can consist of any number of subintervals from 1 to $n + 1$. The subintervals are constrained only by their total length and the distances between them and from them to the endpoints 0 and 1.

The somewhat informal proof will be given by constructing the possible potentials. In one dimension there is no possibility of $\bar{S}$ differing from $\text{supp}(V_*)$, since $\text{supp}(V_*)$ is the set on which the corresponding eigenfunction ψ has its maximum or minimum value, and on the complement ψ is a simple exponential function. Since ψ is not maximized at 0, V must equal 0 on some interval beginning at 0. Since $\psi \in C^1$, its first chance to attain its maximum occurs when $\sin(\sqrt{E_n}\, x) = 1$, i.e., at

$$x = \pi\sqrt{E_n}/2.$$

At that point the eigenfunction may either be constant for a while or continue oscillating until some later maximum or minimum. It is a matter of utter indifference how long the eigenfunction remains constant after reaching a sinusoidal maximum or minimum, so long as the total length of constancy has the correct value. By the Sturmian theorem, the nth eigenfunction must make $(n + 1)/2$ complete sinusoidal oscillations punctuated by intervals on which it is constant. The total length of the oscillations is $(n + 1)\pi/\sqrt{E_n}$, while from the condition that $V_n \in Q(1, c)$ the total length of the intervals of constancy of ψ is c/E_n (see Fig. 1). Therefore

$$(n + 1)\pi/\sqrt{E_n} + c/E_n = 1.$$

The solution of this is

$$E_n = ((n + 1)\pi + ((n + 1)^2\pi^2 + 4c)^{1/2})^2/4.$$

For instance, the first several maximum eigenvalues are

n	$E_{n,\max}$ $c = 1$	$c = 10$
0	11.784 7490	26.027 5168
1	41.454 2947	57.746 7175
2	90.815 4283	107.899 653
3	159.907 417	177.349 813
4	248.736 090	266.364 685
5	357.302 960	375.039 120

The asymptotic form is $E_{n,\max} \sim ((n + 1)\pi)^2$. For compari-

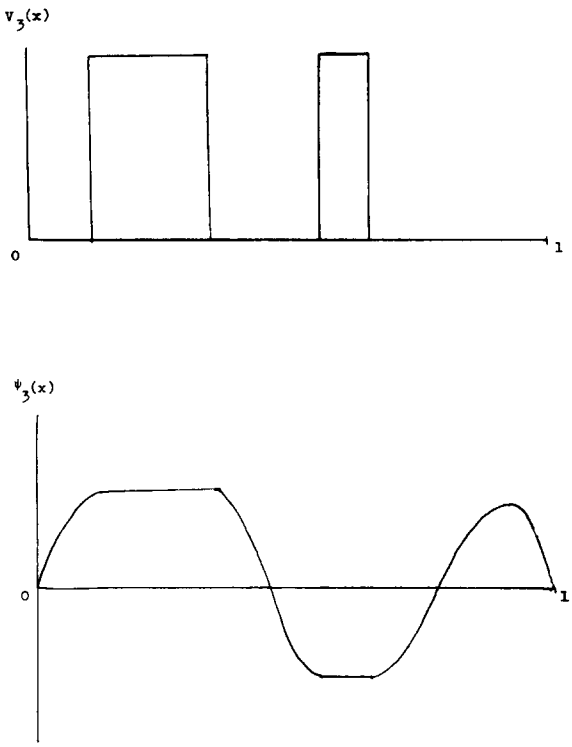

FIG. 1. Typical maximizing potential and eigenfunction for a higher eigenvalue.

son, the bounds on $E_{0,\max}$ from Proposition 1 are, respectively, 11.869 6044 and 29.869 6044 (i.e., $\pi^2 + 2c$), and the lowest eigenvalue with $V = 0$ is $\pi^2 = 9.869\ 6044$. $E_{0,\max}$ has been found by an independent method by Farris.[7]

The maximizing potential for the lowest eigenvalue with Neumann boundary conditions is $V(x) = c$, and the maximizers of the higher Neumann eigenvalues are obtained by an argument analogous to the above.

Similarly, if $n > 1$ and D is a regular figure, such as a cube, sphere, ellipsoid, etc., it is highly probable that the maximal lowest eigenvalue is attained when S is a smaller concentric figure of similar shape, and the maximum eigenvalues can be obtained explicitly in terms of the special functions associated with the separated Laplacian.

This is certainly true of the sphere. Let $p = 1$ and let D be the unit sphere in $\mathbb{R}^m$. The maximizing potential for the lowest eigenvalue is of the form

$$V_*(x) = E_{\max}\, \chi_S(x).$$

The set $\bar{S}$ in this case is again equal to $\mathrm{supp}(V_*)$ and must be a concentric sphere. This is because a spherical average of all rotations of any putative V_* would lead to at least as high an E_*, as seen above. Yet $\mathrm{supp}(V_*)$ cannot be hollow without violating the minimum principle for the superharmonic ground-state eigenfunction on $\mathrm{supp}(V_*)^c$.

It follows that the eigenvalue equation is separable and reduces to the one-dimensional equation

$$-R''(r) - (m-1)R'(r)/r + (V_*(R) - E_{\max})R(r) = 0,$$

which is just a form of Bessel's equation, with solutions

$$R(r) = r^{1-m/2}\mathscr{C}_{m/2-1}(\sqrt{E_{\max} - V_*}\, r)$$

on the interval $[0, r_0]$ on which V_* is constant, where $\mathscr{C}$ is any of the usual Bessel functions of index $m/2 - 1$. Consequently, $E_{\max}$ is the unique solution of the following triple of equations in three unknowns, $E_{\max}$, r_0, and a:

$$E_{\max}\, \omega_m\, r_0^m = c,$$

$$J_{m/2-1}(\sqrt{E_{\max}}) + aY_{m/2-1}(\sqrt{E_{\max}}) = 0 \quad \text{(first zero)},$$

$$\frac{d}{dr}\left(r^{1-m/2}(J_{m/2-1}(\sqrt{E_{\max}}\, r) + aY_{m/2-1}(\sqrt{E_{\max}}\, r))\right)\Big|_{r=r_0} = 0,$$

where ω_m is the volume of the m-sphere. In dimension $m = 3$, the Bessel functions reduce to circular functions, and the equations may be written

$$4\pi r_0^3\, E_{\max}/3 = c,$$

$$\sqrt{E_{\max}} + \phi = \pi,$$

$$\tan(\sqrt{E_{\max}}\, r_0 + \phi) = \sqrt{E_{\max}}\, r_0.$$

These are easy to solve numerically. For example, with $c = 1$,

$$E_{\max} \doteq 11.024\ 7609.$$

(The lowest eigenvalue with $V = 0$ is $\pi^2 \doteq 9.869\ 6044$, and the upper bound from Proposition 1 is $\pi^2 + \pi/2 \doteq 11.440\ 4007$.)

[1]M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, in four volumes (Springer, New York, 1972–1979).
[2]Problem list of A. G. Ramm in H. Samelson, "Queries," Notices Am. Math. Soc. **29**, 326–329 (1982).
[3]B. M. Levitan and M. G. Gasymov, "Determination of a differential equation by two of its spectra," Uspehi Mat. Nauk **19**(2), 3–63 (1964) [Russ. Math. Surveys **19**(2), 1–63 (1964)]. Curiously, Levitan and Gasymov state the theorem in Ambarzumian's original form, assuming that all $E_n = n^2$, although they prove this much more powerful version.
[4]T. Kato, *Perturbation Theory for Linear Operators, Die Grundlehren der mathematischen Wissenschaften*, Vol. 132 (Springer, New York, 1966). The Feynman–Hellmann theorem is not identified as such, but is equation (3.18) on p. 391.
[5]K. Yôsida, *Functional Analysis, Die Grundlehren der mathematischen Wissenschaften*, Vol. 123 (Springer, Heidelberg, 1965).
[6]W. Rudin, *Real and Complex Analysis*, 2nd ed. (McGraw–Hill, New York, 1974).
[7]M. Farris, "A Sturm–Liouville problem with maximal first eigenvalue," preprint, 1982.

Stochastic path-ordered exponentials

Jürgen Potthoff
Fakultät für Physik, Universität Bielefeld, D-4800 Bielefeld 1, Federal Republic of Germany

(Received 20 July 1982; accepted for publication 10 December 1982)

We prove convergence of an approximation of the stochastic product integral for conditional Wiener paths to the solution of a certain stochastic integral equation. This is used to establish the Wiener–Itô representation for the kernel of the semigroup $\exp t\Delta_A$, where $\Delta_A = \Sigma_\mu (\partial_\mu \mathbb{1} + A_\mu)^2$ for functions A_μ with values in the space of anti-Hermitian matrices.

PACS numbers: 02.50.Ey, 02.30. – f

I. INTRODUCTION

The aim of this paper is to construct a (symmetrized) stochastic product integral w.r.t. the D-dimensional conditional Wiener path $\mathbf{Z}$ starting at $\mathbf{x}$ at time zero, ending at $\mathbf{y}$ at time t. The product integral is defined as the limit of a polygonal approximation and we show convergence of this approximation to the solution of a certain stochastic integral equation w.r.t. $\mathbf{Z}$.

The existence of this product integral, which we suggestively denote by $\overleftarrow{\prod}_{s<t} \exp \mathbf{A}(Z_s)\cdot d\mathbf{Z}_s, \overleftarrow{\prod}$ denoting a product whose factors are ordered with increasing time to the left, allows writing the Wiener–Itô representation of the kernel of the semigroup $\exp t\Delta_{\mathbf{A}}$, $t \geqslant 0$, with $\Delta_{\mathbf{A}} = \Sigma_{\mu=1}^{D} (\partial/\partial x_\mu + A_\mu)^2$, on $L^2(\mathbb{R}^D, \mathbb{C}^m)$:

$$(\exp t\Delta_{\mathbf{A}})(x,y) = \int dP_{\mathbf{xy}}^t \overleftarrow{\prod_{s<r}} \exp(\mathbf{A}(\mathbf{Z}_s)\cdot d\mathbf{Z}_s), \tag{1.1}$$

where A is a D-tuple of continuous functions such that div $\mathbf{A}$ is continuous, with values in the space of anti-Hermitian $m \times m$ matrices and $dP_{\mathbf{xy}}^t$ is the conditional Wiener measure.

This formula, which turned out to be very useful in Euclidean quantum field theory and whose proof for the case $m = 1$ can be found in Ref. 1, Chap. V, appears already in several papers.[2–6] A discussion of the proof for $m \geqslant 1$ is found in Refs. 2 and 3; however, there both authors construct the product integral for the Brownian path without fixed endpoint and restrict the integration over these paths [cf. (1.1)] to those with fixed endpoint. Unfortunately the product integral for Brownian paths is defined only up to sets of measure of zero, so that the validity of their discussions is not clear, since the conditioned paths $\mathbf{Z}$ from a set of measure zero.

Stochastic product integrals for Brownian motion have been studied by several authors (see Refs. 7–9 and literature quoted there).

A basic tool of these works is to use the independence of the increments of the Wiener process of the past, i.e., its martingale property, which does not hold for the $\mathbf{Z}$-process. Although Simon[1] has shown how one can overcome this difficulty for defining stochastic integrals w.r.t. $\mathbf{Z}$ by an appropriate decomposition of the increments, this is not sufficient to generalize the proofs presented in Refs. 7 and 9.

Actually, in this paper we have to make use of the ideas of the Strasbourg school[10–13]—in particular Emery has already developed a theory of stochastic product integrals w.r.t. semimartingales and their related integral equations[10] in a very general framework.

On the other hand the $\mathbf{Z}$-process is simple enough (e.g., it is almost surely continuous) to allow for a detailed treatment without going through all the complications provided by the general situation. In this sense part of the present paper can be understood as an illustration (with some modifications) of the ideas found in Ref. 10 and in the beautiful book of Métivier and Pellaumail.[12]

Instead of working directly with the $\mathbf{Z}$-process we prefer to work with the D-dimensional Brownian bridge $\mathbf{W}$, which is related to $\mathbf{Z}$ via

$$\mathbf{Z}_s \doteq (1 - s/t)\mathbf{x} + s/t\,\mathbf{y} + \sqrt{t}\,\mathbf{W}_{s/t}, \quad 0 \leqslant s \leqslant t,$$
$$\int dP_{\mathbf{xy}}^t \doteq (2\pi t)^{-D/2} \exp\left(-\frac{|\mathbf{x}-\mathbf{y}|^2}{2t}\right) E(\cdot), \tag{1.2}$$

where $\doteq$ means equality in sense of probability distributions and $E(\cdot)$ denotes expectations; i.e., $\mathbf{W}_s$, $0 \leqslant s \leqslant 1$, is the Gaussian process [over a probability space $(\Omega, \mathcal{F}, P)$] of mean zero and covariance matrix $E(\mathbf{W}_s \mathbf{W}_t) = \mathbb{1}_D s(1-t)$ for $0 \leqslant s \leqslant t \leqslant 1$, $\mathbb{1}_D$ denoting the D-dimensional unit matrix.[14]

The paper is organized as follows. In Sec. II we discuss some preliminary material; in Sec. III we study integral equations w.r.t. $\mathbf{W}$ and show convergence of the product integral. Finally in Sec. IV we prove the Wiener–Itô representation for the kernel of $\exp t\Delta_{\mathbf{A}}$ as given in (1.1).

II. PRELIMINARY RESULTS[15]

As mentioned in the Introduction the problem in dealing with the Brownian bridge W comes from the dependence of its increments of the past. Simon[1] has shown how to bypass this difficulty using the decomposition

$$W_{t+\Delta t} - W_t = \left(W_{t+\Delta t} - \frac{1-(t+\Delta t)}{1-t} W_t\right) - \Delta t \frac{1}{1-t} W_t, \tag{2.1}$$

so that the increment in ( ) on the rhs is past independent. However, in this paper we need some more detailed information about the dW-integral than is available in Ref. 1, such as continuity of $\int_0^t \cdot dW_s$ in t.

 0022-2488/84/010052-05$02.50

We note that an "integrated version" of (2.1) reads

$$W_t = B_t - \int_0^t ds(1-s)^{-1}W_s, \quad 0\leqslant t\leqslant 1, \tag{2.2}$$

where B_t has the same probability distribution as the standard Brownian motion b_t ($B_t \doteq b_t$), as one easily checks.

Let the underlying probability space of the theory be denoted by $(\Omega,\mathcal{F},P)$ and let the filtration of σ-subalgebras generated by b_t be $(\mathcal{F}_t')_{t>0}$ (i.e., b is an $(\mathcal{F}_t')$-martingale). Then Jeulin[11] shows that B_t is measurable with respect to the enlarged filtration $(\mathcal{F}_t)_{t>0}$, where $\mathcal{F}_t = \mathcal{F}_t' \vee \sigma(b_1)$, $\sigma(b_1)$ denoting the subalgebra generated by b_1. Thus the filtration $(\mathcal{F}_t)_{t>0}$ is the "natural" one in this framework and in fact B is an $(\mathcal{F}_t)$-martingale, so that by (2.2) W_t is an $(\mathcal{F}_t)$-semimartingale.[11,16]

Henceforth measurability is understood w.r.t. $\mathcal{F}$ or $(\mathcal{F}_t)$ depending on the context.

The representation (2.2) allows now for an easy adaption of the construction of stochastic integrals $\int_0^t X_s\,dW_s$ as, e.g., in McKean's book[9] for nonanticipating functionals X of W (i.e., X_s is $\mathcal{F}_s$-measurable for $0\leqslant s\leqslant 1$) satisfying some suitable boundedness condition (see below).

Obviously we have the bound

$$E\left(\left(\int_0^t X_s\,dW_s\right)^2\right)\leqslant 2\left[E\left(\int_0^t X_s^2\,ds\right) + E\left(\left(\int_0^t X_s W_s(1-s)^{-1}ds\right)^2\right)\right], \tag{2.3}$$

and using Hölder's inequality it can be shown that for X such that $E(\int_0^1 |X_s|^{2+\epsilon}ds) < \infty$, for any $\epsilon > 0$, one can define $\int_0^t X_s dW_s$, $0\leqslant t\leqslant 1$ as an integral continuous in t.
All this generalizes now naturally to the case of D-dimensional Brownian bridge $\mathbf{W}$ (i.e., D independent copies of W) and X taking values in some Banach space $\mathcal{H}$ with norm $\|\cdot\|$.

We shall have to use the following

Definition 2.1: A stopping time u is a map $u:\Omega\to[0,1]$ so that $\{\omega;u(\omega)\leqslant t\}\in\mathcal{F}_t$ for every $t\in[0,1]$.

A stochastic interval $[u,v)$, for two stopping times u,v is the set $\{(\omega,t);u(\omega)\leqslant t<v(\omega)\}\subset\Omega\times[0,1]$. $[u,v]$, (u,v), $(u,v]$ are defined similarly.

If X is a process with values in $\mathcal{H}$ and if u is a stopping time, denote $X_u^* = \sup_{0<t<u}\|X_t\|$.

For a D-tuple of processes $\mathbf{X}$, whose components X_μ take values in $\mathcal{H}$, we let $\|X_t\|^2 \equiv \Sigma_\mu \|X_{\mu t}\|^2$ and define X_u^* similarly.

Using the fact that B is a continuous $(\mathcal{F}_t)$-martingale the results of Sec. 6.9 of Ref. 12 imply for $Z_t := \int_0^t \mathbf{X}_s\cdot d\,\mathbf{B}_s$[17] the bound

$$E\left(\sup_{t<u}\|Z_t\|^2\right)\leqslant 8E\left(\int_0^u \|X_s\|^2 ds\right). \tag{2.4}$$

The following theorem is a generalization of the preceding consideration.

Theorem 2.2: Let $\mathbf{X}$ be a D-tuple of $\mathcal{H}$-valued processes so that $E(\int_0^1 \|\mathbf{X}_s\|^{2+\epsilon}ds)$ is finite; then one can define the stochastic integral $\int_0^t \mathbf{X}_s\cdot d\,\mathbf{W}_s$ as a continuous function of t. Moreover, one has the estimate

$$E\left(\sup_{t<u}\left|\left|\int_0^t \mathbf{X}_s\cdot d\,\mathbf{W}_s\right|\right|^2\right) \leqslant E\left(Q_u\int_0^u \|\mathbf{X}_s\|^2(1-s)^{-1/2}ds\right) \tag{2.5}$$

for any stopping time u, where Q denotes the continuous, increasing process

$$Q_t = 16\left(1+\int_0^t |\mathbf{W}_s|^2(1-s)^{-3/2}ds\right), \quad 0\leqslant t\leqslant 1.$$

Remark: Continuity of Q is due to the fact that the integral exists for all $t\in[0,1]$ as a consequence of Hölder continuity of the Brownian bridge W.

(2.5) is similar to what is called "π^*-property" in Ref. 12.

Let us conclude this section by the observation that if $\mathbf{X}$ has the form $\mathbf{X}_s = \mathbf{X}(\mathbf{W}_s)$ then the condition $E(\int_0\|\mathbf{X}_s\|^{2+\epsilon}ds) < \infty$ (in order to define $\int_0^t \mathbf{X}_s\cdot d\,\mathbf{W}_s$) can be replaced by $\mathbf{X}\in L^p_{\text{loc}}$, $p>2$ if $D=1$, and $p>D$ if $D\geqslant 2$, as Hölder's inequality and the use of continuity of $\mathbf{W}$ show.

III. CONVERGENCE OF THE PRODUCT INTEGRAL

For the rest of the paper we let $\mathcal{H}$ be the Banach space of complex $m\times m$ matrices, $\mathbb{1}$ representing the unit matrix equipped with the operator norm $\|\cdot\|$ on $\mathbb{C}^m$.

The central result of this section is to define the product integral by an approximation which is shown to converge to the (unique) solution of a certain stochastic integral equation. Let us begin with a short study of a class of integral equations, which is an adaptation of the very general theory in Ref. 12 to our simple situation.

Consider the equation (let $D=1$ for notational convenience)

$$X_t = \mathbb{1} + \int_0^t dW_s A_s X_s, \quad t\in[0,1] \tag{3.1}$$

for (nonanticipating) A with values in $\mathcal{H}$.

We can state the following

Lemma 3.1: Let A be such that $E(\int_0^1 \|A_s\|^2(1-s)^{-1/2}ds) < \infty$.

Then the integral equation (3.1) admits a unique solution.

The proof of this lemma has two steps. First one shows that (3.1) has a unique solution on a sufficiently small stochastic interval $[0,u]$ (using the Banach fixed point theorem). Then one extends the solution globally by $[0,1]$.

Define a stopping time u by[18]

$$u = \inf\left\{t;t\geqslant 0, Q_t\int_0^t \|A_s\|^2(1-s)^{-1/2}ds > \tfrac{1}{2}\right\}\wedge 1.$$

Let $\mathcal{X}$ be the complete metric space of continuous H-valued processes defined on $[0,u]$, with $X_0 = \mathbb{1}$ for $X\in\mathcal{X}$ and $|||X|||^2 := E(\sup_{t<u}\|X_t\|^2)$ finite. We define a mapping $U:\mathcal{X}\to\mathcal{X}$ by

$$(UX)_t = \mathbb{1} + \int_0^t dW_s A_s X_s. \tag{3.2}$$

By Theorem 2.2, one easily verifies that $\mathcal{D}(U) = \mathcal{X}$:

$$|||UX|||^2 \leqslant 2\left(1 + E\left(\sup_{t<u} Q_t \int_0^t \|A_s\|^2 \|X_s\|^2 (1-s)^{-1/2} ds\right)\right)$$
$$\leqslant 2 + |||X|||^2 < \infty.$$

To prove that U is a contraction let $X, X' \in \mathscr{X}$. Then

$$|||U(X-X')|||^2$$
$$\leqslant E\left(\sup_{t<u} Q_t \int_0^t \|A_s\|^2 \|X_s - X'_s\|^2 (1-s)^{-1/2} ds\right)$$
$$\leqslant \tfrac{1}{2}|||X-X'|||^2,$$

again by Theorem 2.2.

Finally we note that $u > 0$. The condition $E(\int_0^t \|A_s\|^2 (1-s)^{-1/2} ds) < \infty$ implies that $P(\int_0^t \|A_s\|^2 (1-s)^{-1/2} ds > 2^k) \leqslant 2^{-k} \times$ const for every t and k, so that the Borel–Cantelli lemma implies that $\|A_s\|^2 (1-s)^{-1/2}$ is integrable on [0,1] and hence by continuity of Q_t and $\int_0^t \|A_s\|^2 (1-s)^{-1/2} ds$ in t $u > 0$. This concludes the first step.

Note that $|||X|||^2 < \infty$ clearly implies that $\|X_t\| < \infty$ for $t \leqslant u$. Hence, choosing some large $B > 0$, one can extend the solution by the same method as before for all those $\omega \in \Omega$, so that $X^*_u \leqslant B$ and for a new stopping time $u' > u$, so that $Q_t \int_u^t \|A_s\|^2 (1-s)^{-1/2} ds \leqslant \frac{1}{2}$ for $t \leqslant u'$.

This is systematized in the following construction. Define recursively a sequence of stopping times $\{u_k\}_{k>0}$ as follows: $u_0 = 0$; given u_k choose B_k large enough such that $P(X^*_{u_k} > B_k) \leqslant 2^{-k}$. Then if $X^*_{u_k} > B_k$ put $u_{k+1} = u_k$; if $X^*_{u_k} \leqslant B_k$ let

$$u_{k+1} = \inf\left\{t; t \geqslant u_k, Q_t \int_{u_k}^t \|A_s\|^2 (1-s)^{-1/2} ds > \tfrac{1}{2}\right\} \wedge 1.$$

On each stochastic interval one can now apply the contraction mapping principle as above. But as $k \to \infty$ $u_k \to 1$, which proves the lemma.

The lemma is easily generalized to

Theorem 3.2: Consider the D-dimensional Brownian bridge $\mathbf{W}$. Let a D-tuple of nonanticipating functionals $\mathbf{A}$ and a nonanticipating B, $\mathbf{A}$ and B taking values in $\mathscr{H}$, be such that $E(\int_0^1 \|A_s\|^2 (1-s)^{-1/2} ds)$ and $E(\int_0^1 \|B_s\|^2 ds)$ are finite. Then the integral equation

$$X_t = \mathbb{1} + \int_0^t d\mathbf{W}_s \cdot \mathbf{A}_s X_s + \int_0^t ds B_s X_s \tag{3.3}$$

has a unique continuous solution on [0,1].

Remark: As before for $\mathbf{A}_s = \mathbf{A}(\mathbf{W}_s)$, $B_s = B(\mathbf{W}_s)$ the preceding conditions can be replaced by $\mathbf{A} \in L^p_{\text{loc}}$, $B \in L^q_{\text{loc}}$, p as in the remark after Theorem 2.1, $q = 2$ if $D = 1$, $q > D$ if $D \geqslant 2$.

In the following we assume $\mathbf{A}_s = \mathbf{A}(\mathbf{W}_s, s)$ and that $\mathbf{A}$ is C^2 on $\mathbb{R}^D \times [0,1]$, bounded with bounded first and second derivatives.

Define a family of processes on [0,1], indexed by $n \in \mathbb{N}$, as follows:

$$X^n_t = \exp\left[\tfrac{1}{2}(\mathbf{A}_t + \mathbf{A}_{(m-1)/2^n}) \cdot (\mathbf{W}_t - \mathbf{W}_{(m-1)/2^n})\right] \times \prod_{k=1}^{m-1} \exp\left[\tfrac{1}{2}\mathbf{A}_{k/2^n} + \mathbf{A}_{(k-1)/2^n}) \cdot \Delta_k \mathbf{W}\right] \tag{3.4}$$

for

$$t \in \Delta_m := \left[\frac{m-1}{2^n}, \frac{m}{2^n}\right] \quad \text{and}$$
$$\Delta_k \mathbf{W} := \mathbf{W}_{k/2^n} - \mathbf{W}_{(k-1)/2^n}.$$

For later convenience we introduce the following notations: $\Delta_k \mathbf{A} := \frac{1}{2}(\mathbf{A}_{k/2^n} + \mathbf{A}_{(k-1)/2^n}) \cdot \Delta_k \mathbf{W}$; for D vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$, etc. and ∇ the gradient on $\mathbb{R}^D$ $\mathbf{x} \cdot (\nabla \mathbf{y}) \cdot \mathbf{z} = \sum_{\mu,\nu=1}^D x_\mu (\partial_\mu y_\nu) z_\nu$, $((\nabla \mathbf{y}) \cdot \mathbf{z})_\mu = \sum_{\nu=1}^D (\partial_\mu x_\nu) y_\nu$, etc. Also $J^n \mathbf{A}$ denotes the process $(J^n \mathbf{A})_t = \sum_{k=1}^{2^n} 1_{\Delta_k}(t) \mathbf{A}_{(k-1)/2^n}$.

We shall now show that X^n_t converges as $n \to \infty$ uniformly on [0,1] to the solution X_t of Eq. (3.3), B_s being given by $B_s = \frac{1}{2}(\nabla \cdot \mathbf{A}_s + \mathbf{A}_s^2)$. This is done in three steps. First we derive for X^n_t an integral equation of the type previously discussed. Then we show how to reduce the question of convergence of X^n_t to X_t to the question of convergence of their coefficient functions. Finally we prove convergence of the latter.

Proposition 3.3: Let X^n_t be given by (3.4). Then X^n_t is the solution of the integral equation

$$X^n_t = \mathbb{1} + \int_0^t d\mathbf{W}_s \cdot \mathbf{C}_s X^n_s + \int_0^t ds D_s X^n_s, \tag{3.5}$$

where

$$\mathbf{C}_s = \tfrac{1}{2}\{(\nabla \mathbf{A}_s) \cdot (\mathbf{W}_s - (J^n \mathbf{W})_s) + \mathbf{A}_s + (J^n \mathbf{A})_s\},$$
$$D_s = \tfrac{1}{4}\{(\Delta \mathbf{A}_s) \cdot (\mathbf{W}_s - (J^n \mathbf{W})_s) + 2(\nabla \cdot \mathbf{A})_s + \tfrac{1}{2}[(\nabla \mathbf{A}_s) \cdot (\mathbf{W}_s - (J^n \mathbf{W})_s) + \mathbf{A}_s + (J^n \mathbf{A})_s]^2 + 2\left(\frac{\partial}{\partial s}\mathbf{A}\right)_s \cdot (\mathbf{W}_s - (J^n \mathbf{W})_s)\}. \tag{3.6}$$

Proof: The proof is based on an application of Itô's lemma.[19] For $t \in \Delta_m$, $1 \leqslant m \leqslant 2^n$, we compute

$$\int_0^t d\mathbf{W}_s \cdot \mathbf{C}_s X^n_s$$
$$= \sum_{k=1}^m \left\{\frac{1}{2}\int_0^t 1_{\Delta_k}(s) d\mathbf{W}_s \cdot ((\nabla \mathbf{A}_s) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n}) + (\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n})) \exp\left[\tfrac{1}{2}(\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n}) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n})\right]\right\} \prod_{l=1}^{k-1} \exp \Delta_l \mathbf{A} \tag{3.7}$$

using the definition of X^n_t. By Itô's lemma

$$\tfrac{1}{2} d\mathbf{W}_s \cdot ((\nabla \mathbf{A}_s) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n}) + (\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n})) \times \exp\{\tfrac{1}{2}(\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n}) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n})\}$$
$$= d \exp\left[\tfrac{1}{2}(\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n}) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n})\right] - ds 1_{\Delta_k}(s) D_s \exp\left[\tfrac{1}{2}(\mathbf{A}_s + \mathbf{A}_{(k-1)/2^n}) \cdot (\mathbf{W}_s - \mathbf{W}_{(k-1)/2^n})\right]$$

for D as defined before. Inserting this into the rhs of (3.7) gives

$$\int_0^t d\mathbf{W}_s \cdot \mathbf{C}_s X^n_s = X^n_t - \mathbb{1} - \int_0^t ds D_s X^n_s$$

proving the proposition.

Proposition 4.4: Let X_t, X^n_t be as above and u be the following stopping time:

$u = \inf\{t;t>0,$

$$Q_t(\sup_{s<t}(\|X_s\|^2 + \|C_s\|^2 + \|D_s\|^2)) > L\} \wedge 1,$$

L some positive constant. Then we have

$$E\left(\sup_{t<u}\|X_t - X_t^n\|^2\right) \leqslant KE\left(\sup_{t<1}(\|\mathbf{A}_t - \mathbf{C}_t\|^2 + \|B_t - D_t\|^2)\right),$$

where the constant K depends only on L.

Proof: Write

$$\begin{aligned} X_t - X_t^n &= \int_0^t d\,\mathbf{W}_s\cdot(\mathbf{A}_s - \mathbf{C}_s)X_s \\ &\quad + \int_0^t d\,\mathbf{W}_s\cdot\mathbf{C}_S(X_s - X_s^n) \\ &\quad + \int_0^t ds(B_s - D_s)X_s \\ &\quad + \int_0^t ds D_s(X_s - X_s^n), \end{aligned}$$

and by Theorem 2.2 and the definition of u we obtain

$$\begin{aligned} E\left(\sup_{t<u}\|X_t - X_t^n\|^2\right) \leqslant 4\Big\{ & 2LE\left(\sup_{t<q}\|\mathbf{A}_t - \mathbf{C}_t\|^2\right) \\ & + 2LE\left(\int_0^u \|X_s - X_s^n\|^2 \times (1-s)^{-1/2}ds\right) \\ & + LE\left(\sup_{t<1}\|B_t - D_t\|^2\right)\Big\}. \end{aligned}$$

Hence denoting $\phi_t = \sup_{s<t}\|X_s - X_s^n\|^2$ we may bound

$$\begin{aligned} E(\phi_u) \leqslant & 8LE(\sup_{t<1}(\|\mathbf{A}_t - \mathbf{C}_t\|^2 + \|B_t - D_t\|^2)) \\ & + 8LE\left(\int_0^u \phi_s(1-s)^{-1/2}ds\right). \end{aligned} \tag{3.8}$$

The following very simple version of Gronwall's lemma (cf., e.g., Ref. 12) shows that (3.8) implies the proposition:

Let $\{t_k\}_{0<k<k_0}$ be a finite increasing sequence with $t_0 = 0$, $t_{k_0} = 1$ and

$$\int_{t_k}^{t_{k+1}} ds(1-s)^{-1/2} \leqslant (16L)^{-1}.$$

Define a sequence of stopping times $\{v_k\}$ by setting $v_k = t_k \wedge u$. Then (3.8) entails

$$\begin{aligned} E(\phi_{v_{k+1}}) \leqslant & 8LE(\sup_{t<1}(||A_t - C_t||^2 + ||B_t - D_t||^2)) \\ & + 8LE(\phi_{v_k}) + \tfrac{1}{2}E(\phi_{v_{k+1}}), \end{aligned}$$

so that by iteration for every $k \leqslant k_0$,

$$\begin{aligned} E(\phi_{v_{k+1}}) \leqslant & 16LE(\sup_{t<1}(\|\mathbf{A}_t - \mathbf{C}_t\|^2 + \|B_t - D_t\|^2)) \\ & \times \sum_{j=0}^{k_0}(16)^j \end{aligned} \tag{3.9}$$

and (3.9) holds in particular for $E(\phi_u)$.

Proposition 3.5: Let $\mathbf{A}$, B, $\mathbf{C}$, D be as above. Then

$$E(\sup_{t<1}(\|\mathbf{A}_t - \mathbf{C}_t\|^2 + \|B_t - D_t\|^2)) \leqslant \text{const} \times 2^{-n}.$$

Proof: Using the explicit expressions (3.6) and $B_s = \frac{1}{2}(\nabla\cdot\mathbf{A})_s + \mathbf{A}_s^2)$ and Taylor expansion, it suffices to show that

$$E(\sup_{t<1}|\mathbf{W}_t - (J^n\mathbf{W})_t|^2) \leqslant \text{const} \times 2^{-n}.$$

Consider

$$\begin{aligned} & E(\sup_{t<1}|\mathbf{W}_t - (J^n\mathbf{W})_t|^2) \\ & = E(\sup_{\substack{1<k<2^n \\ t\in\Delta_k}}|\mathbf{W}_t - \mathbf{W}_{(k-1)/2^n}|^2) \\ & = E(\sup_{\substack{1<k<2^n \\ t\in\Delta_k}}|\mathbf{W}_{t-(k-1)/2^n}|^2) \\ & = E(\sup_{0<t<2^{-n}}|\mathbf{W}_t|^2) \\ & = E\left(\sup_{0<t<2^{-n}}\left|\int_0^t d\,\mathbf{W}_s\right|^2\right) \\ & \leqslant 2\left(E\left(\int_0^{2^{-n}} ds\right) + E\left(\left(\int_0^{2^{-n}}|\mathbf{W}_s|(1-s)ds\right)^2\right)\right) \\ & \leqslant 6\times 2^{-n}, \end{aligned}$$

where we used (2.3) and (2.4) in the next to last inequality.

Altogether we have found that for u defined as in the hypothesis of Proposition 3.4 the following estimate holds:

$$E(\sup_{t<u}\|X_t - X_t^n\|^2) \leqslant \text{const} \times 2^{-n}.$$

Chebyshev's inequality and the Borel–Cantelli lemma imply now the convergence of X_t^n to X_t uniformly in $t \leqslant u$ as $n \to \infty$. But, for a.e. ω, we have $u = 1$. This follows from the boundedness of the coefficient functions and the continuity properties of Q_t and X_t. We formulate this result in the following

Theorem 3.6: Let $\mathbf{A}$ be a bounded C^2 function with bounded first and second derivatives; then X_t^n (3.4) converges with probability one to the solution X_t of the integral equation

$$X_t = \mathbb{1} + \int_0^t d\,\mathbf{W}_s\cdot\mathbf{A}_s X_s + \frac{1}{2}\int_0^t ds(\nabla\cdot\mathbf{A}_s + \mathbf{A}_s^2)X_s, \tag{3.10}$$

the convergence being uniform in $t\in[0,1]$. The solution of (3.10) is called stochastic product integral or stochastic path ordered exponential, denoted $\overleftarrow{\Pi}_{s<t}\exp \mathbf{A}_s\cdot d\mathbf{W}_s$.

IV. THE WIENER–ITÔ REPRESENTATION

In this last section we shall discuss an application of the results of Sec. III.

For the rest of the paper $\mathbf{A}$ will denote a D-tuple of maps from $\mathbb{R}^D$ into the Banach space of anit-Hermitian $m\times m$ matrices. (The results carry over to the case of real skew-symmetric matrices.) Define the operator $\Delta_{\mathbf{A}} = \Sigma_{\mu=1}^D(\partial/\partial x_\mu + A_\mu)^2$ on the L^2-space of functions on $\mathbb{R}^D$ taking values in $\mathbb{C}^m$ (resp. $\mathbb{R}^m$, in the skew-symmetric case). We quote the

following theorem of Schechter,[20] formulated for scalar **A**, generalized to the matrix-valued situation by Schrader.[6]

Theorem 4.1: Let **A** be such that

(i) $A_\mu \in L^4_{\text{loc}}, \quad 1 \leqslant \mu \leqslant D,$

(ii) $\nabla \cdot \mathbf{A} \in L^2_{\text{loc}},$

(iii) $$\sup_x \int_{|\mathbf{x}-\mathbf{y}|<1} \|\mathbf{A}(\mathbf{y})\| \, |\mathbf{x}-\mathbf{y}|^{-D+1} d^D y < \infty.$$

Then $\Delta_{\mathbf{A}}$ is nonpositive on $L^2(\mathbb{R}^D, \mathbb{C}^m)$ and essentially self-adjoint on $C_0^\infty(\mathbb{R}^D, \mathbb{C}^m)$.

Consider the contraction semigroup $\exp t\Delta_{\mathbf{A}}$, $t \geqslant 0$. By standard methods, e.g., Refs. 1, 21, and 22, we have

$$(\exp t\Delta_{\mathbf{A}})(\mathbf{x},\mathbf{y}) = \lim_{n\to\infty} \int dP^t_{\mathbf{xy}} X^n_T(\mathbf{Z}),$$

$$X^n_t(\mathbf{Z}) = \prod_{m=1}^{2^n} \exp\left[\tfrac{1}{2}(\mathbf{A}(\mathbf{Z}_{mt/2^n}) + \mathbf{A}(\mathbf{Z}_{(mt-t)/2^n})) \cdot (\mathbf{Z}_{mt/2^n} - \mathbf{Z}_{(mt-t)/2^n})\right] \tag{4.1}$$

as an equality of kernels of operators on $L^2(\mathbb{R}^D, \mathbb{C}^m)$, whenever the limit exists.

Using now relation (1.2) it is easy to see that the results of Sec. III carry over to $X^n_t(\mathbf{Z})$; i.e., by Theorem 3.2 $X^n_t(\mathbf{Z})$ converges as $n \to \infty$ to the solution X_t of the equation

$$X_t = \mathbb{1} + \int_0^t d\mathbf{Z}_s \cdot \mathbf{A}_s X_s + \frac{1}{2}\int_0^t ds(\nabla \cdot \mathbf{A}_s + \mathbf{A}_s^2) X_s \tag{4.2}$$

if A is a bounded C^2-function with bounded first and second derivatives.

Furthermore we have $\|X_t\| \leqslant 1$, since **A** is anti-Hermitian (resp. skew symmetric), so that by Lebesgue's dominated convergence theorem the rhs of (4.1) converges to $\int dP_{\mathbf{xy}} X_t(\mathbf{Z})$.

It is easy now to extend this representation to continuous **A** by the following standard argument[1,6]: let $\mathbf{A}_n$ be a sequence of smooth functions converging to **A** in L^p_{loc}, p as remarked after Theorem 2.2, and let X_t, X^n_t resp., denote the solution of (4.2) with the corresponding coefficients. Then $\Delta_{\mathbf{A}_n}$ converges to $\Delta_{\mathbf{A}}$ in strong resolvent sense, hence the semigroup $\exp t\Delta_{\mathbf{A}_n}$ converges strongly to $\exp t\Delta_{\mathbf{A}}$. An argument parallel to the proof of Proposition 3.4 (with coefficient functions multiplied by the characteristic function of a large ball) shows that X^n_t converges to X_t, hence $\int dP^t_{\mathbf{xy}} X^n_t(\mathbf{Z})$ converges to $\int dP^t_{\mathbf{xy}} X_t(\mathbf{Z})$ by the dominated convergence theorem.

Theorem 4.2: Let **A** be a continuous anti-Hermitian (resp. skew-symmetric) matrix, such that div **A** is continuous. Then we have the representation

$$(\exp t\Delta_{\mathbf{A}})(\mathbf{x},\mathbf{y}) = \int dP^t_{xy} \prod_{s<t} \exp \mathbf{A}_s \cdot d\mathbf{Z}_s, \tag{4.3}$$

where $\Pi_{s<t} \exp \mathbf{A} \cdot d\mathbf{Z}_s$ denotes the solution of (4.2).

This theorem has an obvious

Corollary: Denoting by Δ the Laplace operator in $\mathbb{R}^D$, then we have the inequalities

$$\|(\exp t\Delta_{\mathbf{A}})(x,y)\| \leqslant (\exp t\Delta)(x,y), \tag{4.4a}$$

$$\|(m^2 - \Delta_{\mathbf{A}})^{-1}(x,y)\| \leqslant (m^2 - \Delta)^{-1}(x,y), \tag{4.4b}$$

for nonzero, real m.

Inequalities (4.5) are called Kato's inequalities or diamagnetic inequalities; cf. also Refs. 1, 6, 20, and 23–25.

ACKNOWLEDGMENTS

It is a pleasure to thank P. A. Meyer, W. Plass, H. Rost, E. Seiler, and particularly Ph. Blanchard for helpful discussions.

[1] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).
[2] Yu. L. Daleckii, "Continual integrals associated with certain differential equations and systems," Sov. Math. Dokl. **2**, 259–269 (1961).
[3] Z. Haba, "Feynman–Kac formula for Green functions and determinants in Euclidean gauge theories," Wroclaw Preprint 519, 1980.
[4] J. Potthoff, "Euclidean φ^4_3 theory in electromagnetic potential," Bielefeld preprint, 1981 (to appear in Ann. Inst. Henri Poincaré).
[5] J. Potthoff, "Euclidean φ^4_3 theory in an external Yang–Mills field," Thesis, Freie Universität, Berlin, 1980.
[6] R. Schrader, "Towards a constructive approach of a gauge invariant, massive $P(\phi_2)$ theory," Commun. Math. Phys. **58**, 299–312 (1978).
[7] M. Ibero, "Intégrales stochastiques multiplicatives et construction des diffusion sur un groupe de Lie," Bull. Sci. Math. **100**, 175–191 (1976).
[8] H. P. McKean, "Brownian motion on the 3-dimensional rotation group," Mem. Coll. Sci. Kyôto Univ. **33**, 25–38 (1960).
[9] H. P. McKean, *Stochastic Integrals* (Academic, New York, 1969).
[10] M. Emergy, "Stabilité des solutions des équations differentielles stochastique, application aux integrales multiplicatives stochastiques," Z. Wahrscheinlichkeitstheorie **41**, 241–262 (1978).
[11] Th. Jeulin, *Semimartingales et grossisement d'une filtration, Springer Lecture Notes in Mathematics*, Vol. 833 (Springer, Berlin, 1980).
[12] H. Métivier and J. Pellaumail, *Stochastic Integration* (Academic, New York, 1980).
[13] Séminaire de Probabilités, Université de Strassbourg, *Springer Lecture Notes in Mathematics*, Vols. X–XIV (Springer-Verlag, Berlin, 1965, 1975, 1966, 1966, 1966).
[14] **W** presents a D-dimensional Wiener path starting at time zero at the origin, ending there at time one.
[15] To avoid endless repetitions, in the whole paper statements are understood to hold with probability one (a.s).
[16] A semimartingale is a process which admits a decomposition into the sum of a local martingale and a process of finite variation.
[17] $\cdot$ denotes the scalar product in d dimensions.
[18] $u \wedge v := \min(u,v)$.
[19] It is not hard to see that Itô's lemma (see, e.g., Ref. 9) holds for W too.
[20] M. Schechter, "Essential selfadjointness of the Schrödinger operator with magnetic vector potential," J. Funct. Anal. **20**, 93–104 (1975).
[21] D. G. Babitt, "Wiener integral representation for certain semigroups which have infinitesimal generators with matrix coefficients," J. Math. Mech. **19**, 1051–1067 (1970).
[22] E. B. Dynkin, *Markov Processes* (Academic, New York, 1965).
[23] D. C. Brydges, J. Fröhlich, and E. Seiler, "On the construction of quantized gauge fields," Ann. Phys. **121**, 227 (1977).
[24] H. Hess, R. Schrader, and D. Uhlenbrock, "Domination of semigroups and generalization of Kato's inequality," Duke Math. J. **44**, 833–904 (1977).
[25] B. Simon, "Abstract Kato's inequality and the comparison of positivity preserving semigroups," Indiana Math. J. **26**, 1067–1073 (1977).

Path integrals in parametrized theories: Newtonian systems

James B. Hartle
Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637

Karel V. Kuchař
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 10 March 1983; accepted for publication 13 May 1983)

Constraints in dynamical systems typically arise either from gauge or from parametrization. We study Newtonian systems moving in curved configuration spaces and parametrize them by adjoining the absolute time and energy as conjugate canonical variables to the dynamical variables of the system. The extended canonical data are restricted by the Hamiltonian constraint. The action integral of the parametrized system is given in various extended spaces: Extended configuration space or phase space and with or without the lapse multiplier. The theory is written in a geometric form which is manifestly covariant under point transformations and reparametrizations. The quantum propagator of the system is represented by path integrals over different extended spaces. All path integrals are defined by a manifestly covariant skeletonization procedure. It is emphasized that path integrals for parametrized systems characteristically differ from those for gauge theories. Implications for the general theory of relativity are discussed.

PACS numbers: 03.20. + i, 02.30. + g, 03.65. − w

1. MOTIVATION

The most straightforward way to describe an evolving classical system is to give its true dynamical degrees of freedom q^a, p_a, $a = 1,...,n$, as functions of the physical time t. The most straightforward way to describe an evolving quantum system is to give its state ψ on the physical configuration space as a function of the physical time.

The actual classical path of the system extremizes the action functional

$$s[q(t)] = \int_{t'}^{t''} dt\, l(t,q,d_t q) \tag{1.1}$$

in configuration space or the canonical action functional

$$s[q(t), p(t)] = \int_{t'}^{t''} dt\, (p_a\, d_t q^a - h(t,q,p)) \tag{1.2}$$

in phase space. In quantum theory, the state function $\psi(t',q')$ at t' is evolved into the state function $\psi(t'',q'')$ at t'' by the quantum propagator $\langle t'',q''|t',q'\rangle$,

$$\psi(t'',q'') = \int d^n q' \langle t'',q''|t',q'\rangle \psi(t',q'). \tag{1.3}$$

The connection between quantum theory and the classical theory is established when we represent the quantum propagator as an integral over all paths connecting t',q' with t'',q'' in the configuration space,[1]

$$\langle t'',q''|t',q'\rangle\, d^n q' = \int \bar{D}q\, e^{is[q(t)]}, \tag{1.4}$$

or as an integral over all paths connecting t',q' with t'',q'' in the phase space,

$$\langle t'',q''|t',q'\rangle\, d^n q' = \int Dq\, Dp\, e^{is[q(t),\, p(t)]}. \tag{1.5}$$

The transition from classical theory to quantum theory thus amounts to an interpretation of the formal expressions (1.4) or (1.5). To do that, one must explain what is meant by integrating the exponentiated classical action functionals (1.1) or (1.2) and what are the measures $\bar{D}q$ or $Dq\,Dp$ in the space of paths. Both problems can be solved by a skeletonization procedure. In configuration space, the skeletonization of the action functional is obvious: $s[q(t)]$ is replaced by a sum of Hamilton's principal functions for individual segments of the skeletonized path. However, the choice of the skeletonized measure is not obvious. One can use different measures and these measures yield different propagators.[1] This ambiguity corresponds exactly to factor ordering in Hamiltonian quantum mechanics: The Hamilton operators in Schrödinger's equation for the propagators differ by curvature terms of the order $\hbar^2$.

In the phase space path integral, the situation is reversed. The invariant Liouville measure $d^n q\, d^n p$ in the phase space induces a natural measure in the space of skeletonized paths. On the other hand, the skeletonization of the canonical action (1.2) by a sum of phase space principal functions is not unique.[2] Different principal functions yield the same classical dynamics but because nondifferentiable paths are the most significant contributors to the path integral (1.5) they do not yield equivalent quantum dynamics. The advantage of Eq. (1.5) over Eq. (1.4) is that the measure is fixed, and the ambiguity is shifted to the skeletonization of the canonical action where it can be resolved by applying geometric criteria.

The clarity achieved by formulating a physical theory in terms of its true dynamical degrees of freedom is often at the expense of obscuring its fundamental symmetries. Examples of this statement are found in gauge theories and in parametrized theories. The symmetries in these two cases are, of course, gauge invariance and parametrization invariance. For complicated gauge and parametrized systems, it is often impractical or even impossible to exhibit the true dynamical degrees of freedom explicitly. It is thus imperative to have a procedure for passing from the classical version to the quantum version of the theory in its symmetry revealing form. We shall briefly discuss one example of a gauge system

 0022-2488/84/010057-19$02.50

and one example of a parametrized system to get a feeling for the problem.

Though one can easily concoct finite-dimensional gauge theories, the best known specimen of gauge theories is a field system, namely, Maxwell's electrodynamics. The gauge invariance and the Lorentz invariance of this theory are readily seen when the field action is expressed as a functional of the 4-potential $A_\alpha(x)$, $\alpha = 0,1,2,3$. Due to gauge invariance the variables A_α are redundant. However, it is extremely cumbersome to describe the field by its two physical degrees of freedom which are the transverse components of $A_\alpha(x)$, especially in the presence of interactions. The desirability of a quantization procedure which operates at the level of unphysical variables $A_\alpha(x)$ is readily seen. For gauge theories such procedures have been extensively developed.[3] The path integrals are of the same form as Eq. (1.4) or (1.5), but the paths lie in the configuration or phase space augmented by gauge variables. The central issue of these formulations is then specifying the measure which reproduces the physical predictions of the theory. This measure is often quite complicated and difficult to guess from first principles.

As a consequence of gauge invariance, the electric field strength $E^a(x)$ cannot be freely specified, but on each spatial hypersurface it is subject to the constraint

$$C(x^a) = \partial_a E^a = 0. \tag{1.6}$$

In the Hamiltonian version of the theory, $E^a(x)$ is the momentum conjugate to the vector potential $A_a(x)$. The constraint (1.6) is the price we pay for the freedom to perform the gauge transformations.

Another important but quite distinct class of theories with internal symmetries are parametrized theories. The invariance with respect to reparametrization is achieved by adjoining the physical time to the dynamical variables of the system. An arbitrary parameter is then used to locate the system on its dynamical path. Any field theory on a given background can be cast into a parametrized form, but the best known example of a parametrized theory is a finite-dimensional system, namely, the free relativistic particle. Let us discuss the canonical version of the theory. The canonical action (1.2) of the particle is expressed as a functional of the spatial coordinates $q^a(t)$ and their conjugate momenta $p_a(t)$ considered as functions of the Minkowskian time t in a given inertial frame:

$$s[q^a(t), p_a(t)] = \int_{t'}^{t''} dt\,(p_a\, d_t q^a - (\delta^{ab} p_a\, p_b + m^2)^{1/2}). \tag{1.7}$$

In the physical variables q^a, p_a and with the fixed parametrization t, it is difficult to discuss the Lorentz invariance and the reparametrization invariance of the theory. However, if we let t be a function of a parameter τ (not necessarily the proper time) and introduce the Minkowskian time $t = q^0(\tau)$ and energy $-p_0(\tau)$ as dynamical variables, we can write the action (1.7) in the form

$$S[q^\alpha(\tau), p_a(\tau)] = \int_{\tau'}^{\tau''} d\tau\, p_\alpha\, \dot{q}^\alpha, \tag{1.8}$$

which is manifestly invariant both under Lorentz transformations and under reparametrizations of paths, $\tau \to \tau^* = \tau^*(\tau)$. The momenta p_α cannot be varied freely, but they must lie on the mass shell,

$$H = (1/2m)(p_\alpha\, p^\alpha + m^2) = 0. \tag{1.9}$$

The equations of motion follow from extremizing the action (1.8) subject to the constraint (1.9). The constraint (1.9) is the counterpart of the constraint (1.6) in electrodynamics. It is a consequence of the reparametrization invariance in the same way as the constraint (1.6) is a consequence of gauge invariance.

The most important and also the most intricate system in which both gauge and parametrization are subtly intertwined is general relativity. It may be studied as a Hamiltonian theory by foliating space-time with a family of spacelike hypersurfaces. The foliation is specified by giving the lapse function $N(x,t)$ and the shift vector $N^a(x,t)$. The lapse function determines the normal proper time separation $d\sigma = N(x,t)dt$ between two nearby spatial hypersurfaces t and $t + dt$ and the shift vector $N^a(x,t)$ tells us how to displace the point x^a on the hypersurface t so that by launching from the displaced point $x^a + N^a\, dt$ in the direction perpendicular to the hypersurface t we land at the point x^a of the deformed hypersurface $t + dt$. The canonical variables $g_{ab}(t,x)$ and $p^{ab}(t,x)$ are the intrinsic metric and the extrinsic curvature of the hypersurface t. The gauge transformations of the theory are spatial diffeomorphisms on the hypersurfaces of the foliation. The reparametrization is connected with the change of the foliation. Invariance of the theory under gauge transformations implies the supermomentum constraint

$$H_a(x) = -2p^b_{a|b} = 0, \tag{1.10}$$

on the canonical data $g_{ab}(x)$, $p^{ab}(x)$; the reparametrization invariance implies the super-Hamiltonian constraint

$$H(x) = g^{-1/2}(p_{ab}p^{ab} - \tfrac{1}{2}p^2) - g^{1/2}R = 0. \tag{1.11}$$

Here, $g(x) = \det g_{ab}(x)$, the vertical stroke denotes the covariant derivative on the hypersurface and R is the curvature scalar on the hypersurface.

The gauge and reparametrization changes together with the constraints (1.10)–(1.11) imply that the metric field has only 2^{∞^3} degrees of freedom, i.e., $2{\cdot}2^{\infty^3}$ physical field coordinates and conjugate momenta. The remaining 2^{∞^3} coordinates and momenta play the role of an internal time which distinguishes one hypersurface from another by looking at its intrinsic geometry or extrinsic curvature, and of an internal energy. Unfortunately, no one knows how to write an action for general relativity which involves only the two physical degrees of freedom expressed as functions of the physical time. The best we can do is to work with the extended variables g_{ab}, p^{ab}. General relativity comes to us directly only in the gauged and parametrized form. This is our strongest motivation for studying the relation between gauge and parametrized theories in an attempt to understand their similarities and differences.

The similarities are obvious. Both types of invariance imply constraints. In electrodynamics, we have the diver-

gence equation (1.6). In the parametrized relativistic particle theory, we have the restriction (1.9) of the 4-momentum to the mass shell, and in general relativity, we have the constraints (1.10) and (1.11). Further, the constraints generate the changes of extended canonical variables under corresponding transformations. In electrodynamics, we smear the constraint $C(x)$ by an arbitrary test function $\Lambda(x)$,

$$C_\Lambda \equiv \int d^3x\, \Lambda(x) C(x). \tag{1.12}$$

The Poisson bracket of C_Λ with the extended phase space variables $A_a(x)$, $E^a(x)$, $a = 1,2,3$, generates their gauge transformation,

$$\begin{aligned} \delta A_a(x) &= [A_a(x), C_\Lambda] = A_a(x) - \partial_a \Lambda(x), \\ \delta E^a(x) &= [E^a(x), C_\Lambda] = 0. \end{aligned} \tag{1.13}$$

Similarly, for the relativistic particle the constraint (1.9) determines the change of the canonical variables x^α, p_α under displacement $\delta\sigma$ in proper time,

$$\delta x^\alpha = [x^\alpha, H]\,\delta\sigma, \quad \delta p_\alpha = [p_\alpha, H]\,\delta\sigma. \tag{1.14}$$

Finally, in general relativity we smear the super-Hamiltonian (1.11) by the lapse function $N(x)$ and the supermomentum (1.10) by the shift vector $N^a(x)$:

$$\begin{aligned} H_N &\equiv \int d^3x\, N(x) H(x), \\ H_{\mathrm{N}} &\equiv \int d^3x\, N^a(x) H_a(x). \end{aligned} \tag{1.15}$$

The Poisson brackets

$$\delta g_{ab}(x) = [g_{ab}(x), H_{\mathrm{N}}]\,\delta t, \quad \delta p^{ab}(x) = [p^{ab}(x), H_{\mathrm{N}}]\,\delta t \tag{1.16}$$

yield the changes of the canonical variables $g_{ab}(x)$, $p^{ab}(x)$ when the point x^a is displaced by amount $\delta x^a = N^a\,\delta t$ along the hypersurface, while the Poisson brackets

$$\delta g_{ab}(x) = [g_{ab}(x), H_N]\,\delta t, \quad \delta p^{ab}(x) = [p^{ab}(x), H_N]\,\delta t \tag{1.17}$$

yield the changes of $g_{ab}(x)$, $p^{ab}(x)$ when the hypersurface is deformed by the amount $N\,\delta t$ in the normal direction.

There is, however, an important physical distinction between gauge theories and parametrized theories. For gauge theories the changes generated by the constraints do not change the physical state of the system. They change only the gauge in which it is represented. The true physical degrees of freedom do not change. So, in electrodynamics, $A_a(x)$ is changed by the transformation (1.13), but the field strengths $E^a(x)$ and $H^a(x)$ remain unaffected. By contrast, in parametrized theories the changes induced by the constraints are those associated with the dynamical evolution of the system. The true physical degrees of freedom are moved along the dynamical path. This is clearly seen in Eq. (1.14) for a free relativistic particle. In general relativity, the changes (1.16) generated by the supermomentum leave the intrinsic geometry and the extrinsic curvature of the hypersurface unaffected. The quantities like $\int d^3x\, g^{1/2} R$ or $\int d^3x\, g_{ab}\, p^{ab}$ stay the same. On the other hand, the super-Hamiltonian generates the dynamical evolution of the spatial geometry and of the extrinsic curvature under the normal deformation of the hypersurface [Eq. (1.17)].

The different roles which the constraints resulting from gauge invariance and those resulting from reparametrization invariance play in classical theory have fundamental consequences for the quantum theory. This is because in quantum theory time is clearly distinguished from all other variables and cannot be represented by a Hermitian operator. As a result, the path integral procedure developed in gauge theories to achieve the transition from classical mechanics to quantum mechanics is not directly applicable to parametrized theories. In this paper we shall build a correct procedure for a simple class of parametrized systems and show how it differs from the prescription developed for gauge theories. An understanding of where the two prescriptions differ would seem an essential prerequisite to understanding the quantization of general relativity by path integrals.

The finite-dimensional theory which we have chosen as our model is a nonrelativistic system described by the Hamiltonian

$$h = \tfrac{1}{2} g^{ab}(p_a - \phi_a)(p_b - \phi_b) + \phi. \tag{1.18}$$

The potentials ϕ and ϕ_a depend on the configuration variables q^a, $a = 1,\ldots,n$ and on absolute time t. A curved nondegenerate metric g^{ab} is also a function of these variables. We study a nonrelativistic theory because it contains an easily and uniquely identifiable time variable. We consider curved configuration spaces because the parametrized version of the theory can be expressed in terms of a degenerate curved metric in $n+1$ dimensions and so bears structural similarity to general relativity which is our ultimate theory of interest. To emphasize this similarity, we shall express our results in a manifestly covariant manner using this extended metric. We can thus clearly exhibit the geometric structure these theories possess.

Our starting point is the path integral (1.5) in the physical phase space with the canonical action (1.2) containing the Hamiltonian (1.18). We interpret this path integral by a manifestly covariant skeletonization procedure which leads to the Schrödinger equation for the quantum propagator without additional curvature term. This choice fixes the factor ordering and the quantum theory. Our ending points are path integrals for the same propagator over associated spaces. The simplest of these is the integral (1.4) over paths in the physical configuration space. More important, however, are path integrals corresponding to the parametrized version of the theory.

We parametrize the system by adjoining time and a conjugate momentum to the variables $\{q^a, p_b\}$, forming thus an enlarged configuration space $\{Q^A\}$, $A = 0,\ldots,n$ and phase space $\{Q^A, P_A\}$. The quantum propagator can be expressed as a path integral in the enlarged phase space or in the enlarged configuration space. Each case divides into two, corresponding to the classical choice of how the constraint connected with reparametrization invariance is enforced. It can be enforced either explicitly on the variations of an action or implicitly using a Lagrange multiplier. This possibility is reflected quantum-mechanically in two forms of the path integral for each space of variables: one in which the action is

free from multipliers but the measure includes a δ function of the constraint and a second in which the action contains a multiplier and the measure includes an integration over it. There are thus four forms of the path integral for parametrized theories with the basic Hamiltonian (1.18). This may seem an unnecessary proliferation of possibilities, but each of the four forms of the classical action corresponding to these choices can be actually constructed in general relativity. They are displayed in Table I. It therefore seems appropriate to consider all of them in the simple nonrelativistic systems under consideration.

Our results are thus the six forms for the path integral for the system described by the Hamiltonian (1.18)—two in terms of physical variables and four in terms of extended variables. They are specified by six actions displayed in Table II (Sec. 3) and by six measures summarized in Table III (Sec. 10). They are six equivalent ways for passing from the classical theory to the quantum theory. None of the parametrized versions of this passage correspond to the standard procedures for quantizing gauge theories. We shall discuss this in detail in Sec. 9. This only underlines once again the depth of the issues involved in quantizing gravity.

2. PARAMETRIZED NEWTONIAN SYSTEMS

Our immediate goal is to reformulate classical dynamics of a Newtonian system in an extended phase space. In this process, absolute time and energy are adjoined as conjugate canonical variables to the dynamical variables of the system. The absolute time loses thereby its privileged role in parametrizing paths, and it is replaced by an arbitrary label time. For this reason, the process is called parametrization. With absolute time lifted among the configuration variables, one can introduce arbitrary coordinates in the configuration space-time. This underscores the geometric content of the parametrized theory. To reduce the theory back to its humble physical origins, one should learn how to identify the original physical variables from the geometric structures and reinstate them as privileged variables into the action. This inverse process is summarily called a deparametrization. Our goal is thereby set: First, parametrize the physical theory and geometrize it; second, deparametrize the geometric theory and return to the physical starting point.

We assume that the Newtonian space-time is endowed by a privileged foliation of hypersurfaces whose leaves are instants of absolute time. We label the hypersurfaces by a parameter t, not necessarily coinciding with the pace of a standard clock. We assume that each hypersurface carries a positive-definite metric. We do not insist, however, that this metric be flat or time-independent. We introduce into the space-time an arbitrary congruence of world lines transversal to the time foliation and label them by three coordinates. The congruence represents a choice of reference frame.

The dynamical system which we have in mind might be a single point particle or a system of such particles subject to holonomic though in general rheonomic (time-dependent) constraints. (These constraints have nothing to do with the Hamiltonian constraint we introduce later.) Knowing the masses of the particles and the constraints to which they are subject, we can express the kinetic energy of the system in terms of the generalized coordinates q^a, $a = 1,2,...,n$, and generalized velocities $\dot{q}^a$ and deduce thus the instantaneous metric $g_{ab}(t,q)$ induced in the configuration space $\{q^a\}$ of the system. The system is also subject to forces derivable from a scalar potential $\phi(t,q)$ and a vector potential $\phi_a(t,q)$. We do not need to distinguish "true" forces from "ficticious" forces, which are already contained in the expression for the

TABLE I. Alternative forms of the action for general relativity.

	Canonical variables	Multipliers	Action	Lagrangian, Hamiltonian, constraints
Extended canonical action, conditional	g_{ab}, p^{ab}		$S[g_{ab}, p^{ab}] = \int dt \int d^3x\, p^{ab}\dot{g}_{ab}$,	$H = g^{-1/2}(p_{ab}p^{ab} - \frac{1}{2}p^2) - g^{1/2}R$
			$H(x) = 0 = H_a(x)$	$H_a = -2p^b_{a\|b}$
Extended canonical action, with lapse and shift multipliers	g_{ab}, p_{ab}	N, N^a	$S[g_{ab}, p^{ab}, N, N^a] = \int dt \int d^3x(p^{ab}\dot{g}_{ab} - H(g_{ab}, p^{ab}, N, N^a))$	$H = N(x)H(x) + N^a(x)H_a(x)$
Extended Lagrangian action, with lapse and shift multipliers	g_{ab}	N, N^a	$S[g_{ab}, N, N^a] = \int dt \int d^3x L(g_{ab}, \dot{g}_{ab}, N, N^a)$	$L = Ng^{1/2}[(K_{ab}K^{ab} - K^2) + R] = (-{}^4g)^{1/2}\,{}^4R$ + (divergence terms)
				$K_{ab} = \frac{1}{2}N^{-1}(-\dot{g}_{ab} + N_{a\|b} + N_{b\|a})$
Extended Lagrangian action, without the lapse multipliers[a]	g_{ab}	N^a	$S[g_{ab}, N^a] = \int dt \int d^3x\, L(g_{ab}, \dot{g}_{ab}, N^a)$	$L = [gR(U_{ab}U^{ab} - U^2)]^{1/2}$
				$U_{ab} = \dot{g}_{ab} - N_{a\|b} - N_{b\|a}$

[a] The elimination of N^a would lead to the homogeneous Lagrangian action without multipliers. The elimination cannot be carried out explicitly.

kinetic energy. We thus include both types of terms into the potentials ϕ, ϕ_a.

An elementary example of such a system would be a charged particle moving on an expanding curved surface placed in an external electromagnetic field. The generalized coordinates q^a might be any curvilinear coordinates on the surface. Another example, closer to actual systems studied in nonrelativistic quantum mechanics, would be a charged rigid rotator in an external electromagnetic field. The generalized coordinates q^a might be the Euler angles. The kinetic energy of the rigid rotator expressed as a quadratic form of generalized velocities indicates that the configuration space of the rotator is curved, but the metric is time-independent.

The dynamical evolution of the system takes place in the physical phase space $\{q^a, p_a\}$ which is a cotangent bundle over the physical configuration space $\{q^a\}$. The evolution of physical variables is governed by the canonical action

$$s[q,p] = \int dt\,(p_a\, d_t q^a - h(t,q,p)) \tag{2.1}$$

with the Hamiltonian

$$h(t,q,p) = \tfrac{1}{2} g^{ab}(p_a - \phi_a)(p_b - \phi_b) + \phi. \tag{2.2}$$

A new choice of the time labeling, $t^* = t^*(t)$, or a change of the reference frame changes the Hamiltonian (2.2) into another Hamiltonian of the same type. The only features of the Newtonian system which are important for our purposes are the existence of a privileged foliation of the configuration space-time by leaves of absolute time and the fact that the Hamiltonian of the system is a quadratic function of canonical momenta. There is no need to introduce other features usually associated with Newtonian physics like the presence of the Galilei group.

We now parametrize a possible path along which the system moves in the phase space $\{q^a, p_a\}$ by an arbitrary label time τ and adjoin the originally chosen absolute time $t(\tau)$ to the configuration variables $q^a(\tau)$:

$$Q^A = \{t, q^a\}, \quad Q^A = Q^A(\tau), \quad A = 0,1,2,...,n. \tag{2.3}$$

The action (2.1) takes the form

$$s[Q^A, p_a] = \int d\tau\,(p_a \dot{q}^a - h(Q,p)\dot{t}) \tag{2.4}$$

when written in the τ-parametrization. The dot denotes a derivative with respect to the label time τ. Numerically, the expression (2.4) is equal to the expression (2.1) and so variation with respect to q^a, p_a yields equivalent equations of motion. Moreover, the variation of the parametrized action (2.4) with respect to t also yields a correct equation, namely, the energy balance equation

$$\dot{h} = \partial_t h\, \dot{t}. \tag{2.5}$$

The integrand of the action functional (2.4) is linear in the velocities $\dot{Q}^A = \{\dot{t}, \dot{q}^a\}$. By introducing a momentum $p_0 = -h$ canonically conjugate to t and by putting

$$P_A = \{p_0, p_a\} \tag{2.6}$$

we cast the action (2.4) into a suggestive form

$$S[Q^A, P_A] = \int d\tau\, P_A \dot{Q}^A. \tag{2.7}$$

However, the variables P_A cannot be varied freely, because p_0 is a mere abbreviation for the function $-h(Q^A, p_a)$. To obtain correct equations of motion, we must vary the action (2.7) under the constraint

$$H^{(o)} \equiv p_0 + h(Q^A, p_a) = 0. \tag{2.8}$$

In this way, a constraint on the variables of the enlarged phase space enters into the theory. It is called the Hamiltonian constraint.

The actual path of the system extremizes the action functional (2.7) in comparison to all neighboring paths which lie on the constraint surface (2.8). In other words, the actual path is selected by the conditions

$$H^{(o)}(Q,P) = 0,$$
$$\delta S[Q,P] = 0 \quad \forall\ \delta Q,\ \delta P\colon\ \delta H^{(o)} = 0. \tag{2.9}$$

Equations (2.9) constitute a conditional variational principle.

The constraint function $H^{(o)}$ is a quadratic function of extended momenta P_A. This property is preserved if we multiply the constraint by an arbitrary function $\Lambda(Q^A) > 0$ of extended configuration variables,

$$H \equiv \Lambda(Q) H^{(o)}, \quad H = 0. \tag{2.10}$$

The constraint function $H(Q,P)$ is called a super-Hamiltonian of the system.

We shall now write the constraints (2.8) or (2.10) in a manifestly covariant notation. We introduce a covector field

$$t_A = t_{,A}(Q) = (1;0,...,0) \tag{2.11}$$

normal to the instants of absolute time and a vector field

$$u^A = (1;0,...,0) \tag{2.12}$$

tangent to the world lines $q^a = \text{const}$ of our "configuration reference frame." We collect the potentials into a space-time covector field

$$\phi_A = (-\phi;\phi_a) \tag{2.13}$$

and complete the spatial metric g^{ab} into a degenerate space-time metric

$$g^{AB} = \begin{vmatrix} 0, & 0 \\ 0, & g^{ab} \end{vmatrix}. \tag{2.14}$$

The metric g^{AB} has the signature $(0; +,..., +)$ and t_A is its degeneracy direction,

$$g^{AB} t_B = 0. \tag{2.15}$$

Of course,

$$u^A t_A = 1. \tag{2.16}$$

When g^{AB} and u^A are given, Eqs. (2.15) and (2.16) determine t_A.

The super-Hamiltonian (2.8) can now be written in a manifestly covariant form

$$H^{(o)} = u^A(P_A - \phi_A) + \tfrac{1}{2} g^{AB}(P_A - \phi_A)(P_B - \phi_B). \tag{2.17}$$

After scaling the fields u^A and g^{AB} by the factor $\Lambda(Q)$,

$$G^{AB} \equiv \Lambda g^{AB}, \quad U^A \equiv \Lambda u^A, \tag{2.18}$$
$$H = U^A(P_A - \phi_A) + \tfrac{1}{2} G^{AB}(P_A - \phi_A)(P_B - \phi_B). \tag{2.19}$$

Up to now, the absolute time variable $Q^0 = t$ was clearly separated from the configuration variables $Q^a = q^a$. At this stage, however, we can easily mix the space-time variables Q^A by an arbitrary transformation $Q^{A*}(Q^B)$, inducing thereby a transformation of the conjugate momenta:

$$Q^{A*} = Q^{A*}(Q^B), \quad P_{A*} = Q^B{}_{A*} P_B, \tag{2.20}$$

$$Q^{A*}{}_B \equiv \frac{\partial Q^{A*}}{\partial Q^B}, \quad Q^B{}_{A*} \equiv \frac{\partial Q^B}{\partial Q^{A*}}.$$

When we transform U^A (or u^A) as a vector, G^{AB} (or g^{AB}) as a tensor, and ϕ_A as a covector, the constraint (2.19) [or (2.17)] preserves its form. We shall omit the asterisks with the understanding that Eqs. (2.17)–(2.19) are written in general coordinates. The action principle (2.9) then yields the actual motion of the system in general coordinates.

In the special coordinates $Q^A = \{t, q^a\}$, the coefficients u^A, g^{AB} assume the simplified form (2.12),(2.14). This implies that the scaled coefficients U^A, G^{AB} cannot be arbitrary functions of general coordinates Q^A. In a permissable parametrized Newtonian theory, U^A and G^{AB} must be subject to two sets of restrictions which ensure that the physical theory can be recovered by deparametrization. These restrictions are:

(I) The metric G^{AB} must be degenerate, with signature $(0; +,..., +)$. The degeneracy direction T_A,

$$G^{AB} T_B = 0, \quad T_B \neq 0, \tag{2.21}$$

must be surface-forming. This happens if and only if the metric G^{AB} satisfies the integrability condition (Appendix A)

$$\delta_{AA_1 \cdots A_n} G^{A[B,C]} G^{A_1 B_1} \cdots G^{A_n B_n} = 0. \tag{2.22}$$

(II) The inner product $U^A T_A$ cannot vanish and, for a future-oriented T_A, it must be positive,

$$U^A T_A > 0. \tag{2.23}$$

Equation (2.23) implies that T_A can be normalized so that

$$U^A T_A = 1. \tag{2.24}$$

The parametrized Newtonian system is characterized by a quadratic super-Hamiltonian (2.19) whose coefficients U^A and G^{AB} satisfy our restrictions (I) and (II). We complete our demonstration that the physical and parametrized versions of the theory are equivalent by showing how to deparametrize the system. To do this, we have to find the absolute time function and return back to the physical Hamiltonian (2.2).

Notice first that the quadratic function (2.19) determines the coefficient G^{AB} uniquely, but the coefficients U^A and ϕ_A only up to a gauge transformation

$$*U^A = U^A + G^{AB}\psi_B, \qquad *\phi_A = \phi_A + \psi_A, \tag{2.25}$$

generated by a gauge variable ψ_A which satisfies the condition

$$\tfrac{1}{2} G^{AB}\psi_A \psi_B + U^A \psi_A = 0. \tag{2.26}$$

The transformation (2.25)–(2.26) expresses an arbitrary change of the configuration reference frame. We have discussed the influence of such a gauge transformation on quantum description of a Newtonian system in an earlier paper.[4] Here, we shall simply assume that one reference field U^A is chosen within the equivalence class (2.25)–(2.26).

Return now to the problem of how to reconstruct the physical Hamiltonian. For the metric G^{AB} with signature $(0; +,..., +)$ all solutions T_A of Eq. (2.21) fill a ray. The integrability condition (2.22) ensures that at least one solution t_A within this ray is a gradient of a scalar function,

$$\exists\ t(Q):\ t_A = t_{,A}. \tag{2.27}$$

In fact, all solutions which are gradients are related to one another by the transformations $t^* = t^*(t)$. We select one which increases to the future, i.e., which satisfies the condition

$$U^A t_{,A} \equiv \Lambda(Q) > 0 \tag{2.28}$$

for our time function $t(Q)$. We then scale the super-Hamiltonian (2.19) down by the factor Λ^{-1}, scaling G^{AB} down to g^{AB} and U^A to u^A by Eq. (2.18). Equation (2.28) then implies Eq. (2.16). Of course, the scaled metric satisfies Eq. (2.15).

We can now introduce within the reference frame u^A comoving coordinates q^a as any n functionally independent solutions $q^a(Q)$ of the equations

$$u^A q^a{}_{,A} = 0. \tag{2.29}$$

We take the time function (2.27) and the comoving coordinates (2.29) as our special coordinates $Q^A = \{t, q^a\}$. Equations (2.16) and (2.29) then ensure that u^A in special coordinates has the components (2.12). Similarly, Eq. (2.15) ensures that the rescaled metric g^{AB} has the components (2.14). Therefore, the rescaled super-Hamiltonian (2.17) reduces back to the form (2.8), where h is our old Hamiltonian (2.2). When we solve the constraint (2.8) with respect to p_0, substitute this solution into the action (2.7), and parametrize paths by the absolute time t, we return back to the physical action (2.1). In this way, we regain the physical action from the parametrized action (2.7) subject to the super-Hamiltonian constraint (2.8).

3. ALTERNATIVE FORMS OF THE ACTION

We have transformed the canonical action (2.1)–(2.2) on the physical phase space into a constrained action (2.7)–(2.10), (2.17), (2.19) on the extended phase space. Besides these forms of the action, there are still others which are frequently used in dynamical considerations. In particular, one can adjoin the Hamiltonian constraint (2.10) to the extended phase space action (2.7) by a lapse multiplier, and one can cast the parametrized action into a Lagrangian form, either on the physical or on the extended configuration space, and either including or excluding the lapse multiplier.

We have argued in the Introduction that any of these forms could serve as the starting point for the transition to quantum theory by path integrals. However, only in the physical phase space do we have a universal prescription for the measure. All other path integrals should be thus derived from the path integral in the physical phase space. To proceed, we must first understand how the various forms of the action are connected to each other. We shall study this clas-

sical problem now and postpone its application to path integrals to subsequent sections.

In the beginning, we replace the conditional variational principle by a free variational principle by adjoining the constraint (2.17) to the action (2.7) by a Lagrange multiplier $N^{(o)}$,

$$S[Q,P,N^{(o)}] = \int d\tau\,(P_A \dot{Q}^A - N^{(o)} H^{(o)}), \tag{3.1}$$

or the scaled constraint (2.19) by a Lagrange multiplier N,

$$S[Q,P,N] = \int d\tau\,(P_A \dot{Q}^A - NH). \tag{3.2}$$

All the variables Q^A, P_A, $N^{(o)}$ or Q^A, P_A, N may now be varied freely.

The physical meaning of the multipliers $N^{(o)}$ or N follows from the Euler–Lagrange equations. By varying Eq. (3.2) in the momenta P_A, we get

$$\dot{Q}^A = N(U^A + G^{AB}(P_B - \phi_B)). \tag{3.3}$$

We multiply Eq. (3.3) by a degeneracy covector T_A, Eqs. (2.21), (2.23), and calculate N:

$$N = (T_B U^B)^{-1} T_A \dot{Q}^A. \tag{3.4}$$

In the special coordinates $Q^A = \{t, q^a\}$, Eq. (3.4) reduces to

$$N = \Lambda^{-1}\dot{t} \tag{3.5}$$

by virtue of Eq. (2.28). The same sequence of steps starting from the action (3.1) leads to the equation

$$N^{(o)} = (t_{,B} u^B)^{-1} t_{,A} \dot{Q}^A = \dot{t}. \tag{3.6}$$

We thus see that the multiplier $N^{(o)}$ equals the rate of change $\dot{t}$ of the absolute time t with respect to the label time τ. For this reason, it is called the lapse function. We shall loosely use this name also for the scaled multiplier (3.5).

The action (3.2) is the best starting point for further rearrangements. We group its arguments into several classes:

extended configuration variables Q^A

$= \{$physical time t, physical coordinates $q^a\}$,

extended momenta variables P_A

$= \{$physical Hamiltonian $- p_0$, physical momenta $p_a\}$,

Lagrange multiplier $= \{$lapse function $N\}$.

By eliminating one or more classes of variables from the action, we cast it into a number of alternative forms which lead to equivalent sets of equations of motion. The transition from the extended action (3.2) to the physical action (2.1) has this character: It is achieved by using the equations of motion to eliminate the lapse multiplier N and the time–energy pair t, p_0 from the action. One can proceed one step further and eliminate all momenta variables from the canonical action (2.1). One arrives then at the physical Lagrangian action

$$S[q] = \int dt\, l(t,q,d_t q) \tag{3.7}$$

by the Legendre dual transformation

$$l(t,q,d_t q) = [\,p_a\, d_t q^a - h(t,q,p)\,]_{p = p(t,q,d_t q)}, \tag{3.8}$$

$$d_t q^a = \frac{\partial h}{\partial p_a}.$$

Because the physical Hamiltonian is nondegenerate, the second equation uniquely determines the generalized momenta p_a in terms of the generalized velocities $d_t q^a$, $p_a = p_a(t,q,d_t q)$. For the Newtonian system (2.2),

$$l(t,q,d_t q) = \tfrac{1}{2} g_{ab} d_t q^a d_t q^b + \phi_a d_t q^a - \phi,$$
$$p_a = g_{ab} d_t q^b + \phi_a. \tag{3.9}$$

Start now from the parametrized canonical action (3.2) instead of from the physical canonical action (2.1). Try to eliminate the momenta P_A, but leave the lapse function N in the action. This time, however, the expression (3.3) for the velocities $\dot{Q}^A$ in terms of the momenta P_A is not invertible because the metric G^{AB} is degenerate. One can, however, go most of the way by defining the covariant metric G_{AB} by the equations

$$U^B G_{BA} = 0, \quad G^{AB} G_{BC} = \delta^A_C - U^B T_C, \tag{3.10}$$

where T_C is the normalized degeneracy covector (2.21), (2.24). The metric G_{AB} is again degenerate, with signature $(0; +,\ldots, +)$. After introducing the abbreviations

$$\phi_{\|} \equiv \phi_A U^A, \quad P_{\|} \equiv P_A U^A, \tag{3.11}$$

we express the momenta P_A in terms of the velocities $\dot{Q}^A$ and a single scalar $P_{\|}$

$$P_A = N^{-1} G_{AB} \dot{Q}^B + (\phi_A + \phi_{\|} T_A) + P_{\|} T_A. \tag{3.12}$$

After the Legendre transformation

$$L \equiv [P_A \dot{Q}^A - NH]_{P_A = P_A(Q,\dot{Q},N,P_{\|})}$$
$$= \tfrac{1}{2} N^{-1} G_{AB} \dot{Q}^A \dot{Q}^B + N\phi_{\|} + (\phi_A - \phi_{\|} T_A)\dot{Q}^A$$
$$+ P_{\|}(T_A \dot{Q}^A - N) \tag{3.13}$$

$P_{\|}$ stays in the action as another Lagrange multiplier. However, it can be eliminated by using the Euler–Lagrange equation obtained by varying the lapse multiplier N,

$$P_{\|} - \phi_{\|} + \tfrac{1}{2} N^{-2} G_{AB} \dot{Q}^A \dot{Q}^B = 0. \tag{3.14}$$

This leads to the Lagrangian

$$L(Q,\dot{Q},N) = (N^{-1} - \tfrac{1}{2} N^{-2} T_C \dot{Q}^C) G_{AB} \dot{Q}^A \dot{Q}^B + \phi_A \dot{Q}^A \tag{3.15}$$
$$= -\tfrac{1}{2}(N^{-1}(T_C \dot{Q}^C)^{1/2} - (T_C \dot{Q}^C)^{-1/2})^2 G_{AB} \dot{Q}^A \dot{Q}^B$$
$$+ \tfrac{1}{2}(T_C \dot{Q}^C)^{-1} G_{AB} \dot{Q}^A \dot{Q}^B + \phi_A \dot{Q}^A. \tag{3.16}$$

It is not difficult to check that by varying Q^A and N we obtain correct equations of motion. In special coordinates $Q^A = \{t, q^a\}$ with the lapse function $N^{(o)} = \Lambda N$ the Lagrangian (3.15) reduces to

$$L(t,q,\dot{q},N^{(o)}) = (N^{(o)\,-1} - \tfrac{1}{2} N^{(o)\,-2}\dot{t})$$
$$\times g_{ab} \dot{q}^a \dot{q}^b + \phi_a \dot{q}^a - \phi \dot{t}. \tag{3.17}$$

As a final transformation, we eliminate the lapse function N from the extended Lagrangian (3.16). The Euler–Lagrange equation obtained by varying N can be solved for N, with the result

$$N = T_A \dot{Q}^A. \tag{3.18}$$

This expression replicates Eq. (3.4) which was obtained from the canonical action. By substituting it back into the Lagrangian (3.16), we get the reduced Lagrangian

$$L(Q,\dot{Q}) = \tfrac{1}{2}(T_C \dot{Q}^C)^{-1} G_{AB} \dot{Q}^A \dot{Q}^B + \phi_A \dot{Q}^A, \tag{3.19}$$

TABLE II. Alternative forms of the action.

	Action	Lagrangian, Hamiltonian, super-Hamiltonian	
Physical canonical action	$s[q,p]=\int dt\,(p_a d_t q^a - h(t,q,p))$	$h=\frac{1}{2}g^{ab}(p_a-\phi_a)(p_b-\phi_b)+\phi$	
Physical Lagrangian action	$s[q]=\int dt\, l(t,q,d_t q)$	$l=\frac{1}{2}g_{ab}\,d_t q^a\, d_t q^b+\phi_a\, d_t q^a-\phi$	
Extended canonical action, conditional	$S[Q,P]=\int d\tau\, P_A\dot{Q}^A,\quad H=0$	General coordinates	$H=U^A(P_A-\phi_A)+\frac{1}{2}G^{AB}(P_A-\phi_A)(P_B-\phi_B)$
Extended canonical action, with lapse multiplier	$S[Q,P,N]=\int d\tau\,(P_A\dot{Q}^A-NH)$	Special coordinates, rescaled	$H^{(0)}=p_0+\frac{1}{2}g^{ab}(p_a-\phi_a)(p_b-\phi_b)+\phi$
Extended Lagrangian action, homogeneous	$S[Q]=\int d\tau\, L(Q,\dot{Q})$	General coordinates	$L(Q,\dot{Q})=\frac{1}{2}(T_C\dot{Q}^C)^{-1}G_{AB}\dot{Q}^A\dot{Q}^B+\phi_A\dot{Q}^A$
		Special coordinates	$L(t,q,\dot{t},\dot{q})=\frac{1}{2}\dot{t}^{-1}g_{ab}\dot{q}^a\dot{q}^b+\phi_a\dot{q}^a-\phi\dot{t}$
Extended Lagrangian action, with lapse multiplier	$S[Q,N]=\int d\tau\, L(Q,\dot{Q},N)$	General coordinates	$L(Q,\dot{Q},N)=-\frac{1}{2}(N^{-1}(T_C\dot{Q}^C)^{1/2}-(T_C\dot{Q}^C)^{-1/2})^2\times G_{AB}\dot{Q}^A\dot{Q}^B+L(Q,\dot{Q})$
		Special coordinates	$L(t,q,\dot{t},\dot{q},N)=-\frac{1}{2}(N^{-1}\dot{t}^{1/2}-\dot{t}^{-1/2})^2\times g_{ab}\dot{q}^a\dot{q}^b+L(t,q,\dot{t},\dot{q})$

which is a homogeneous function of the first degree in the extended velocities $\dot{Q}^A$. In special coordinates $Q^A=\{t,q^a\}$, the homogeneous Lagrangian assumes the form

$$L(t,q,\dot{t},\dot{q})=\tfrac{1}{2}\dot{t}^{-1}g_{ab}\dot{q}^a\dot{q}^b+\phi_a\dot{q}^a-\phi\dot{t}. \tag{3.20}$$

We display a summary of our results for the alternative forms of the action in Table II.

4. PATH INTEGRALS IN PHYSICAL PHASE SPACE

The canonical action (2.1)–(2.2) on physical phase space is a logical starting point for path integration because the privileged Liouville measure $d^n q\, d^n p$ in this space induces a natural measure in space of skeletonized paths. We represent the quantum propagator by a path integral on the physical phase space following the procedure of Ref. 2. In subsequent sections, we transform this path integral into equivalent path integrals corresponding to alternative forms of the action. In this process, nontrivial and often quite complicated measures are induced in alternative spaces of paths.

The Hilbert space of our dynamical system is the space of scalar state functions $\psi(q,t)$ with the scalar product

$$\langle \underset{1}{\psi}|\underset{2}{\psi}\rangle=\int d^n q\, g^{1/2}(t,q)\underset{1}{\psi}{}^*(t,q)\underset{2}{\psi}(t,q). \tag{4.1}$$

Positions q^a and momenta p_a are represented by Hermitian operators

$$\mathbf{q}^a=q^a,\quad \mathbf{p}_a=-ig^{-1/4}\partial_a g^{1/4}. \tag{4.2}$$

The classical Hamiltonian (2.2) is turned into a covariant operator

$$\begin{aligned}\mathbf{h}&=\tfrac{1}{2}g^{-1/4}(\mathbf{q})(\mathbf{p}_a-\phi_a(\mathbf{q}))g^{1/2}g^{ab}(\mathbf{q})\\&\quad\times(\mathbf{p}_b-\phi_b(\mathbf{q}))g^{-1/4}(\mathbf{q})+\phi(\mathbf{q})\\&=-\tfrac{1}{2}\Delta+i(\phi^a\partial_a+\tfrac{1}{2}\phi^a{}_{|a})+\phi+\tfrac{1}{2}\phi^a\phi_a,\end{aligned} \tag{4.3}$$

which is again Hermitian under the norm (4.1). The state function $\psi(t,q)$ is evolved in time by the Schrödinger equation

$$ig^{-1/4}\partial_t(g^{1/4}\psi)=\mathbf{h}\psi. \tag{4.4}$$

The general solution of Eq. (4.4) is provided by the quantum propagator $\langle t'',q''|t',q'\rangle$,

$$\psi(t'',q'')=\int d^n q'\,\langle t'',q''|t',q'\rangle\psi(t',q'). \tag{4.5}$$

This propagator is a scalar in q'' and a scalar density in q'. It satisfies the Schrödinger equation

$$ig''^{-1/4}\partial_{t''}(g''^{1/4}\langle t'',q''|t',q'\rangle)=\mathbf{h}''\langle t'',q''|t',q'\rangle \tag{4.6}$$

with the boundary condition

$$\langle t'',q''|t'',q'\rangle=\delta(q''|q'). \tag{4.7}$$

We represent the quantum propagator by an integral over all phase space paths $q(t)$, $p(t)$ which start in the configuration q' at t' and end in the configuration q'' at t'',

$$\langle t'',q''|t'',q'\rangle\, d^n q'=\int Dq\, Dp\, e^{is[q(t),p(t)]}. \tag{4.8}$$

Here, $s[q(t),p(t)]$ is the canonical action integral (2.1) and $Dq\,Dp$ is a measure in the space of phase space paths.

We interpret the formal expression (4.8) by a skeletonization procedure in which the time between t' and t'' is sliced into small intervals and the measure becomes the product of the Liouville phase space measures on each slice. In the integrand, we need to skeletonize the action for each path in phase space. We replace the action functional by a sum of principal functions for getting from one phase space point on the skeletonized path to the next. These principal functions cannot be the Hamilton principal functions, because Hamilton's principal functions are determined by the initial and

final configurations and do not depend on momentum. A correct construction was discussed in Ref. 2. Evaluate the canonical action (2.1) along the actual path $q^a(t)$ in configuration space and the momentum path $p_a(t)$ found by transporting an arbitrary initial momentum along the configuration space path by a specified rule. There results a principal function $s(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)})$ which depends on the initial and final configurations and on the initial momentum. By summing such principal functions for all segments of the phase space path, one arrives at an action function which is manifestly covariant under point transformations

$$q^{a^*} = q^{a^*}(t,q), \quad p_{a^*} = \frac{\partial q^b(t,q^*)}{\partial q^{a^*}} p_b. \tag{4.9}$$

There are, in fact, a variety of such skeletonization procedures, depending on which rule is used to transport the momentum along the actual classical path. Each gives a different quantum mechanical propagator. We shall use the rule of geodesic deviation transport. There are compelling reasons for such a choice: (1) *A fortiori*, the momentum vector is Lie propagated by a flow of actual configuration paths; (2) *a posteriori*, the Schrödinger equation (4.4) does not contain any curvature term.

Let us now describe this procedure in detail. The skeletonized phase space path $t_{(K)}, q_{(K)}, p_{(K)}$, $K = 0,1,...,N$, starts at the configuration q' at t' and ends in the configuration q'' at t'',

$$t_{(0)} = t', \quad q_{(0)} = q', \quad t_{(N)} = t'', \quad q_{(N)} = q''. \tag{4.10}$$

The canonical action integral $s[q(t), p(t)]$ is replaced by a chain

$$\sum_{K=0}^{N-1} s(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)}) \tag{4.11}$$

of phase space principal functions $s(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)})$. The skeletonized measure $Dq\,Dp$ is taken as the product

$$\prod_{K=0}^{N-1} (2\pi)^{-n}\, d^n q_{(K)}\, d^n p_{(K)} \tag{4.12}$$

of invariant Liouville measures on phase space. There is one such measure at each time $t_{(K)}$, $K = 0,1,...,N-1$, with the exception of the final time $t_{(N)}$. The integration is performed over all of the momenta $p_{(K)}$, $K = 0,1,...,N-1$, but only over the interpolated coordinates $q_{(I)}$, $I = 1,...,N-1$. The differential $d^n q'$ thus remains unused in the integral (4.8) and appears on both sides of the equation. The asymmetric way in which q integrations and p integrations are performed reflects the fact that the paths have fixed boundary configurations but free boundary momenta. The path integral (4.8) is defined as a limit of the described $[Nn\cdot(N-1)n]$-fold integral (q' integration omitted) as $N \to \infty$ while the skeletonization is infinitely refined. That is, if

$$\Delta t_{\mathrm{MAX}} \equiv \max_{K=0,...,N-1} |t_{(K+1)} - t_{(K)}|, \tag{4.13a}$$

then

$$\int Dq\,Dp\, e^{is[q(t),p(t)]} \equiv \lim_{\Delta t_{\mathrm{MAX}} \to 0} \int \prod_{K=0}^{N-1} d^n q_{(K)}\, d^n p_{(K)} \times C(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)}). \tag{4.13b}$$

The biscalar

$$C(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)}) \equiv (2\pi)^{-n} e^{is(t_{(K+1)}, q_{(K+1)}|t_{(K)}, q_{(K)}, p_{(K)})} \tag{4.14}$$

we call the classical propagator.

The phase space principal function $s(t'', q''|t', q', p')$ is defined as the canonical action integral (2.1) evaluated along the actual configuration path $q(t)$ between t', q' and t'', q'' given by the equations

$$\nabla_t(g_{ab}\, d_t q^b) = F_a \equiv B_{ab}\, d_t q^b + E_a,$$
$$B_{ab} = \partial_a \phi_b - \partial_b \phi_a, \quad E_a = -\partial_a \phi - \partial_t \phi_a, \tag{4.15}$$

with the momentum p_a propagated from its initial value $p_{a'}$ by the equation of geodesic deviation with a force term,

$$\nabla_t^2(p_a - \phi_a) + R_{abcd}\, d_t q^b (p^c - \phi^c)\, d_t q^d = \nabla_t F_a,$$
$$[\nabla_t(p_a - \phi_a - g_{ab}\, d_t q^b)]_{t=t'} = 0. \tag{4.16}$$

The phase space principal function $s(t'', q''|t', q', p')$ is a biscalar under point transformations (4.9). It is a quadratic function of the initial momenta.

At each step of the skeletonization procedure, the corresponding phase space principal function enters into the classical propagator (4.14). In the limit (4.13), we need to know each function only up to terms linear in the time interval $\Delta t_{(K)} = t_{(K+1)} - t_{(K)}$ and quadratic in the instantaneous geodesic separation $\sigma_{t_{(K)}}(q_{(K+1)}|q_{(K)})$.

To write such an approximate form of the phase space principal function $s(t'', q''|t', q', p')$, we introduce the configuration space Hamilton principal function $s(t'', q''|t', q')$. This function is the extremum of $s(t'', q''|t', q', p')$ with respect to p' and it satisfies the Hamilton–Jacobi equations

$$\partial_{t''} s + h(t'', q^{a''}, p_{a''} = \partial_{a''} s) = 0,$$
$$-\partial_{t'} s + h(t', q^{a'}, p_{a'} = -\partial_{a'} s) = 0. \tag{4.17}$$

From the Hamilton principal function, we can find the initial velocity $d_t q^{a'}$ on the actual path from t', q' to t'', q'':

$$d_t q^{a'} = -g^{a'b'}(\partial_{b'} s + \phi_{b'}). \tag{4.18}$$

This velocity is of the order $\sigma_{t'}/\Delta t$. The approximate form of the phase space principal function can be written in the suggestive form

$$s(t'', q''|t', q', p') \approx (p_{a'}\, d_t q^{a'} - \tfrac{1}{2}\bar{g}^{a'b'}(p_{a'} - \phi_{a'})(p_{b'} - \phi_{b'}) - \phi')\,\Delta t. \tag{4.19}$$

The coefficient

$$\bar{g}^{a'b'}(t'', q''|t', q') = g^{a'b'} - \tfrac{1}{3} R^{a'}{}_{c'}{}^{b'}{}_{d'}(\Delta t\, d_t q^{c'})(\Delta t\, d_t q^{d'}) \tag{4.20}$$

differs from the metric $g^{a'b'}(t', q')$ by a Riemann curvature term which is brought in by the geodesic deviation transport. This term is of the order $\sigma_{t'}^2$. The function (4.19) is constructed in the following way: (I) The initial value of the canonical Lagrangian $p_{a'}\, d_t q^{a'} - h(t', q', p')$ is multiplied by the time interval $\Delta t = t'' - t'$; (II) the initial velocity $d_t q^{a'}$ is expressed

as a function of the boundary data t',q' and t'',q'', Eq. (4.18); (III) the metric in the initial Hamiltonian is replaced by the tensor–scalar coefficient (4.20). We call the modified Hamiltonian $\bar{h}(t'',q''|t',q',p')$.

The description of the phase space integral is now complete. The approximate form (4.19)–(4.20) of the phase space principal function can be used in each classical propagator (4.14) and the path integral defined as the limit (4.13). One can prove[1] that the quantum propagator (4.8) represented by this path integral satisfies the Schrödinger equation (4.6) with the boundary condition (4.7). The geodesic deviation transport which induces the modification (4.20) of the metric ensures that no scalar curvature potential appears in the Schrödinger equation.

5. PATH INTEGRALS IN PHYSICAL CONFIGURATION SPACE

We pass from the phase space path integral (4.13)–(4.14) to a path integral on the physical configuration space by performing momenta integrations. The $K+1$ step in the skeletonization process starts at $t_{(K)},q_{(K)},p_{(K)}$ and ends at $t_{(K+1)},q_{(K+1)}$. Generically, we call

$$t_{(K)}=t,\quad q_{(K)}=q,\quad p_{(K)}=p$$

and (5.1)

$$t_{(K+1)}=\bar{t},\quad q_{(K+1)}=\bar{q}.$$

The phase space principal function (4.19) at each step can be completed into a square,

$$s(\bar{t},\bar{q}|t,q,p)=-\tfrac{1}{2}\bar{g}^{ab}\pi_a\pi_b\,\Delta t+l(t,q,d_tq)\Delta t. \tag{5.2}$$

Here,

$$\Delta t=\bar{t}-t, \tag{5.3}$$

$$\pi_a=p_a-g_{ab}d_tq^b-\phi_a \tag{5.4}$$

and $l(t,q,d_tq)$ is the physical Lagrangian (3.9). The initial velocity d_tq^a is still expressed through the configuration space boundary data $\bar{q},\bar{t},q,t$:

$$d_tq^a=-g^{ab}(t,q)[\partial_b s(\bar{t},\bar{q}|t,q)+\phi_b(t,q)]. \tag{5.5}$$

Let (4.14) be the phase space classical propagator from t,q,p to $\bar{t},\bar{q}$,

$$C(\bar{t},\bar{q}|t,q,p)=(2\pi)^{-n}e^{is(\bar{t},\bar{q}|t,q,p)}. \tag{5.6}$$

We define the configuration space classical propagator as an integral of Eq. (5.6) over the momenta,

$$C(\bar{t},\bar{q}|t,q)\equiv\int d^np\,C(\bar{t},\bar{q}|t,q,p). \tag{5.7}$$

The integration over p can be replaced by integration over π. This leads to the Gaussian integral

$$\int d^n\pi\,e^{-(1/2)i\Delta t\bar{g}^{ab}\pi_a\pi_b}=((2\pi)^{-1}i\Delta t)^{-n/2}\bar{g}^{1/2}, \tag{5.8}$$

where

$$\bar{g}(\bar{t},\bar{q}|t,q)\equiv\det\bar{g}_{ab}. \tag{5.9}$$

Up to the first order terms in Δt,

$$s(\bar{t},\bar{q}|t,q)=l(t,q,d_tq)\,\Delta t. \tag{5.10}$$

This leads to the configuration space classical propagator

$$C(\bar{t},\bar{q}|t,q)=(2\pi i\Delta t)^{-n/2}\bar{g}^{1/2}e^{is(\bar{t},\bar{q}|t,q)}. \tag{5.11}$$

By integrating over all the momenta $p_{(K)}$, $K=0,1,\ldots,N-1$, we transform the quantum propagator (4.8),(4.13) to a configuration space form

$$\langle t'',q''|t',q'\rangle\,d^nq'=\int\bar{D}q\,e^{is[q(t)]}$$
$$\equiv\lim_{\Delta t_{MAX}\to 0}\int\prod_{K=0}^{N-1}d^nq_{(K)}\,C(t_{(K+1)},q_{(K+1)}|t_{(K)},q_{(K)}). \tag{5.12}$$

The integration takes place over the interpolated positions $q_{(I)}$, $I=1,\ldots,N-1$.

The Lagrangian action integral $s[q(t)]$ in Eq. (5.12) gets skeletonized by a chain of Hamilton's principal functions

$$s[q(t)]\approx\sum_{K=0}^{N-1}s(t_{(K+1)},q_{(K+1)}|t_{(K)},q_{(K)})$$
$$\approx\sum_{K=0}^{N-1}l(t_{(K)},q_{(K)},d_tq_{(K)})\Delta t_{(K)}, \tag{5.13}$$

and the measure $\bar{D}q$ is skeletonized by the product

$$\bar{D}q\approx\prod_{K=0}^{N-1}d^nq_{(K)}\,(2\pi i\,\Delta t_{(K)})^{-1/2}$$
$$\times\bar{g}^{1/2}(t_{(K+1)},q_{(K+1)}|t_{(K)},q_{(K)}). \tag{5.14}$$

Each determinant $\bar{g}^{1/2}$ can be expressed as

$$\bar{g}^{1/2}=g^{1/2}(t_{(K)},q_{(K)})(1+R_{ab}(q_{(K)})$$
$$\times\Delta t_{(K)}d_tq^a{}_{(K)}\,\Delta t_{(K)}\,d_tq^b{}_{(K)}). \tag{5.15}$$

Under this measure, the quantum propagator (5.12) satisfies the Schrödinger equation without any curvature term.

6. PATH INTEGRALS IN EXTENDED PHASE SPACE

We shall now express the quantum propagator by path integrals in extended phase space. There are two ways of doing this corresponding classically to whether the constraints are enforced explicitly or implicitly through a lapse multiplier. We begin by replacing each classical propagator $C(\bar{t},\bar{q}|t,q,p)$ by an extended propagator $C(\bar{Q}|Q,P)$ such that, in special coordinates (2.3),(2.6),

$$C(\bar{t},\bar{q}|t,q,p)=\int dQ^0\,dP_0\,C(\bar{Q}|Q,P). \tag{6.1}$$

The procedure then closely follows the parametrization process of classical action. First, we take the absolute time as a prescribed function $t(\tau)$ of a label time $\tau\in[\tau',\tau'']$ respecting the boundary conditions

$$t(\tau')=t',\quad t(\tau'')=t''. \tag{6.2}$$

To first order in $\Delta\tau$,

$$\Delta t=\dot{t}\,\Delta\tau,\quad \Delta t\equiv\bar{t}-t,\quad \Delta\tau\equiv\bar{\tau}-\tau, \tag{6.3}$$

and, as a consequence of Eqs. (4.19) and (4.14),

$$C(\bar{t},\bar{q}|t,q,p)=(2\pi)^{-n}e^{i(p_a\dot{q}^a-\bar{h}(\bar{t},\bar{q}|t,q,p)\dot{t})\Delta\tau}. \tag{6.4}$$

The initial velocity $\dot{q}^a$ in Eq. (6.4) is again expressed as a function of the boundary configuration data [cf. Eq. (5.5)]:

$$\dot{q}^a=\dot{t}\,d_tq^a=-\dot{t}g^{ab}(\partial_bs+\phi_b). \tag{6.5}$$

We adjoin to it the quantity $\dot{t}$ and write

$$\dot{Q}^A = \dot{t}(\tau)\{1, -g^{ab}(\partial_b s + \phi_b)\}. \tag{6.6}$$

In the expression (6.4), the variables q and $\bar{q}$ are arbitrary, but t is considered as a given function of τ, $t(\tau)$. To remove this asymmetry, we consider both t and q as independent variables $Q^A = \{t,q\}$, but multiply the classical propagator (6.4) by a delta function $\delta(Q^0 - t(\tau))$. From now on, s and ϕ_b in Eq. (6.6) are also considered as functions of Q^A, though $\dot{t}(\tau)$ is still a prescribed function of τ.

We also extend the momenta variables by adding a variable p_0, $P_A = \{p_0, p_a\}$, and write the phase factor in Eq. (6.4) as the linear combination $P_A \dot{Q}^A \, \Delta\tau$. To ensure that p_0 is $-\bar{h}$, we multiply the classical propagator by the delta function $\delta(\overline{H}^{(o)})$ of the modified Hamiltonian constraint (2.8),

$$\overline{H}^{(o)} = p_0 + \bar{h}(\bar{t},\bar{q}|t,q,p). \tag{6.7}$$

These changes lead to the following classical propagator on extended phase space:

$$C(\overline{Q}|Q,P) = (2\pi)^{-n}\delta(Q^0 - t(\tau))\delta(\overline{H}^{(o)})e^{iP_A\dot{Q}^A\Delta\tau}. \tag{6.8}$$

Integration of this propagator with respect to the newly introduced variables Q^0 and P_0 reduces it to the old propagator, Eq. (6.1).

The new propagator (6.8) can be written in a manifestly covariant form. We introduce fields $t(Q)$ and $u^A(Q)$ by Eqs. (2.11) and (2.12) and a degenerate tensor–scalar $\bar{g}^{AB}(\overline{Q}|Q)$ related to the coefficient (4.20) by a counterpart of Eq. (2.14). The super-Hamiltonian $\overline{H}^{(o)}$ is thereby cast to the form (2.17) with $\bar{g}^{AB}$ in place of g^{AB}.

In the same vein, Eq. (6.6) assumes the form

$$\dot{Q}^A = \dot{t}(u^A - g^{AB}(\partial_B S + \phi_B)). \tag{6.9}$$

The Hamilton–Jacobi equations which determine the Hamilton principal function

$$S(\overline{Q}|Q) = s(\bar{t},\bar{q}|t,q) \tag{6.10}$$

are obtained by substituting $P_A = -\partial_A S$ and $P_{\bar{A}} = \partial_{\bar{A}} S$ into the Hamiltonian constraint at the initial and the final boundaries,

$$\begin{aligned} &-u^A(\partial_A S + \phi_A) + \tfrac{1}{2}g^{AB}(\partial_A S + \phi_A)(\partial_B S + \phi_B) = 0, \\ &u^{\bar{A}}(\partial_{\bar{A}} S - \phi_{\bar{A}}) + \tfrac{1}{2}g^{\bar{A}\bar{B}}(\partial_{\bar{A}} S - \phi_{\bar{A}})(\partial_{\bar{B}} S - \phi_{\bar{B}}) = 0. \end{aligned} \tag{6.11}$$

The classical propagator (6.8) then takes on a manifestly covariant appearance

$$C(\overline{Q}|Q,P) = (2\pi)^{-n}\delta(t(Q) - t(\tau))\delta(\overline{H}^{(o)})e^{iP_A\dot{Q}^A\Delta\tau}. \tag{6.12}$$

We can now mix the extended phase space variables Q^A, P_A by an arbitrary point transformation (2.20) and transform the classical propagator as a biscalar without changing its general form (6.12).

In a final step, we scale $\overline{H}^{(o)}$ into $\overline{H}$ by a positive scalar factor $\Lambda(Q)$ as in Eqs. (2.18)–(2.19). In terms of the scaled quantities (2.18), S again satisfies the Hamilton–Jacobi equations (6.11), but the scaling factor enters into Eq. (6.9) by which $\dot{Q}^A$ is interpreted in terms of the boundary configurations,

$$\dot{Q}^A = \Lambda^{-1}\dot{t}\,[U^A - G^{AB}(\partial_B S + \phi_B)]. \tag{6.13}$$

Because $\delta(\overline{H}^{(o)}) = \delta(\Lambda^{-1}\overline{H}) = \Lambda\,\delta(\overline{H})$, the scaling factor also explicitly appears in the modulus of the classical propagator (6.12), which becomes

$$C(\overline{Q}|Q,P) = (2\pi)^{-n}\Lambda(Q)\delta(t(Q) - t(\tau))\delta(\overline{H})e^{iP_A\dot{Q}^A\Delta\tau}. \tag{6.14}$$

The absolute time function $t(Q)$ is covariantly characterized by Eqs. (2.21) and (2.27). The scaling factor Λ in expressions (6.13) and (6.14) can then be interpreted by Eq. (2.28) or, alternatively, as the Poisson bracket

$$\Lambda(Q) = [t(Q),H] = [t(Q),\overline{H}]. \tag{6.15}$$

This completes a covariant characterization of the classical propagator (6.14).

The quantum propagator can be represented by a path integral in the extended phase space,

$$\begin{aligned} \langle Q''|Q'\rangle\delta(t(Q') - t')\,d^{n+1}Q' &= \int \overline{D}Q\,\overline{D}P\,e^{iS[Q,P]} \\ &= \lim_{\Delta\tau_{MAX}\to 0}\int\prod_{K=0}^{N-1} d^{n+1}Q_{(K)}\,d^{n+1}P_{(K)} \\ &\quad\times C(Q_{(K+1)}|Q_{(K)},P_{(K)}). \end{aligned} \tag{6.16}$$

The integrations are performed over all the extended momenta $P_{(K)}$, $K = 0,1,\ldots,N-1$, but only over the interpolated extended coordinates $Q_{(I)}$, $I = 1,\ldots,N-1$. Due to Eq. (6.1), we obtain in this way our old quantum propagator (4.8),(4.13).

The new form (6.16) of the path integral corresponds to the conditional form of the action, Table II, line 3. The skeletonized measure

$$\begin{aligned} \overline{D}Q\,\overline{D}P \approx \prod_{K=0}^{N-1} &d^{n+1}Q_{(K)}\,d^{n+1}P_{(K)} \\ &\times(2\pi)^{-n}\Lambda(Q_{(K)})\delta(t(Q_{(K)}) - t(\tau_{(K)})) \\ &\times\delta(\overline{H}(Q_{(K+1)}|Q_{(K)},P_{(K)})) \end{aligned} \tag{6.17}$$

contains a product of delta functions $\delta(\overline{H}(Q_{(K+1)}|Q_{(K)}, P_{(K)}))$ which enforce the Hamiltonian constraint at each instant $\tau_{(K)}$ of the skeletonized time. However, these constraints are not simply classical Hamiltonian constraints at $\tau_{(K)}$, but modified constraints in which the metric $G^{AB}(Q_{(K)})$ is replaced by the tensor–scalar coefficient $\overline{G}^{AB}(Q_{(K+1)}|Q_{(K)})$. This modification is necessary for the quantum propagator (6.16) to satisfy the Schrödinger equation without an additional scalar curvature potential. If the measure contained the unmodified super-Hamiltonian $H(Q_{(K)},P_{(K)})$, the Schrödinger equation would acquire the potential $\frac{1}{12}R$.

Besides the delta functions of super-Hamiltonians, the measure also contains the delta functions $\delta(t(Q_{(K)}) - t(\tau_{(K)}))$. These delta functions ensure that the instants of the label time τ correspond to the leaves of absolute time t. The configurations which the system has to select at an instant τ are thus all simultaneous in the absolute sense. The labeling of the leaves of absolute time, however, is provided by an arbitrary parameter τ. Finally, the factor $\Lambda(Q)$ in the measure takes care of an arbitrary scaling of the Hamiltonian constraint.

In addition to the choice of measure, one must also specify how to skeletonize the action functional

$S[Q,P] = \int_\tau^{\tau''} d\tau\, P_A \dot{Q}^A$. Our skeletonization says that $S[Q,P]$ is to be replaced by the sum

$$S[Q,P] \approx \sum_{K=0}^{N-1} P_{(K)A} \dot{Q}_{(K)}{}^A \, \Delta\tau_{(K)}, \tag{6.18}$$

in which $\dot{Q}_{(K)}{}^A$ is the actual extended velocity at $\tau_{(K)}$ on the actual path from $Q_{(K)}$ to $Q_{(K+1)}$. This actual velocity can be derived from the Hamilton principal function $S(Q_{(K+1)}|Q_{(K)})$ by Eq. (6.13).

It is easy to introduce the lapse multiplier and pass from the conditional form of the path integral to an unconditional one. We just interpret each $\delta(\overline{H})$ as the Fourier integral

$$\delta(\overline{H}) = \int dN\, \Delta\tau\, (2\pi)^{-1} e^{-iN\overline{H}\,\Delta\tau}. \tag{6.19}$$

In other words, we extend the classical propagator $C(\overline{Q}\,|Q,P)$ into the Q, P, N space by the prescription

$$C(\overline{Q}\,|Q,P,N) = (2\pi)^{-(n+1)}\, \Delta\tau\, \Lambda(Q)\delta(t(Q) - t(\tau)) \times e^{i(P_A \dot{Q}^A - NH)\,\Delta\tau} \tag{6.20}$$

and connect it with the old propagator by the equation

$$C(\overline{Q}\,|Q,P) = \int DN\, C(\overline{Q}\,|Q,P,N). \tag{6.21}$$

The quantum propagator (6.16) can then be represented by a path integral in the Q, P, N space,

$$\begin{aligned}\langle Q''|Q'\rangle \delta(t(Q') - t')\, d^{n+1}Q' &= \int \overline{D}Q\, \overline{D}P\, \overline{D}N\, e^{iS[Q,P,N]} \\ &\equiv \lim_{\Delta\tau_{MAX}\to 0} \int \prod_{K=0}^{N-1} d^{n+1}Q_{(K)}\, d^{n+1}P_{(K)}\, dN_{(K)} \\ &\quad \times C(Q_{(K+1)}|Q_{(K)},P_{(K)},N_{(K)}).\end{aligned} \tag{6.22}$$

The integration takes place over all $N_{(K)}$, $K = 0,1,\ldots,N-1$. This corresponds to the fact that the lapse function is a Lagrange multiplier which, like the momenta P_A, can be freely specified at the ends.

The skeletonized measure has the form

$$\begin{aligned}\overline{D}Q\, \overline{D}P\, \overline{D}N \approx \prod_{K=0}^{N-1} & d^{n+1}Q_{(K)}\, d^{n+1}P_{(K)}\, dN_{(K)} \\ & \times \Delta\tau_{(K)}\, \Lambda(Q_{(K)})(2\pi)^{-(n+1)} \\ & \times \delta(t(Q_{(K)}) - t(\tau_{(K)})).\end{aligned} \tag{6.23}$$

The product $N_{(K)}\, \Delta\tau_{(K)}\, \Lambda(Q_{(K)})$ which enters into the measure is unchanged when we use a different label time; in fact, $N\,\Delta\tau\,\Lambda$ is to be interpreted as the interval Δt of the absolute time, Eq. (3.5).

Finally, the action functional (3.2) is replaced by the sum

$$\begin{aligned}S[Q,P,N] \approx \sum_{K=0}^{N-1} & (P_{(K)A} \dot{Q}_{(K)}{}^A \\ & - N_{(K)} \overline{H}(Q_{(K+1)}|Q_{(K)},P_{(K)}))\, \Delta\tau_{(K)}.\end{aligned} \tag{6.24}$$

Here, $\dot{Q}_{(K)}{}^A$ is again given by Eq. (6.13) and $\overline{H}$ is the modified super-Hamiltonian.

We have thereby transformed the path integral in physical phase space into two equivalent forms in the extended phase space, one with and one without the lapse multiplier.

7. PATH INTEGRALS IN EXTENDED CONFIGURATION SPACE

The path integral in physical configuration space was obtained from the path integral in physical phase space by evaluating all integrals over the momenta. Similarly, by integrating the extended classical propagator (6.14) over the extended momentum variables we cast the path integral into a form corresponding to the homogeneous Lagrangian on the extended configuration space. To do this, we introduce for convenience mechanical energy and momenta

$$\Pi_A = P_A - \phi_A \tag{7.1}$$

as new variables. The extended classical propagator (6.14) assumes the form

$$\begin{aligned}C(\overline{Q}\,|Q,\Pi) = (2\pi)^{-n} \Lambda(Q)\delta(t(Q) - t(\tau))\delta(\overline{H}) e^{i\Pi_A \dot{Q}^A\,\Delta\tau} \\ \times e^{i\phi_A \dot{Q}^A\,\Delta\tau},\end{aligned} \tag{7.2}$$

with

$$\overline{H} = U^A \Pi_A + \tfrac{1}{2}\overline{G}^{AB}\Pi_A \Pi_B. \tag{7.3}$$

Let Q^a_A be n linearly independent covectors perpendicular to U^A,

$$U^A Q^a_A = 0, \quad a = 1,\ldots,n. \tag{7.4}$$

The projected coefficient

$$\overline{G}^{ab} = \overline{G}^{AB} Q^a_A Q^b_B \tag{7.5}$$

is nondegenerate. The covectors $\{T_A, Q^a_A\}$ form a basis in the cotangent space. We split Π_A into a longitudinal and transversal parts according to

$$\Pi_A = \Pi_\parallel T_A + \Pi_a Q^a_A. \tag{7.6}$$

The Jacobian $J = \det \partial\{\Pi_A\}/\partial\{\Pi_\parallel, \Pi_a\}$ of the transformation (7.6) from the variables $\{\Pi_\parallel, \Pi_a\}$ to the variables Π_A is (see Appendix B)

$$J = (1/n!)\delta^{AA_1\cdots A_n} T_A Q^{a_1}_{A_1} \cdots Q^{a_n}_{A_n} \delta_{a_1\cdots a_n}, \tag{7.7}$$

where $\delta_{a_1\cdots a_n}$ is the alternating symbol. As a consequence,

$$C(\overline{Q}\,|Q,P)\, d^{n+1}P = J(Q) C(\overline{Q}\,|Q,\Pi_\parallel,\Pi_a) d\Pi_\parallel\, d^n\Pi. \tag{7.8}$$

In the new variables,

$$\overline{H} = \Pi_\parallel + \tfrac{1}{2}\overline{G}^{ab}\Pi_a \Pi_b \tag{7.9}$$

and the integration with respect to $\Pi_\parallel$ is easily performed. We get

$$\begin{aligned}C(\overline{Q}\,|Q,\Pi_a) &= \int d\Pi_\parallel\, J(Q) C(\overline{Q}\,|Q,\Pi_\parallel,\Pi_a) \\ &= (2\pi)^{-n} J(Q) \Lambda(Q) \delta(t(Q) - t(\tau)) \\ &\quad \times e^{i\phi_A \dot{Q}^A\,\Delta\tau} e^{i[\Pi_a \dot{Q}^a - (1/2)(T_C \dot{Q}^C)\overline{G}^{ab}\Pi_a \Pi_b]\,\Delta\tau},\end{aligned} \tag{7.10}$$

where we have introduced the abbreviation

$$\dot{Q}^a \equiv Q^a_A \dot{Q}^A. \tag{7.11}$$

The terms in Π_a can be completed into a square,

$$\begin{aligned}(\Pi_a \dot{Q}^a - (T_C \dot{Q}^C)\tfrac{1}{2}\overline{G}^{ab}\Pi_a \Pi_b)\,\Delta\tau \\ = (-\tfrac{1}{2}(T_C \dot{Q}^C)\overline{G}^{ab}\widetilde{\Pi}_a \widetilde{\Pi}_b + \tfrac{1}{2}(T_C \dot{Q}^C)^{-1}\overline{G}_{ab}\dot{Q}^a \dot{Q}^b)\,\Delta\tau,\end{aligned} \tag{7.12}$$

where

$$\tilde{\Pi}_a = \Pi_a - (T_C \dot{Q}^C)^{-1} \overline{G}_{ab} \dot{Q}^b. \tag{7.13}$$

Moreover,

$$\overline{G}_{ab} \dot{Q}^a \dot{Q}^b = \overline{G}_{AB} \dot{Q}^A \dot{Q}^B = G_{AB} \dot{Q}^A \dot{Q}^B. \tag{7.14}$$

One can replace the modified coefficient $\overline{G}_{AB}$ by the metric G_{AB} because, in special coordinates $Q^A = \{t, q_a\}$, $(R_{acbd}\, \Delta\tau \dot{q}^c \Delta\tau \dot{q}^d) \dot{q}^a \dot{q}^b = 0$. As a result,

$$C(\overline{Q}\,|Q,\tilde{\Pi}_a) = (2\pi)^{-n} J(Q) \Lambda(Q) \delta(t(Q) - t(\tau)) \times e^{-(1/2)i(T_C \dot{Q}^C)\, \Delta\tau\, \overline{G}^{ab} \tilde{\Pi}_a \tilde{\Pi}_b} e^{iL(Q,\dot{Q})\, \Delta\tau}, \tag{7.15}$$

where $L(Q,\dot{Q})$ is the homogeneous Lagrangian (3.19). The Gaussian integral over $\tilde{\Pi}_a$ gives

$$(2\pi)^{-n} \int d^n \tilde{\Pi}\, e^{-(1/2)i(T_C \dot{Q}^C)^{-1}\, \Delta\tau\, \overline{G}^{ab} \tilde{\Pi}_a \tilde{\Pi}_b} = (2\pi i T_C \dot{Q}^C \Delta\tau)^{-n/2} \overline{G}^{1/2} \tag{7.16}$$

with

$$\overline{G} \equiv \det \overline{G}_{ab}. \tag{7.17}$$

The product

$$J\overline{G}^{1/2} \equiv \overline{D}^{-1/2} \tag{7.18}$$

can be written directly in terms of the degenerate coefficient $\overline{G}^{AB}$ (Appendix B):

$$\overline{D} = (1/n!) \delta_{AA_1 \cdots A_n} U^A U^B \overline{G}^{A_1 B_1} \cdots \overline{G}^{A_n B_n} \delta_{BB_1 \cdots B_n}. \tag{7.19}$$

This sequence of steps yields the classical propagator in extended configuration space,

$$\begin{aligned} C(\overline{Q}\,|Q) &= \int d^{n+1} P\, C(\overline{Q}\,|Q,P) \\ &= \int d^n \tilde{\Pi}\, C(\overline{Q}\,|Q,\tilde{\Pi}_a) \\ &= (2\pi i T_C \dot{Q}^C \Delta\tau)^{-n/2} \\ &\quad \times \overline{D}^{-1/2} (\overline{Q}\,|Q) \Lambda(Q) \delta(t(Q) - t(\tau)) e^{iL(Q,\dot{Q})\, \Delta\tau}. \end{aligned} \tag{7.20}$$

Note that by the interpretation (6.13) of $\dot{Q}^C$ we have

$$T_C \dot{Q}^C = \Lambda^{-1}(Q) \dot{t}(\tau). \tag{7.21}$$

From Eq. (6.16), we obtain a representation of quantum propagator by a path integral in the extended configuration space,

$$\langle Q'' | Q' \rangle \delta(t(Q') - t')\, d^{n+1} Q' = \int \overline{D}Q\, e^{iS[Q]} \equiv \lim_{\Delta\tau_{\mathrm{MAX}} \to 0} \int \prod_{K=0}^{N-1} d^{n+1} Q_{(K)}\, C(Q_{(K+1)}\,|Q_{(K)}). \tag{7.22}$$

The integration takes place only over the interpolated extended coordinates $Q_{(I)}$, $I = 1,\ldots,N-1$. The homogeneous Lagrangian action $S[Q] = \int_{\tau'}^{\tau''} d\tau\, L(Q,\dot{Q})$ is skeletonized by the prescription

$$S[Q] \approx \sum_{K=0}^{N-1} (\tfrac{1}{2}(T_C(Q_{(K)}) \dot{Q}^C_{(K)})^{-1} \times G_{AB}(Q_{(K)}) \dot{Q}^A_{(K)} \dot{Q}^B_{(K)} + \phi_A(Q_{(K)}) \dot{Q}^A_{(K)}\, \Delta\tau_{(K)}). \tag{7.23}$$

The velocities $\dot{Q}_{(K)}$ are interpreted in terms of the configuration data at the ends of each step in Eq. (6.13). Note that the coefficient G_{AB} in Eq. (7.23) is the ordinary degenerate metric unmodified by the curvature term. The modified metric coefficient enters only into the measure, but not into the phase of the path integral (7.22). The measure is skeletonized by the product

$$\overline{D}Q \approx \prod_{K=0}^{N-1} d^{n+1} Q\, (2\pi i T_C(Q_{(K)}) \dot{Q}^C_{(K)}\, \Delta\tau_{(K)})^{-n/2} \times D^{-1/2}(Q_{(K+1)}\,|Q_{(K)}) \Lambda(Q_{(K)}) \times \delta(t(Q_{(K)}) - t(\tau_{(K)})). \tag{7.24}$$

The modified metric coefficient appears in the determinant (7.19).

In the special coordinates (2.3) all previous expressions considerably simplify. The Jacobian (7.7) reduces to

$$J = \Lambda^{-1}, \tag{7.25}$$

the determinant (7.19) goes over to

$$\overline{D} = \Lambda^{n+2} \overline{g}^{-1}, \quad \overline{g} \equiv \det \overline{g}_{ab}, \tag{7.26}$$

and the classical propagator assumes the form

$$C(\bar{t},\bar{q}\,|t,q) = (2\pi i \dot{t}(\tau)\, \Delta\tau)^{-n/2} \times \overline{g}^{1/2}(\bar{t},\bar{q}\,|t,q) \delta(t - t(\tau)) e^{iL(t,q,\dot{t},\dot{q})\, \Delta\tau}. \tag{7.27}$$

Here, $L(t,q,\dot{t},\dot{q})$ is the homogeneous Lagrangian (3.20). In the expression (7.24), t is an independent variable, while $t(\tau)$ and $\dot{t}(\tau)$ are prescribed functions of τ. The velocity $\dot{q}^a$ is interpreted as a function of t,q, and $\bar{t},\bar{q}$ by Eq. (6.5). The measure (7.24) in path integral (7.22) reduces to

$$\overline{D}t\, \overline{D}q \approx \prod_{K=0}^{N-1} dt_{(K)}\, d^n q_{(K)}\, [2\pi i t(\tau_{(K)})\, \Delta\tau_{(K)}]^{-n/2} \times \overline{g}^{1/2}(t_{(K+1)}, q_{(K+1)}\,|t_{(K)}, q_{(K)}) \delta(t_{(K)} - t(\tau_{(K)})), \tag{7.28}$$

while the Lagrangian action $S[t,q]$ gets skeletonized by

$$S[t,q] \approx \sum_{K=0}^{N-1} [\tfrac{1}{2} \dot{t}^{-1}(\tau_{(K)}) g_{ab}(t_{(K)}, q_{(K)}) \dot{q}^a_{(K)} \dot{q}^b_{(K)} + \phi_a(t_{(K)}, q_{(K)}) \dot{q}^a_{(K)} - \phi(t_{(K)}, q_{(K)}) \dot{t}(\tau_{(K)})]\, \Delta\tau_{(K)}. \tag{7.29}$$

When we perform the integrations over $t_{(I)}$, $I = 1,\ldots,N-1$, and parametrize the paths by absolute time, $t(\tau) = \tau$, the path integral (7.22) reduces back to the path integral (5.12) in physical configuration space.

8. PATH INTEGRALS IN EXTENDED CONFIGURATION SPACE WITH LAPSE

The only form of the action remaining in Table I is the Lagrangian action on extended configuration space with the lapse multiplier, Eqs. (3.15)–(3.16). We now represent the quantum propagator by a path integral whose phase is this action.

We start from the classical propagator (7.10) in which the integration over longitudinal part of the momentum $\Pi_{\|}$ has been performed, but which still depends on the transversal momenta Π_a. Instead of integrating over all transversal momenta Π_a [which would lead us back to the classical propagator (7.20)], we decompose Π_a into a component parallel to the velocity $\dot{Q}^a$ and $n-1$ components perpendicular to $\dot{Q}^a$. We choose a basis Q^α_a, $\alpha = 1,\ldots,N-1$, in the subspace perpendicular to $\dot{Q}^a$,

$$\dot{Q}^a Q^\alpha_a = 0, \tag{8.1}$$

and write

$$\Pi_a = N^{-1}\dot{Q}_a + \Pi_\alpha Q^\alpha_a. \tag{8.2}$$

The Jacobian of this transformation is (Appendix B)

$$\det \frac{\partial\{\Pi_a\}}{\partial\{N,\Pi_\alpha\}} = \det \begin{vmatrix} -N^{-2}\dot{Q}_a \\ Q^\alpha_a \end{vmatrix} = -N^{-2}\tilde{J}, \tag{8.3}$$

with

$$\tilde{J} = (1/(n-1)!)\delta^{aa_1\cdots a_{n-1}}\dot{Q}_a Q^{\alpha_1}_{a_1}\cdots Q^{\alpha_{n-1}}_{a_{n-1}}\delta_{\alpha_1\cdots\alpha_{n-1}}. \tag{8.4}$$

The last equation is the counterpart of Eq. (7.7) in a space of lower dimension.

Expressing the phase of the propagator (7.10) in terms of our new variables, we find

$$\begin{aligned}
&[\Pi_a\dot{Q}^a - \tfrac{1}{2}(T_C\dot{Q}^C)\overline{G}^{ab}\Pi_a\Pi_b + \phi_A\dot{Q}^A]\,\Delta\tau \\
&\quad = -\tfrac{1}{2}(T_C\dot{Q}^C)\overline{G}^{\alpha\beta}\Pi_\alpha\Pi_\beta\,\Delta\tau \\
&\qquad + [(N^{-1} - \tfrac{1}{2}T_C\dot{Q}^C N^{-2})G_{ab}\dot{Q}^a\dot{Q}^b + \phi_A\dot{Q}^A]\,\Delta\tau \\
&\quad = -\tfrac{1}{2}(T_C\dot{Q}^C)\overline{G}^{\alpha\beta}\Pi_a\Pi_\beta\,\Delta\tau + L(Q,\dot{Q},N)\,\Delta\tau.
\end{aligned} \tag{8.5}$$

The metric coefficient $\overline{G}^{\alpha\beta}$ is the projection

$$\overline{G}^{\alpha\beta} = \overline{G}^{ab}Q^\alpha_a Q^\beta_b \tag{8.6}$$

and $L(Q,\dot{Q},N)$ is the Lagrangian (3.15) with the lapse function N. The propagator (7.10) thereby assumes the form

$$\begin{aligned}
&C(\overline{Q}\,|Q,N,\Pi_\alpha) \\
&\quad = -N^{-2}\tilde{J}(Q)C(\overline{Q}\,|Q,\Pi_a) \\
&\quad = (2\pi)^{-n}(-N^{-2})\tilde{J}(Q)J(Q)\Lambda(Q)\delta(t(Q)-t(\tau)) \\
&\qquad \times e^{iL(Q,\dot{Q},N)\,\Delta\tau}e^{-(1/2)i(T_C\dot{Q}^C)G^{\alpha\beta}\Pi_\alpha\Pi_\beta\Delta\tau}.
\end{aligned} \tag{8.7}$$

We now evaluate the Gaussian integral over the momenta Π_α and find

$$\begin{aligned}
&\int d^{n-1}\Pi\, e^{-(1/2)i(T_C\dot{Q}^C)\,\Delta\tau\, G^{\alpha\beta}\Pi_\alpha\Pi_\beta} \\
&\quad = (2\pi)^{(n-1)/2}(iT_C\dot{Q}^C\Delta\tau)^{-(n-1)/2}\det{}^{1/2}\overline{G}_{\alpha\beta}.
\end{aligned} \tag{8.8}$$

Taking into account Eqs. (B24) and (7.14),

$$\tilde{J}J\det{}^{1/2}\overline{G}_{\alpha\beta} = \overline{D}^{-1/2}(G_{AB}\dot{Q}^A\dot{Q}^B)^{1/2}. \tag{8.9}$$

This sequence of operations leads us to the classical propagator

$$\begin{aligned}
&C(\overline{Q}\,|Q,N) = \int d^{n-1}\Pi\, C(\overline{Q}\,|Q,N,\Pi_\alpha) \\
&(-N^{-2})(2\pi)^{-1}(2\pi iT_C\dot{Q}^C\Delta\tau)^{-(n-1)/2}\overline{D}^{-1/2}(\overline{Q}\,|Q) \\
&\quad \times(G_{AB}\dot{Q}^A\dot{Q}^B)^{1/2}\Lambda(Q)\delta(t(Q)-t(\tau))e^{iL(Q,\dot{Q},N)\,\Delta\tau}.
\end{aligned} \tag{8.10}$$

All velocities $\dot{Q}^A$ in Eq. (8.10) are expressed in terms of boundary data, Eq. (6.13).

We can now represent the quantum propagator by the path integral

$$\begin{aligned}
&\langle Q''|Q'\rangle\delta(t(Q')-t')\,d^{n+1}Q' = \int \overline{D}Q\,\overline{D}N\,e^{iS[Q,N]} \\
&\quad \equiv \lim_{\Delta\tau_{\mathrm{MAX}}\to 0}\int\prod_{K=0}^{N-1} d^{n+1}Q_{(K)}\,dN_{(K)} \\
&\qquad \times C(Q_{(K+1)}|Q_{(K)},N_{(K)}).
\end{aligned} \tag{8.11}$$

The integral in Eq. (8.11) is over all $N_{(K)}$, $K = 0,1,\ldots,N-1$, but only over the interpolated $Q_{(I)}$, $I = 1,\ldots,N$. This corresponds to the fact that the momentumlike multiplier N has free ends.

The Lagrangian action $S[Q,N]$ is skeletonized by the prescription

$$\begin{aligned}
S[Q,N] \approx \sum_{K=0}^{N-1}\{&(N^{-1}_{(K)} - \tfrac{1}{2}T_C(Q_{(K)})\dot{Q}^C_{(K)}N^{-2}_{(K)}) \\
&\times G_{AB}(Q_{(K)})\dot{Q}^A_{(K)}\dot{Q}^B_{(K)} + \phi_A(Q_{(K)})\dot{Q}^A_{(K)}\}\,\Delta\tau,
\end{aligned} \tag{8.12}$$

where $\dot{Q}_{(K)}$ are again interpreted in terms of the configuration data $Q_{(K)}, Q_{(K+1)}$ at the boundaries of each step by Eq. (6.13).

The measure is skeletonized by the product

$$\begin{aligned}
\overline{D}Q\,\overline{D}N \approx \prod_{K=0}^{N-1} &d^{n+1}Q_{(K)}\,dN_{(K)}\,(-N^{-2}_{(K)}) \\
&\times(2\pi)^{-1}[2\pi iT_C(Q_{(K)})\dot{Q}^C_{(K)}\,\Delta\tau_{(K)}]^{-(n-1)/2} \\
&\times\overline{D}^{-1/2}(Q_{(K+1)}|Q_{(K)})(G_{AB}(Q_{(K)})\dot{Q}^A_{(K)}\dot{Q}^B_{(K)})^{1/2} \\
&\times\Lambda(Q_{(K)})\delta(t(Q_{(K)}) - t(\tau_{(K)})).
\end{aligned} \tag{8.13}$$

The modified metric coefficient enters into the measure (8.13) through the determinant (7.19).

These expressions simplify considerably in the special coordinate system, but, before showing this, let us recover the path integral in the extended configuration space by performing the integrations over $N_{(K)}$. To do this, we write the phase of the classical propagator (8.10) in the form

$$\begin{aligned}
&L(Q,\dot{Q},N)\,\Delta\tau = L(Q,\dot{Q})\,\Delta\tau \\
&\quad - \tfrac{1}{2}(N^{-1}(T_C\dot{Q}^C)^{1/2} - (T_C\dot{Q}^C)^{-1/2})^2 G_{AB}\dot{Q}^A\dot{Q}^B\Delta\tau,
\end{aligned} \tag{8.14}$$

where $L(Q,\dot{Q})$ is the homogeneous Lagrangian (3.19). We replace N by a new variable

$$M = N^{-1}(T_C\dot{Q}^C)^{1/2} - (T_C\dot{Q}^C)^{-1/2} \tag{8.15}$$

and write

$$\begin{aligned}
C(\overline{Q}\,|Q,N)\,dN &= C(\overline{Q}\,|Q,N)(-N^2(T_C\dot{Q}^C)^{-1/2})\,dM \\
&= dM\,e^{-(1/2)iG_{AB}\dot{Q}^A\dot{Q}^B\Delta\tau M^2} \\
&\quad\times(2\pi)^{-1}(T_C\dot{Q}^C)^{-1/2}(G_{AB}\dot{Q}^A\dot{Q}^B)^{1/2} \\
&\quad\times(2\pi iT_C\dot{Q}^C\Delta\tau)^{-(n-1)/2}\overline{D}^{-1/2}\Lambda \\
&\quad\times\delta(t(Q)-t(\tau))e^{iL(Q,\dot{Q})\,\Delta\tau}.
\end{aligned} \tag{8.16}$$

Integration over M yields the Gaussian integral

$$\int dM\,e^{-(1/2)iG_{AB}\dot{Q}^A\dot{Q}^B\Delta\tau M^2} = \pi^2(\tfrac{1}{2}i\,\Delta\tau)^{-1/2}(G_{AB}\dot{Q}^A\dot{Q}^B)^{-1/2}, \tag{8.17}$$

and the classical propagator (8.16) reduces back to the classical propagator (7.20).

The classical propagator (8.10) again simplifies in special coordinates (2.3). Taking into account Eqs. (7.21), (7.26) and rescaling the lapse multiplier,

$$N^{(o)} = AN, \tag{8.18}$$

we get

$$C(\bar{t},\bar{q}|t,q,N^{(o)})\,dN^{(o)}$$
$$= -dN^{(o)}N^{(o)-2}(2\pi)^{-1}(2\pi i\dot{t}(\tau)\Delta\tau)^{-(n-1)/2}$$
$$\times\bar{g}^{1/2}(\bar{t},\bar{q}|t,q)[g_{ab}(q)\dot{q}^a\dot{q}^b]^2\delta(t-t(\tau))e^{iL(t,q,\dot{q},N^{(o)})\Delta\tau}, \tag{8.19}$$

where $L(t,q,\dot{q},N^{(o)})$ is the action (3.17).

The velocity $\dot{q}^a$ is again interpreted by Eq. (6.5). The measure (8.13) in the path integral (8.11) reduces thereby to

$$\bar{D}Q\,\bar{D}N^{(o)} \approx \prod_{K=0}^{N-1} dt_{(K)}\,d^nq_{(K)}\,dN^{(o)}_{(K)}$$
$$-N^{(o)-2}_{(K)}(2\pi)^{-1}(2\pi i\dot{t}(\tau_{(K)})\Delta\tau_{(K)})^{-(n-1)/2}$$
$$\times\bar{g}^{1/2}(t_{(K+1)},q_{(K+1)}|t_{(K)},q_{(K)})$$
$$\times(g_{ab}(q_{(K)})\dot{q}^a_{(K)}\dot{q}^b_{(K)})^{1/2}\delta(t_{(K)}-t(\tau_{(K)})). \tag{8.20}$$

This completes our program. We have represented the quantum propagator by path integrals corresponding to all action functionals enumerated in Table II.

9. PATH INTEGRALS: PARAMETRIZATION VERSUS GAUGE

We have now learned how to write the quantum propagator for a parametrized system as a path integral in extended phase space. Our prescription, Eqs. (6.16)–(6.17), recognizes the need to enforce the Hamiltonian constraint and to select a definite parametrization of the path. These two aims are achieved by the delta functions $\delta(\bar{H}(Q_{(K+1)}|Q_{(K)},P_{(K)}))$ and $\delta(t(Q_{(K)})-t(\tau_{(K)}))$ in the skeletonized measure.

A similar need arises in gauge theories. One must enforce the constraints generating gauge transformations, and one should fix the gauge when writing the path integral in the space of redundant variables. It is of interest to compare the algorithm which we have obtained for a parametrized theory with the standard prescription for gauge theories.

Let us first review the basic structure of gauge theories. To bring out the issues clearly, we consider again our old finite-dimensional nonrelativistic system. We can turn it into a gauge theory by adjoining an additional spurious gauge coordinate ϕ to the physical coordinates q^a. This brings us to the extended configuration space $\{q^a,\phi\}$. As q^a is kept fixed and ϕ is varied, we move along a fiber over q^a. We interpret all points in such a fiber as different descriptions of the same physical state.

As the state of the system evolves in time, the choice of the gauge variable remains arbitrary. In other words, the velocity

$$d_t\phi(t)=\lambda(t) \tag{9.1}$$

can be freely prescribed at each step of the dynamical evolution. Equation (9.1) can be obtained by varying the action

$$\sigma[\phi,\pi;\lambda] = \int_{t'}^{t''} dt\,(\pi\,d_t\phi-\lambda\pi) \tag{9.2}$$

with respect to the gauge momentum π. By varying (9.2) with respect to λ and ϕ, we learn that the momentum π is constrained to vanish,

$$\pi=0, \tag{9.3}$$

and continues to vanish in the course of time.

The evolution of the system in the extended phase space $\{q^a,\phi,p_a,\pi\}$ is then described by the action S which is the sum of the physical action (2.1) and the gauge action (9.2)

$$S[q^a,\phi,p_a,\pi;\lambda] = \int_{t'}^{t''} dt\,(p_a d_t q^a+\pi\,d_t\phi-h(t,q,p)-\lambda\pi). \tag{9.4}$$

After an arbitrary point transformation in the extended phase space,

$$Q^A=Q^A(q^a,\phi),\quad p_a=Q^A{}_{,a}P_A,\quad \pi=Q^A{}_{,\phi}P_A, \tag{9.5}$$

the action (9.5) assumes the form

$$S[Q^A,P_A;\lambda] = \int_{t'}^{t''} dt\,(P_A d_t Q^A - h(Q,P)-\lambda\pi(Q,P)). \tag{9.6}$$

The action (9.6) can be modified in two ways without changing the equations of motion. The constraint (9.3) can be scaled by an arbitrary factor $\Lambda(Q)\neq 0$,

$$\Pi=\Lambda(Q)\pi(Q,P), \tag{9.7}$$

and it can be adjoined to the physical Hamiltonian h,

$$\tilde{h}=h+[k^A(Q)P_A+k(Q)]\pi(Q,P). \tag{9.8}$$

We have chosen the coefficients $\Lambda(Q)$ and $k^A(Q)P_A+k(Q)$ so that the new constraint Π is still linear in the momenta P_A and the new Hamiltonian $\tilde{h}$ is still quadratic in the momenta P_A.

The constraints (9.3) or (9.7) generate the gauge transformation of the canonical variables Q^A,P_A. Such a transformation does not change the physical state of the system. To single out a particular representative for each physical state, one can introduce a gauge fixing condition

$$\Phi(Q^A,P_A)=0. \tag{9.9}$$

Here, Φ is any function which yields a unique value of the gauge coordinate ϕ when Eqs. (9.3) and (9.5) are taken into account.

We can write now the standard prescription for the quantum propagator as a path integral in the extended phase space $\{Q^A,P_A\}$ of the gauge theory. The propagator has the form (4.13) with the classical propagator

$$C(\bar{Q}|Q,P)=(2\pi)^{-n}\delta(\Phi)\delta(\Pi)|[\Phi,\Pi]|e^{iS(\bar{t},\bar{Q}|t,Q,P)}, \tag{9.10}$$

corresponding to the skeletonized canonical action with the Hamiltonian (9.8).

The prescription (9.10) is superficially similar in form to our result (6.14) for the classical propagator of a parametrized theory. The gauge constraint $\Pi=0$ plays the role of the super-Hamiltonian constraint $H=0$ and the gauge fixing condition (9.9) replaces the condition

$$t(Q)-t(\tau)=0, \tag{9.11}$$

which selects the parametrization of path. [Due to Eq. (6.15), the factor Λ in the measure (6.17) has the meaning of the Poisson bracket between the expression (9.11) and the super-Hamiltonian H]. However, there are two important differences:

(I) The gauge fixing condition does not need to contain any reference to time. On the other hand, the condition (9.11) selecting the parametrization must introduce a prescribed function $t(\tau)$ of τ.

(II) In gauge theories, any function (9.9) of the extended coordinates and momenta is permissible. On the other hand, in a parametrized theory $t(Q)$ is a definite function on the extended configuration space. For our Newtonian system, the time function $t(Q)$ is obtained by the reconstruction procedure discussed in Sec. 2.

To see that the distinction (I) is vital, let us blindly apply a condition (9.9) appropriate for a gauge theory to our parametrized theory. In the simplest case, this is achieved by putting $t(\tau) = 0$ and identifying $t(Q)$ with $\Phi(Q,P)$. Of course, our derivation of Eqs. (6.16)–(6.17) for the quantum propagator is no longer valid because $t(\tau) = 0$ implies $t' = 0 = t''$. When we insist that the expression (6.16)–(6.17) represents the quantum propagator from t' to $t'' > t'$ even for $t(\tau) = 0$, we predictably end with an absurd result. On the other hand, when we put $t(\tau) = 0$ and simultaneously restrict ourselves to $t' = 0 = t''$, the expression (6.16)–(6.17) for the quantum propagator equally predictably yields a correct triviality: It reduces to the delta function because the dynamics is frozen at a single instant of time.

The distinction (I) reflects the fundamental physical difference between gauge theories and parametrized theories. The constraints which follow from gauge invariance generate gauge changes of the extended phase space variables. These are unobservable; the physical state of the system is unchanged. The constraints which follow from reparametrization invariance generate the dynamics of the system. They are observable and the physical state does change. It makes sense to fix a gauge to get one representative to a physical state. It makes no sense to fix the time.

The distinction (II) is more subtle. It means that the slices of a constant label time τ coincide with the leaves of the absolute time foliation. Such a restriction follows naturally from our derivation of Eqs. (6.16)–(6.17) for the quantum propagator. There is no simple modification of this derivation which would introduce a different foliation, e.g.,

$$\Phi(Q,\tau) = 0. \tag{9.12}$$

In fact, the Schrödinger equation ceases to be a first-order equation in the foliation label when we allow the general foliation (9.12) and the Hilbert space interpretation loses thereby its meaning. We thus consider it highly unlikely that the general foliation (9.12) would yield the correct quantum propagator when used instead of the absolute time foliation (9.11) in the expressions (6.16)–(6.17) for the path integral. We emphasize yet again that the choice of the time variable is a central decision in forming quantum theories and that, once made, it cannot be easily altered without altering the theory.

10. SUMMARY

The representation of the quantum propagator by a path integral of the exponentiated canonical action on the physical phase space is a natural starting point for quantum mechanics. The measure in the space of paths is induced by the invariant Liouville measure in phase space. The geometrically privileged transport of momentum by actual classical paths of the system leads to the skeletonization of the canonical action by the chain of phase-space principal functions. This privileged skeletonization removes the ambiguity connected with the factor ordering.

Unfortunately, not all classical theories are easily formulated in terms of the true physical degrees of freedom. Both gauge theories and parametrized theories use redundant variables. The dynamical evolution of the system takes place in extended spaces of variables. General relativity is the most prominent example of a system in which the simultaneous presence of gauge and parametrization makes it extremely difficult to return back to the physical phase space. It is thus essential to represent the quantum propagator by integrals over paths in such extended spaces of variables.

We have accomplished this program for parametrized Newtonian systems moving in curved configuration spaces. Our point of departure was the path integral in the physical phase space of the system. We arrived at equivalent path integrals in alternative spaces by extending or restricting the variables.

The extension of variables was always done so that integration over the new variables yielded the integral we have started from. Typical devices for ensuring this property are delta functions introduced into the measure or representations of known functions by integrals over a parameter. The restriction of variables was always carried out by integrating over them. Typically, the integrals involved were Gaussian integrals in the momenta which can be explicitly evaluated. Such integrals lead to nontrivial measures in spaces of remaining variables.

We summarize our results in Table III, which is a continuation of our Table II for the alternative forms of the action. In the first column, we write down a symbolic expression for the path integrals. The symbolic expression is interpreted by skeletonizing the measure and skeletonizing the action. In the second column, we enter the measure associated with a segment of skeletonized path between the gate $dX = dX_{(K)}$ at $X = X_{(K)}$ and the gate $d\overline{X} = dX_{(K+1)}$ at $\overline{X} = X_{(K+1)}$ in the space $\{X\}$ of appropriate variables. The total measure is the product of such elementary measures at all gates, $K = 0,1,\ldots,N-1$. In the following column, we give the number of the equation which introduces this measure in the main text. Some of the measures are quite complicated and do not follow a clearly recognizable pattern. On the other hand, the classical action is always skeletonized in the same manner: For each step of the skeletonized path, we write the initial value $L_{(K)}$ of the appropriate Lagrangian and multiply it by the interval $\Delta\tau_{(K)} = \tau_{(K+1)} - \tau_{(K)}$ of time. The initial values of velocities which enter into the Lagrangian must be expressed in terms of the configuration data at the boundaries of each step. This is achieved by using the appropriate Hamilton principal function obeying the standard Hamilton–Jacobi equations. Moreover, in the phase space versions of the theory, the initial metric entering into the Lagrangian must be replaced by a tensor–scalar coefficient which takes into account the geodesic deviation trans-

TABLE III. Alternative forms for path integrals.

Type of action	Quantum propagator represented by the path integral	Elementary measure of a segment of path	Skeletonized measure: Eq. number	Skeletonized action: Eq. number
Physical canonical action	$\langle t'', q''\|t', q'\rangle\, d^n q' = \int Dq\, Dp\, e^{iS[q,p]}$	$d^n q\, d^n p (2\pi)^{-n}$	(4.12)	(4.11)
Physical Lagrangian action	$\langle t'', q''\|t', q'\rangle\, d^n q' = \int \bar{D}q\, e^{iS[q]}$	$d^n q\, (2\pi i\, \Delta t)^{-n/2}\, \bar{g}^{1/2}(\bar{t}, \bar{q}\|t, q)$	(5.14)	(5.13)
Extended canonical action, conditional	$\langle Q''\|Q'\rangle \delta(t(Q') - t')\, d^{n+1} Q' = \int \bar{D}Q\, \bar{D}P\, e^{iS[Q,P]}$	$d^{n+1} Q\, d^{n+1} P (2\pi)^{-n} \Lambda(Q)$ $\delta(t(Q) - t(\tau))\delta(\bar{H}(\bar{Q}\|Q, P))$	(6.17)	(6.18)
Extended Lagrangian action, with lapse multiplier	$\langle Q''\|Q'\rangle \delta(t(Q') - t')\, d^{n+1} Q'\, dN' = \int \bar{D}Q\, \bar{D}P\, \bar{D}N\, e^{iS[Q,N]}$	$d^{n+1} Q\, d^{n+1} P\, dN\, \Delta t\, \Lambda(Q)$ $(2\pi)^{-(n+1)} \delta(t(Q) - t(\tau))$	(6.23)	(6.24)
Extended Lagrangian action, homogeneous	$\langle Q''\|Q'\rangle \delta(t(Q') - t')\, d^{n+1} Q' = \int \bar{D}Q\, e^{iS[Q]}$	$d^{n+1} Q\, [2\pi i T_C(Q) \dot{Q}^C \Delta\tau]^{-n/2}$ $\bar{D}^{-1/2}(\bar{Q}\|Q)\, \Lambda(Q) \delta(t(Q) - t(\tau))$	(7.24)	(7.23)
		$dt\, d^n q [2\pi i \dot{t}(\tau) \Delta\tau]^{-n/2}\, \bar{g}(\bar{t}, \bar{q}\|t, q) \delta(t - t(\tau))$	(7.28)	(7.29)
Extended Lagrangian action, with lapse multiplier	$\langle Q''\|Q'\rangle \delta(t(Q') - t')\, d^{n+1} Q'\, dN' = \int \bar{D}Q\, \bar{D}N\, e^{iS[Q,N]}$	$D^{n+1} Q\, dN (-N^{-2}) (2\pi)^{-1}$ $(2\pi i T_C(Q) \dot{Q}^C \Delta\tau)^{-(n-1)/2}$ $\bar{D}^{-1/2}(\bar{Q}\|Q) (G_{AB}(Q) \dot{Q}^A \dot{Q}^B)^{1/2}$ $\Lambda(Q) \delta(t(Q) - t(\tau))$	(8.13)	(8.12)
		$dt\, d^n q\, dN (-N^{(0)-2}) (2\pi)^{-1}$ $(2\pi i \dot{t}(\tau) \Delta\tau)^{-(n-1)/2}\, \bar{g}^{1/2}(\bar{t}, \bar{q}\|t, q)$ $(g_{ab}(q) \dot{q}^a \dot{q}^b)^{1/2} \delta(t - t(\tau))$	(8.20)	

port of momenta. The classical action is skeletonized by the sum $\Sigma_{K=0}^{N-1} L_{(K)}\, \Delta\tau_{(K)}$ of such contributions. Because the procedure follows a well-defined algorithm, there is no need to enter the skeletonized action into our table. We refer merely to the equation where it is discussed in the paper.

While the measures are often complicated, they have one feature in common—the occurrence of $\delta(t(Q) - t(\tau))$ which fixes the integrations to the leaves of absolute time that flows from the initial instant t' to the final instant t''. The specific form of this delta function is characteristic of parametrized theories and reflects the privileged role time plays in quantum mechanics.

ACKNOWLEDGMENT

This work was supported in part by NSF Grants PHY 81-06909, PHY 80-26043, and PHY 81-07384.

APPENDIX A: INTEGRABILITY CONDITIONS ON THE DEGENERATE METRIC G^{AB}

A degenerate metric G^{AB} with signature $(0; +, \ldots, +)$ has a unique degeneracy direction, i.e., the solutions T_A to the equation

$$G^{AB} T_B = 0 \tag{A1}$$

fill a ray. The ray determines a foliation if and only if it is surface forming,

$$M_{ABC} \equiv T_A T_{[B,C]} + T_B T_{[C,A]} + T_C T_{[A,B]} = 0. \tag{A2}$$

To be so, the metric G^{AB} cannot be arbitrary, but it must satisfy certain integrability conditions which we are now going to derive.

Note that the equation

$$T_A X^A = 0 \tag{A3}$$

has n linearly independent solutions Q^A_a, $a = 1, \ldots, n$ and that the metric G^{AB} is nondegenerate on the vector subspace spanned by Q^A_a:

$$G^{AB} = G^{ab} Q^A_a Q^B_b, \quad \det G^{ab} \neq 0. \tag{A4}$$

Let U^A be an arbitrary vector linearly independent of Q^A_a, i.e.,

$$T_A U^A \neq 0. \tag{A5}$$

The vectors $\{U^A, Q^A_a\}$ form a basis. Because G^{ab} is nondegenerate, any equation $M_A = 0$ can be replaced by an equivalent set of equations

$$G^{AB} M_B = 0, \quad U^B M_B = 0. \tag{A6}$$

Handling each index of the completely antisymmetric tensor M_{ABC} in this way, we can replace Eq. (A2) by an equivalent system of equations:

$$G^{KA} G^{LB} G^{MC} M_{ABC} = 0, \tag{A7}$$

$$G^{KA} G^{LB} U^C M_{ABC} = 0. \tag{A8}$$

Due to Eq. (A1), the condition (A7) is identically satisfied.

Further, because of Eqs. (A1) and (A5), the condition (A8) reduces to

$$G^{KA}G^{LB}T_{[A,B]} = 0. \tag{A9}$$

Using Eq. (A1) again, we cast Eq. (A9) into the form

$$G^{A[K,L]}T_A = 0, \tag{A10}$$

where

$$G^{AK,L} \equiv G^{AK}{}_{,B}G^{BL}. \tag{A11}$$

From Eqs. (A3) and (A5) we see that

$$\exists\ H^{KLa}:\ G^{A[K,L]} = H^{KLa}Q^A_a. \tag{A12}$$

An alternative way of writing Eq. (A12) is

$$\delta_{AA_1\cdots A_n}G^{A[B,C]}G^{A_1B_1}\cdots G^{A_nB_n} = 0. \tag{A13}$$

Here $\delta_{AA_1\cdots A_n}$ is a completely antisymmetric tensor density of weight -1 with $\delta_{012\cdots n} = 1$. Note that in a Newtonian space-time we cannot introduce the more usual Levi-Civita pseudotensor $\epsilon_{AA_1\cdots A_n}$ because the metric G^{AB} is degenerate.

Equation (A13) is equivalent to the condition (A12) which is a necessary and sufficient condition for the degeneracy covector T_A determined by Eq. (A1) to be surface-forming.

APPENDIX B: DETERMINANTS WITH DEGENERATE METRICS

The metric G^{AB} is degenerate, and its determinant thus vanishes. However, we can project G^{AB} into the subspace orthogonal to the degeneracy direction T_A and take the determinant of the projected metric.

For a given G^{AB} and U^A, Eqs. (2.21) and (2.24) have a unique solution T_A. Furthermore, the equation

$$U^A X_A = 0 \tag{B1}$$

has n linearly independent solutions Q^a_A, $a = 1,\ldots,n$:

$$U^A Q^a_A = 0. \tag{B2}$$

The covectors $\{T_A, Q^a_A\}$ form a cobasis. Of course, Q^a_A can be changed by a transformation

$$Q^{a^*}_A = \Lambda^{a^*}_b(Q)Q^b_A. \tag{B3}$$

We introduce the alternating symbol $\delta_{a_n\cdots a_n}$ which transforms as a tensor density of weight -1 under the Λ transformations (B3). Besides it, we have at our disposal the alternating symbol $\delta^{AA_1\cdots A_n}$, which transforms as a tensor density of weight 1 under transformations of extended coordinates.

The projection

$$G^{ab} \equiv G^{AB}Q^a_A Q^b_B \tag{B4}$$

of the degenerate metric G^{AB} is nondegenerate, and we can write its determinant as

$$G^{-1} = (1/n!)\delta_{a_1\cdots a_n}G^{a_1b_1}\cdots G^{a_nb_n}\delta_{b_1\cdots b_n}. \tag{B5}$$

In terms of the original metric,

$$G^{-1} = (1/n!)\delta_{a_1\cdots a_n}Q^{a_1}_{A_1}\cdots Q^{a_n}_{A_n} \times G^{A_1B_1}\cdots G^{A_nB_n}\delta_{b_1\cdots b_n}Q^{b_1}_{B_1}\cdots Q^{b_n}_{B_n}. \tag{B6}$$

Study now the expression

$$D = (1/n!)\delta_{AA_1\cdots A_n}U^A U^B G^{A_1B_1}\cdots G^{A_nB_n}\delta_{BB_1\cdots B_n}. \tag{B7}$$

The tensor density $U^A\delta_{AA_1\cdots A_n}$ has two properties: (1) It is completely antisymmetric in $A_1,\ldots,A_n$, and (2) it is orthogonal to U^A. As a consequence, we must have

$$U^A\delta_{AA_1\cdots A_n} = J^{-1}\delta_{a_1\cdots a_n}Q^{a_1}_{A_1}\cdots Q^{a_n}_{A_n}. \tag{B8}$$

To determine the proportionality factor J^{-1}, we multiply Eq. (B8) by $\delta^{BA_1\cdots A_n}$. Because

$$\delta_{AA_1\cdots A_n}\delta^{BA_1\cdots A_n} = n!\delta^B_A, \tag{B9}$$

we get

$$n!U^B = J^{-1}\delta_{a_1\cdots a_n}Q^{a_1}_{A_1}\cdots Q^{a_n}_{A_n}\delta^{BA_1\cdots A_n}. \tag{B10}$$

Multiplication by T_B yields

$$J = (1/n!)\delta_{a_1\cdots a_n}T_A Q^{a_1}_{A_1}\cdots Q^{a_n}_{A_n}\delta^{AA_1\cdots A_n}. \tag{B11}$$

By introducing Eqs. (B8) and (B11) into the expression (B7), we learn that

$$G^{-1} = J^2 D. \tag{B12}$$

Any covector Π_A can be split into a part along T_A and a part perpendicular to U^A,

$$\Pi_A = \Pi_\parallel T_A + \Pi_A Q^a_A. \tag{B13}$$

Equation (B13) can be considered as a transformation from the variables $\Pi_\parallel, \Pi_a$ to the variables Π_A. The Jacobi matrix of this transformation is

$$\frac{\partial\{\Pi_A\}}{\partial\{\Pi_\parallel,\Pi_a\}} = \begin{vmatrix} T_A \\ Q^a_A \end{vmatrix}. \tag{B14}$$

We see that J is nothing else but the Jacobian of the transformation (B13).

We can replace the metric G^{AB} by the tensor–scalar coefficient $\overline{G}^{AB}$ and introduce appropriate quantities (B4), (B5), and (B7). We place bars over symbols denoting these quantities: $\overline{G}^{ab}, \overline{G}, \overline{D}$. The modified quantities are again connected by the equation

$$\overline{G}^{-1} = J^2\overline{D}. \tag{B15}$$

Mutatis mutandis, the same line of reasoning applies to nondegenerate metrics. Take a regular metric G^{ab}, $a = 1,\ldots,n$, and a vector $\dot{Q}^a$. Let Q^α_a, $\alpha = 1,\ldots,n-1$, be a basis in cotangent space orthogonal to $\dot{Q}^a$:

$$\dot{Q}^a Q^\alpha_a = 0. \tag{B16}$$

Project the metric G^{ab},

$$G^{\alpha\beta} \equiv G^{ab}Q^\alpha_a Q^\beta_b. \tag{B17}$$

The projected metric $G^{\alpha\beta}$ is again regular, and we can introduce its inverse $G_{\alpha\beta}$. Greek indices are raised by $G^{\alpha\beta}$ and lowered by $G_{\alpha\beta}$. Similarly, Latin indices are raised by G^{ab} and lowered by G_{ab}. With this convention,

$$G_{ab} = G_{\alpha\beta}Q^\alpha_a Q^\beta_b + \dot{Q}^{-2}\dot{Q}_a\dot{Q}_b, \tag{B18}$$

with

$$\dot{Q}^2 \equiv g_{ab}\dot{Q}^a\dot{Q}^b. \tag{B19}$$

We take the determinant of Eq. (B18). Because $G_{\alpha\beta}Q^\alpha_a Q^\beta_b$ and $\dot{Q}_a\dot{Q}_b$ are degenerate matrices,

$$G = (1/n!)\delta^{aa_1\cdots a_{n-1}} G_{ab} G_{a_1 b_1} \cdots G_{a_{n-1} b_{n-1}} \delta^{bb_1\cdots b_{n-1}}$$
$$= [1/(n-1)!]\dot{Q}^{-2}\delta^{aa_1\cdots a_{n-1}}\delta^{bb_1\cdots b_{n-1}}$$
$$\times \dot{Q}_a \delta^{aa_1\cdots a_{n-1}} Q^{\alpha_1}_{a_1}\cdots Q^{\alpha_{n-1}}_{a_{n-1}}$$
$$\times \dot{Q}_b \delta^{bb_1\cdots b_{n-1}} Q^{\beta_1}_{b_1}\cdots Q^{\beta_{n-1}}_{b_{n-1}} G_{\alpha_1\beta_1}\cdots G_{\alpha_{n-1}\beta_{n-1}}. \quad \text{(B20)}$$

As in Eqs. (B8) and (B11),

$$\dot{Q}_a \delta^{aa_1\cdots a_{n-1}} Q^{\alpha_1}_{a_1}\cdots Q^{\alpha_{n-1}}_{a_{n-1}} = \tilde{J}\delta^{\alpha_1\cdots\alpha_{n-1}}, \quad \text{(B21)}$$

with

$$\tilde{J} = [1/(n-1)!]\delta^{aa_1\cdots a_{n-1}}\dot{Q}_a Q^{\alpha_1}_{a_1}\cdots Q^{\alpha_{n-1}}_{a_{n-1}}\delta_{\alpha_1\cdots\alpha_{n-1}}. \quad \text{(B22)}$$

As a result,

$$G = \tilde{J}^2\dot{Q}^{-2}\det G_{\alpha\beta}. \quad \text{(B23)}$$

We multiply Eq. (B12) by Eq. (B23) and conclude that

$$\tilde{J}J \det{}^{1/2} G_{\alpha\beta} = (G_{AB}\dot{Q}^A\dot{Q}^B)^{1/2} D^{-1/2}. \quad \text{(B24)}$$

[1]The literature on the implementation of quantum dynamics by path integrals for nonrelativistic and relativistic systems in curved and flat configuration spaces is extensive and too large to be cited here. A useful general survey with extensive references to the original literature is L. S. Shulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1982).

[2]K. Kuchař, J. Math. Phys. **24**, 2122 (1983).

[3]L. Faddeev, Teor. Mat. Fiz. **1**, 3 (1969); L. Faddeev and V. Popov, Phys. Lett. B **25**, 30 (1967); L. Faddeev and V. Popov, Usp. Fiz. Nauk **111**, 427 (1973) [Sov. Phys. Usp. **16**, 777 (1974)]; E. S. Fradkin, and G. A. Vilkovisky "Quantization of Relativistic Systems with Constraints, Equivalence of Canonical and Covariant Formalisms in the Quantum Theory of the Gravitational Field," CERN Report TH-2332, 1977; L. Faddeev and A. Slavnov, *Gauge Fields: Introduction to Quantum Theory* (Benjamin, Reading, MA, 1980).

[4]K. Kuchař, Phys. Rev. D **22**, 1285 (1980).

Quantum energy-entropy inequalities: A new method for proving the absence of symmetry breaking

M. Fannes,[a] P. Vanheuverzwijn,[b] and A. Verbeure
Instituut voor Theoretische Fysica, Universiteit Leuven, B-3030 Leuven, Belgium

(Received 25 January 1983; accepted for publication 10 June 1983)

For quantum systems we develop a new method, based on a general energy-entropy inequality, to rule out spontaneous breaking of symmetries. The main advantage of our scheme consists in its clear-cut physical significance and its new areas of applicability; in particular we can handle discrete symmetry groups as well as continuous ones. Finally a few illustrations are discussed.

PACS numbers: 03.65. − w, 05.50. + q, 02.20. + b

I. INTRODUCTION

In the case of classical lattice systems we derived recently[1] correlation inequalities expressing the balance between energy and entropy for an equilibrium state. These inequalities were shown to reproduce easily the sharpest results concerning spontaneous magnetization in long range Ising models[2] and they gave a more direct and intuitive understanding of the underlying physics. Maybe even more important is the applicability to continuous as well as to discrete symmetry groups. In particular we proved translation invariance for one-dimensional systems under very weak conditions on the potential.[1]

Here we are concerned with the quantum-mechanical situation. The well-known method to prove absence of symmetry breaking is based on the Bogoliubov inequality. The first results along this line are the celebrated theorems of Mermin–Wagner[3] and Hohenberg.[4] Recently there was a revival of interest in the field. The best results along this line can be found in Ref. 5. It is important to remark that this method is restricted to continuous symmetry groups as the occurrence of an infinitesimal generator is essential for the method. On the contrary our method allows also for discrete symmetries. To stress this fact we will concentrate on the applications to discrete symmetries.

Our main tool is the correlation inequality [see formula (2) below] which has a clear physical significance as being an expression for the change of free energy under a dissipative perturbation of the equilibrium state.[1,6]

One should mention here also the results based on relative entropy considerations.[7] This technique as well allows for the treatment of discrete symmetries; however, our method based on the inequality seems to us more direct and intuitive.

II. ABSENCE OF SYMMETRY BREAKING

Let $(\mathscr{A},\alpha_t)$ be a C^*-dynamical system, i.e., $\mathscr{A}$ a C^*-algebra and α_t ($t\in\mathbb{R}$) is a strongly continuous one-parameter group of *-automorphisms of $\mathscr{A}$. A state ω of $\mathscr{A}$ satisfies the KMS condition for the evolution α_t at inverse temperature β, if $\omega(x\,\alpha_{i\beta}(y)) = \omega(yx)$ for all x,y in a norm dense, α_t-invariant *-subalgebra of $\mathscr{A}$. Let $\mathfrak{H}$ be the GNS representation space of the state ω and $\Omega\in\mathfrak{H}$ the cyclic vector; we denote by $\mathscr{M}$ the von Neumann algebra $\mathscr{A}''$ and by H the infinitesimal generator of the time evolution on $\mathfrak{H}$. As ω is time invariant we have $\Omega\in\mathscr{D}(H)$ (domain of H) and $H\Omega = 0$.

If

$$H = \int_{-\infty}^{\infty} \lambda\, dE(\lambda)$$

is the spectral decomposition of the Hamiltonian H, define for all $x\in\mathscr{M}$ the measures on $\mathbb{R}$

$$d\mu_x(\lambda) = (x\Omega, dE(\lambda)x\Omega),$$
$$d\nu_x(\lambda) = (x\Omega, dE(-\lambda)x^*\Omega).$$

As ω is a KMS state the measures μ_x and ν_x are equivalent with Radon–Nikodym derivative

$$\frac{d\mu_x(\lambda)}{d\nu_x(\lambda)} = e^{\beta\lambda} \tag{1}$$

(see, e.g., Ref. 8, Proposition 5.3.14).

We start with an easy derivation of an inequality for KMS states which was stated implicitly for the first time in Ref. 9.

For all $x\in\mathscr{M}$ such that $x\Omega\in\mathscr{D}(H)$,

$$\begin{aligned}\frac{\beta(x\Omega,Hx\Omega)}{(x\Omega,x\Omega)} &= \frac{\int \beta\lambda\, d\mu_x(\lambda)}{\int d\mu_x(\lambda)} \\ &= -\ln\exp -\frac{\int \beta\lambda\, d\mu_x(\lambda)}{\int d\mu_x(\lambda)} \\ &\geqslant -\ln\frac{\int e^{-\beta\lambda}\, d\mu_x(\lambda)}{\int d\mu_x(\lambda)} \\ &= -\ln\frac{\int d\nu_x(\lambda)}{\int d\mu_x(\lambda)} \\ &= \ln\frac{(x\Omega,x\Omega)}{(x^*\Omega,x^*\Omega)}\end{aligned}$$

by the Jenssen inequality. Hence

$$\beta\omega(x^*Hx\Omega) \geqslant \omega(x^*x)\ln(\omega(x^*x)/\omega(xx^*)). \tag{2}$$

Lemma II.1: Let I be a finite interval of $\mathbb{R}$. If for $0\neq x\in\mathscr{M}\cap\mathscr{D}([H,\cdot])$ $\operatorname{supp}\mu_x \subset I$ then if ω satisfies the KMS condition,

$$0 \leqslant \beta\omega(x^*Hx) - \omega(x^*x)\ln\frac{\omega(x^*x)}{\omega(xx^*)} \leqslant \beta\omega(x^*x)\Delta,$$

where Δ is the length of the interval I.

Proof: Let $I = [\lambda_1,\lambda_2]$, $\lambda_i\in\mathbb{R}$; using (1) and (2) we compute

[a] Bevoegdverklaard navorser NFWO, Belgium.
[b] Aangesteld navorser NFWO, Belgium.

 0022-2488/84/010076-03$02.50

$$0 \leqslant \beta\omega(x^*Hx) - \omega(x^*x)\ln\frac{\omega(x^*x)}{\omega(xx^*)}$$

$$= \beta\int \lambda\, d\mu_x(\lambda) - \int d\mu_x(\lambda)\ln\frac{\int d\mu_x(\lambda)}{\int e^{-\beta\lambda}d\mu_x(\lambda)}$$

$$\leqslant \beta\lambda_2\int d\mu_x(\lambda) - \int d\mu_x(\lambda)\ln\frac{1}{e^{-\beta\lambda_1}}$$

$$= \beta(\lambda_2 - \lambda_1)\int d\mu_x(\lambda). \quad \blacksquare$$

Now we proceed to our main objective, namely, the development of a theory for the absence of spontaneous symmetry breaking. We suppose that we have a symmetry represented by a *-automorphism τ of $\mathscr{A}$ satisfying the following conditions:

(a) τ is approximately inner, i.e., there exists a sequence $(u_n)_{n\geqslant 1}$ of unitaries in $\mathscr{A}$ such that for all $x\in\mathscr{A}$,

$$\lim_{n\to\infty} \|\tau(x) - u_n^* x u_n\| = 0.$$

This condition implies

$$\lim_{n\to\infty} \omega(u_u^* x u_n) = \omega(\tau(x)) \tag{3}$$

for all states ω of $\mathscr{A}$. This notion of approximately inner automorphism has been introduced in Ref. 10. As far as the physics is concerned it means that the automorphism can be approximated by local unitary transformations.

(b) As τ represents a symmetry of the system we have $[\alpha_t,\tau] = 0$ for all $t\in\mathbb{R}$. Furthermore, we suppose that the local approximations almost commute with α_t in the sense that for all m: $u_m\in\mathscr{D}([H,\cdot])$ and

$$K = \sup_m \|[H,u_m^*]\| < \infty. \tag{4}$$

This is essentially the condition used in Ref. 7.

Theorem II.2: Let ω be a KMS state with respect to the evolution α_t at inverse temperature β; let τ be a symmetry as above. Then there exists a constant C such that for all $x\in\mathscr{A}$, $\omega(xx^*)\leqslant C\omega(\tau(xx^*))$ holds.

Proof: For $f\in C_c^\infty(\mathbb{R})$ (the space of infinitely differentiable functions with compact support) and for any $x\in\mathscr{A}$ we denote

$$x(f) = \int dt\,\hat{f}(t)\alpha_t(x),$$

where $f(\lambda) = \int dt\,\hat{f}(t)e^{it\lambda}$.

For $\epsilon > 0$ one finds a decomposition of the identity by a sequence $(h_n)_{n\geqslant 1}$ of positive functions in C_c^∞ such that pointwise $\Sigma_{n\geqslant 1} h_n^2 = 1$ and such that the support of each h_n is contained in an interval of length ϵ.

By a straightforward computation one gets

$$\omega(xx^*) = \int d\nu_x(-\lambda) = \sum_n \int h_n(\lambda)^2 d\nu_x(-\lambda)$$

$$= \sum_n \omega(x(h_n)x(h_n)^*). \tag{5}$$

Substitute in the correlation inequality (2) the observable x by $u_m^* x(h_n)$ for each n such that $x(h_n)\neq 0$; adding and subtracting a term and using time invariance

$$\omega(x(h_n)^*x(h_n))\ln\frac{\omega(x(h_n)x(h_n)^*)}{\omega(u_m^*x(h_n)x(h_n)^*u_m)}$$

$$-\beta\omega(x(h_n)^*u_m[H,u_m^*]x(h_n))$$

$$\leqslant \beta\omega(x(h_n)^*Hx(h_n)) - \omega(x(h_n)^*x(h_n))\ln\frac{\omega(x(h_n)^*x(h_n))}{\omega(x(h_n)x(h_n)^*)}$$

$$\leqslant \beta\epsilon\omega(x(h_n)^*x(h_n)),$$

where the last inequality is obtained from Lemma II.1 as the support of h_n is contained in an interval of length less than ϵ. Hence by (4),

$$\omega(x(h_n)x(h_n)^*) \leqslant e^{(K+\epsilon)\beta}\omega(u_m^*x(h_n)x(h_n)^*u_m),$$

and by (3)

$$\omega(x(h_n)x(h_n)^*) \leqslant e^{(K+\epsilon)\beta}\omega(\tau(x(h_n)x(h_n)^*)).$$

As $[\tau,\alpha_t] = 0$ one has $\tau(x(f)) = (\tau x)(f)$; hence after summation over n, using (5) one gets

$$\omega(xx^*) \leqslant e^{\beta(K+\epsilon)}\omega(\tau(xx^*)). \quad \blacksquare$$

At this point it might be interesting to remark that this result of absolute continuity of states is obtained through the use of the correlation inequality. It is worthwhile to mention the work of Araki[11] and of Sakai.[12] They are interested in the problem of unicity of KMS states. Sakai is also working towards a result expressing absolute continuity of states but by explicit calculations using the Gibbs form of the state. Araki's technique is based on the notion of relative entropy and leads to quasiequivalence of states.

Finally one gets as an easy consequence the invariance of the equilibrium states under the symmetry group.

Corollary II.3: Under the conditions of Theorem II.2

$$\omega\circ\tau = \omega.$$

Proof: It is sufficient to prove the corollary for extremal KMS states. Suppose that ω is such an external state. Then, as $[\tau,\alpha_t] = 0$, $\omega\circ\tau$ is also an extremal KMS state. By Theorem II.2 and a well-known property (Ref. 8, Theorem 5.3.29) there exists $T\in\mathscr{A}''\cap\mathscr{A}'$ such that

$$\omega(\tau(x)) = \langle\Omega_\omega|Tx\Omega_\omega\rangle.$$

As ω is extremal $T = 1$ and therefore $\omega = \omega\circ\tau$. $\blacksquare$

III. ILLUSTRATION

We prove the absence of breaking of translation symmetry in one-dimensional lattice systems for long-range interactions. This result was announced in Ref. 13. The algebra of observables is the usual tensor product algebra

generated by the local algebras $\mathscr{A}_\Lambda = \bigotimes_{k\in\Lambda}\mathscr{B}(\mathfrak{H})$, where $\mathfrak{H}$ is a finite-dimensional Hilbert space.

Consider the local Hamiltonian

$$H_N = \sum_{-N\leqslant i<j\leqslant N}\sum_{rs} J_{rs}(|i-j|)\sigma_i^r\sigma_j^s + \sum_r h_r\sum_{i=-N}^{N}\sigma_i^r,$$

where $\{\sigma_i^r | r = 1,\ldots,d\}$ are the spin matrices for the lattice site i; the interaction energies $J_{rs}(k)$ satisfy

$$\sum_{k=1}^{\infty}|J_{rs}(k)| < \infty. \tag{6}$$

This condition guarantees a good thermodynamic behavior of the system.

Now we want to apply Theorem II.2. The symmetry τ is the translation over one lattice site, i.e., $\tau(\sigma_i^r) = \sigma_{i+1}^r$. Note that τ is approximately inner since it can be approximated by τ_m standing for the cyclic translation of the lattice interval $[-m, +m]$ such that

$$\tau_m(\sigma_i^r) = \sigma_{i+1}^r \quad \text{if } -m \leqslant i < m,$$
$$\tau_m(\sigma_m^r) = \sigma_{-m}^r,$$
$$\tau_m(\sigma_j^r) = \sigma_j^r \quad \text{if } |j| > m.$$

It is easy to check that there exist unitary operators u_m such that $\tau_m(x) = u_m^* x u_m$ for all elements of $\mathscr{A}$. Clearly for all $x \in \cup_\Lambda \mathscr{A}_\Lambda$ one has $\tau(x) = \tau_m(x)$ when m is large enough. Therefore formula (3) holds. Furthermore, because of condition (6) the time evolution automorphisms α_t are well defined as[8]

$$\alpha_t(x) = \lim_N e^{-itH_N} x e^{-itH_N}$$

on the C^*-algebra generated by $\cup_\Lambda \mathscr{A}_\Lambda$ and clearly $[\alpha_t, \tau] = 0$. Suppose now that for all $r,s = 1,\ldots,d$,

$$\sum_{k=1}^{\infty} k \left| J_{rs}(k) - J_{rs}(k-1) \right| < \infty; \tag{7}$$

then

$$\sup_m \| [H, u_m^*] \| = \sup_m \| u_m [H, u_m^*] \|$$
$$= \sup_m \| \lim_N (\tau_m^{-1}(H_N) - H_N) \|$$
$$\leqslant \sum_{rs} \left\{ 2 \sum_{k=1}^{\infty} k \left| J_{rs}(k) - J_{rs}(k-1) \right| + 12 \sum_{k=1}^{\infty} |J_{rs}(k)| \right\} < \infty.$$

Hence (4) is satisfied and by Theorem II.2 each KMS state ω satisfies

$$\omega(xx^*) \leqslant C\omega(\tau(xx^*)).$$

By Corollary II.3 $\omega = \omega \circ \tau$ and we proved that any equilibrium state of the system is translation invariant if the interaction energies satisfy condition (7). It is instructive to realize that in the ferromagnetic or antiferromagnetic case [i.e., the $J_{rs}(k)$ have the same sign] condition (7) follows from condition (6) if the function $k \to J_{rs}(k)$ is monotonic for large k.

Finally we remark that, although we considered here only a one-dimensional system, our method extends to higher-dimensional ones, e.g., it provides a short proof of the absence of breaking of internal symmetries in two-dimensional quantum lattice systems.[7] Furthermore, the proof of Theorem II.2 relies on an estimate for $\omega(x^* u_m [H, u_m^*] x)$ given by condition (4). Depending on the particular model under consideration more refined estimates might be obtained weakening condition (4) on the interaction and hence extending the range of applicability of the theorem.

[1] M. Fannes, P. Vanheuverzwijn, and A. Verbeure, J. Stat. Phys. **29**, 545–558 (1982).
[2] B. Simon and A. D. Sokal, J. Stat. Phys. **25**, 679 (1981).
[3] N. D. Mermin and H. Wagner, Phys. Rev. Lett. **17**, 1133 (1966).
[4] P. C. Hohenberg, Phys. Rev. **158**, 383 (1967).
[5] C. A. Bonato, J. F. Perez, and A. Klein, J. Stat. Phys. **29**, 159 (1982).
[6] M. Fannes and A. Verbeure, J. Math. Phys. **19**, 558 (1978).
[7] J. Fröhlich and C. E. Pfister, Commun. Math. Phys. **81**, 277 (1981).
[8] D. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics II* (Springer-Verlag, New York, 1981).
[9] G. Roepstorff, Commun. Math. Phys. **46**, 253 (1976).
[10] R. T. Powers and S. Sakai, Commun. Math. Phys. **39**, 273 (1975).
[11] H. Araki, Commun. Math. Phys. **44**, 1 (1975).
[12] S. Sakai, J. Functional Analysis **27**, 203 (1976).
[13] M. Fannes, P. Vanheuverzwijn, and A. Verbeure, "Quantum Energy-Entropy Balance and Breaking of Symmetries," Preprint KUL-TF-82/19.

Quantum measuring processes of continuous observables

Masanao Ozawa
Department of Information Sciences, Tokyo Institute of Technology, Oh-Okayama, Meguro-ku, Tokyo 152, Japan

(Received 3 May 1983; accepted for publication 23 June 1983)

The purpose of this paper is to provide a basis of theory of measurements of continuous observables. We generalize von Neumann's description of measuring processes of discrete quantum observables in terms of interaction between the measured system and the apparatus to continuous observables, and show how every such measuring process determines the state change caused by the measurement. We establish a one-to-one correspondence between completely positive instruments in the sense of Davies and Lewis and the state changes determined by the measuring processes. We also prove that there are no weakly repeatable completely positive instruments of nondiscrete observables in the standard formulation of quantum mechanics, so that there are no measuring processes of nondiscrete observables whose state changes satisfy the repeatability hypothesis. A proof of the Wigner–Araki–Yanase theorem on the nonexistence of repeatable measurements of observables not commuting conserved quantities is given in our framework. We also discuss the implication of these results for the recent results due to Srinivas and due to Mercer on measurements of continuous observables.

PACS numbers: 03.65.Bz, 02.50. + s

1. INTRODUCTION

In the last decade, some attempts were developed to construct a satisfactory theory of the quantum mechanical measurement of an observable with continuous spectrum.[1–9] However, we have found no satisfactory solution of the fundamental problem to determine the state changes caused by measurements of continuous observables. In spite of these difficulties in continuous spectrum, the theory for discrete spectrum has a conventionally accepted solution since the pioneering work of von Neumann.[10]

Let $A = \Sigma_i \lambda_i P_i$ be an observable with simple discrete spectrum $\lambda_1, \lambda_2, \cdots$. Then von Neumann[10] showed the following:

(1) By the repeatability hypothesis, the state change $\rho \to \rho'$ caused by the measurement of A is determined as $\rho' = \Sigma_i P_i \rho P_i$.

(2) The above state change $\rho \to \rho'$ is compatible with the Hamiltonian formalism in the description of the measuring process in terms of the time evolution of the composite system of the observed system and the measuring apparatus.

In the present paper, we shall show the following:

(1) The description of measuring processes has a satisfactory generalization to continuous observables.

(2) Every measuring process determines a state change caused by the measurement.

(3) There are no measuring processes of a nondiscrete observable whose state changes satisfy the repeatability hypothesis.

In order to clarify the present situation, we shall review some developments on the problem so far. In the early stage, Umegaki and Nakamura[11] showed that the state change $\rho \to \rho' = \Sigma_i P_i \rho P_i$ is just an example of Umegaki's noncommutative conditional expectations[12] onto the von Neumann algebra generated by A, and they conjectured that the state change caused by the measurement of a continuous observable would also be such a noncommutative conditional expectation. However, it is shown by Areveson[13] that such conditional expectations do not exist for continuous observables. In view of these results, Davies and Lewis[1] established the mathematical concept of instruments which enables us to treat statistical correlations of outcomes of successive measurements, and formulate the repeatability hypothesis for continuous observables. They conjectured the nonexistence of repeatable instruments for continuous observables and proposed the more flexible approach to measurements of continuous observables abandoning repeatability hypothesis. Recently, Srinivas[8] generalized the concept of instruments and showed the existence of such generalized instruments for continuous observables which satisfy the repeatability hypothesis. He proposed a generalized collapse postulate which determines such repeatable generalized instruments to describe the state changes caused by measurements of continuous observables. More recently, Mercer[9] considered a wider class of state transformations than conditional expectations and proposed the state change should be described by such a transformation with the locality introduced by him. It is a remarkable fact that these attempts are concerned only with the first half of von Neumann's work cited above. An operator theoretical analysis on von Neumann's second result was done by Kraus.[14] He established the complete positivity of state changes caused by the general measuring processes, but his result is concerned only with the yes–no measurements.

In this paper, we shall show that the state changes determined by measuring processes naturally correspond to completely positive instruments and vice versa. We prove Davies and Lewis's conjecture for completely positive instruments, i.e., completely positive instruments cannot be weakly repeatable unless the corresponding observable is discrete. These results show that Srinivas's generalized col-

 0022-2488/84/010079-09$02.50

lapse postulate cannot be compatible for continuous observables with the Hamiltonian description of measuring processes. We shall also show that if they can be realized by some measuring processes, Mercer's local transition maps correspond to repeatable measurements, and hence they cannot exist for continuous observables.

The nonexistence of repeatable measuring processes of continuous observables suggests that we should investigate the approximately repeatable measuring processes as models of measurements in quantum mechanics. Moreover, this direction of investigation is appropriate not only for continuous observables. Indeed, even in measurements of discrete observables, it is known that the repeatable measurement is impossible unless observed quantity commutes with conserved quantity under some conservation law (see Refs. 15 and 16, also Sec. 8). The author believes that, in future investigations on really existing approximately repeatable measurements, our framework of measuring processes will provide a nice basis. However, we shall discuss these problems elsewhere.

In Sec. 2, we give some preliminaries on semiobservables and conditional expectations. Our concept of observed quantities allows the nonorthogonal resolutions of identity, called semiobservables. In Sec. 3, we generalize von Neumann's measuring processes to continuous observables and show that every measuring process determines the state change caused by the measurement. In Sec. 4, we provide a dilation theorem and a decomposition theorem of completely positive instruments which are useful in the later sections. In Sec. 5, we shall establish the one-to-one correspondence between measuring processes and completely positive instruments. If the observed quantity is a usual one, the obtained correspondence is reduced to very simple form by the decomposition theorem, that is, measuring processes are determined by their transition $\rho\to\rho'$. In Sec. 6, we study the repeatability hypothesis and prove the nonexistence of weakly repeatable completely positive instruments for nondiscrete observables in the standard formulation of quantum mechanics. In Sec. 7, we study the local transition maps and prove the nonexistence of local transition maps corresponding to measuring processes of nondiscrete observables. In Sec. 8, we shall give a proof of the Wigner–Araki–Yanase theorem in our framework, which states the nonexistence of repeatable measuring processes of the observables which do not commute with the conserved quantity. In Sec. 9, we shall give a characterization of the measuring processes discussed in the conventional measurement theory among our general measuring processes.

2. OBSERVABLES AND CONDITIONAL EXPECTATIONS

Let $\mathscr{H}$ be a Hilbert space. Denote by $\mathscr{L}(\mathscr{H})$ the algebra of bounded operators on $\mathscr{H}$ and by $\mathscr{T}(\mathscr{H})$ the space of trace class operators on $\mathscr{H}$. A *state* ρ on $\mathscr{H}$ is a positive trace one operator on $\mathscr{H}$. Denote by $\mathscr{S}(\mathscr{H})$ the space of all states on $\mathscr{H}$. Let $(\Omega,\mathscr{B})$ be a Borel space. A *semiobservable* X on $\mathscr{H}$ with value space $(\Omega,\mathscr{B})$ is a positive operator valued measure $X:\mathscr{B}\to\mathscr{L}(\mathscr{H})$ such that $X(\Omega)=1$. An *observable* X is a semiobservable which is projection valued. Denote by $\mathscr{B}(\mathbb{R}^n)$ the Borel σ-field of $\mathbb{R}^n$. By the spectral theory, we shall identify an observable X on $\mathscr{H}$ with value space $(\mathbb{R}^n,\mathscr{B}(R^n))$ and the corresponding mutually commutable family $\{x_1,\ldots,x_n\}$ of self-adjoint operators on $\mathscr{H}$ such that

$$x_i=\int_{\mathbb{R}}\lambda X(\mathbb{R}\times\cdots\times d\hat{\lambda}\times\cdots\times\mathbb{R}). \tag{2.1}$$

An observable X with value space $(\mathbb{R},\mathscr{B}(R))$ is called *bounded* if $x=\int_{\mathbb{R}}\lambda X(d\lambda)$ is bounded. Let X be a semiobservable on $\mathscr{H}$ with value space $(\Omega,\mathscr{B})$. If the system is in the state ρ at the instant before a measurement of X, then the probability distribution $\mathrm{Prob}(X\in B;\rho)$ of the outcomes of this measurement is given by

$$\mathrm{Prob}(X\in B;\rho)=\mathrm{Tr}[\rho X(B)], \tag{2.2}$$

for any B in $\mathscr{B}$. For a semiobservable X, we shall denote by $X(\mathscr{B})$ the range of X, i.e., $X(\mathscr{B})=\{X(B);B\in\mathscr{B}\}$. A *conditional expectation* T on $\mathscr{L}(\mathscr{H})$ onto a von Neumann algebra $\mathscr{M}$ on $\mathscr{H}$ is a normal completely positive linear map T on $\mathscr{L}(\mathscr{H})$ with range $\mathscr{M}$ such that $T(axb)=aT(x)b$ for all a,b in $\mathscr{M}$, x in $\mathscr{L}(\mathscr{H})$. It is known[17] that an ultraweakly continuous linear map T on $\mathscr{L}(\mathscr{H})$ is a conditional expectation if and only if it is a projection of norm 1 onto $\mathscr{M}$.

Let $\mathscr{K}$ be another Hilbert space. Let σ be a state on $\mathscr{K}$. Then the formula

$$\mathrm{Tr}[\rho E_\sigma(x)]=\mathrm{Tr}[(\rho\otimes\sigma)x], \tag{2.3}$$

where $x\in\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$ and $\rho\in\mathscr{T}(\mathscr{H})$, defines a normal completely positive linear map $E_\sigma:\mathscr{L}(\mathscr{H}\otimes\mathscr{K})\to\mathscr{L}(\mathscr{H})$ such that $E_\sigma(a\otimes 1)=a$ for any a in $\mathscr{L}(\mathscr{H})$. Thus the formula $x\to E_\sigma(x)\otimes 1$, for x in $\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$, defines a conditional expectation on $\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$ onto $\mathscr{L}(\mathscr{H})\otimes\mathbb{C}1$. It is easily seen that the map E_σ is the adjoint of the map $\rho\to\rho\otimes\sigma$ from $\mathscr{T}(\mathscr{H})$ into $\mathscr{T}(\mathscr{H}\otimes\mathscr{K})$. The formula

$$\mathrm{Tr}[E_{\mathscr{K}}(\phi)a]=\mathrm{Tr}[\phi(a\otimes 1)], \tag{2.4}$$

where $\phi\in\mathscr{T}(\mathscr{H}\otimes\mathscr{K})$ and $a\in\mathscr{L}(\mathscr{H})$, defines a completely positive linear map $E_{\mathscr{K}}:\mathscr{T}(\mathscr{H}\otimes\mathscr{K})\to\mathscr{T}(\mathscr{H})$, which is called the *partial trace* over $\mathscr{K}$. The partial trace $E_{\mathscr{K}}$ also satisfies that for any ξ,η in $\mathscr{H}$, and any orthogonal basis $\{\psi_i\}$, we have

$$(E_{\mathscr{K}}(\rho)\xi,\eta)=\sum_i(\rho(\xi\otimes\psi_i),\eta\otimes\psi_i), \tag{2.5}$$

for any ρ in $\mathscr{T}(\mathscr{H}\otimes\mathscr{K})$. It is easily seen that the adjoint of $E_{\mathscr{K}}$ is the map $a\to a\otimes 1$ from $\mathscr{L}(\mathscr{H})$ into $\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$.

The following lemmas can be verified by easy computations.

Lemma 2.1: Let $\rho\in\mathscr{T}(\mathscr{H})$, $\sigma\in\mathscr{T}(\mathscr{H}\otimes\mathscr{K})$, and $b\in\mathscr{L}(\mathscr{K})$. If we have $\mathrm{Tr}[a\rho]=\mathrm{Tr}[(a\otimes b)\sigma]$ for any $a\in\mathscr{L}(\mathscr{H})$, then we have

$$\rho=E_{\mathscr{K}}[(1\otimes b)\sigma]. \tag{2.6}$$

Lemma 2.2: Let $T:\mathscr{T}(\mathscr{H})\to\mathscr{T}(\mathscr{H})$ be a bounded linear map, and let $U\in\mathscr{L}(\mathscr{H}\otimes\mathscr{K})$, $b\in\mathscr{L}(\mathscr{K})$, and $\sigma\in\mathscr{T}(\mathscr{K})$. Then

$$T(\rho)=E_{\mathscr{K}}[U(\rho\otimes\sigma)U^*(1\otimes b)], \tag{2.7}$$

for any ρ in $\mathscr{T}(\mathscr{H})$ if and only if

$$T^*(a)=E_\sigma[U^*(a\otimes b)U], \tag{2.8}$$

for any a in $\mathcal{L}(\mathcal{H})$.

Lemma 2.3: Let $\sigma = \Sigma_i \lambda_i |\xi_i\rangle\langle\xi_i|$ be the spectral decomposition of σ in $\mathcal{S}(\mathcal{H})$. Then

$$E_\sigma[A] = \sum_i \lambda_i E_{|\xi_i\rangle\langle\xi_i|}[A], \tag{2.9}$$

for any A in $\mathcal{L}(\mathcal{H}\otimes\mathcal{K})$, where the sum is convergent in the weak operator topology.

3. MEASURING PROCESSES

In order to determine the possible transformations of states associated with the measurement of an observable, we shall consider the description of the measuring process in terms of the interaction between the observed system and the apparatus, which is a generalization of von Neumann's description of the measuring process for an observable with discrete spectrum (Ref. 10, Chap. IV). Our mathematical formulation of the measuring process is as follows.

Definition 3.1: Let $\mathcal{H}$ be a Hilbert space and X be a semiobservable on $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. A *measuring process* M of X is a 4-tuple $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ consisting of a Hilbert space $\mathcal{K}$, an observable $\tilde{X}$ on $\mathcal{K}$ with value space $(\Omega,\mathcal{B})$, a state σ on $\mathcal{K}$, and a unitary operator U on $\mathcal{H}\otimes\mathcal{K}$ satisfying the relation

$$X(B) = E_\sigma[U^*(1\otimes\tilde{X}(B))U] \tag{3.1}$$

for any B in $\mathcal{B}$.

Now we shall explain the physical interpretation of the measuring process $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ of a semiobservable X of a Hilbert space $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. The Hilbert space $\mathcal{H}$ and $\mathcal{K}$ describe, respectively, the measured system I and the apparatus II. The semiobservable X is to be measured by this measuring process. The observable $\tilde{X}$ is to show the value of X on a scale in the apparatus which is actually measured by the observer, i.e., $\tilde{X}$ is the position of the pointer on this scale. The state σ is the initially prepared state of the apparatus. The measurement is carried out by the interaction between the observed system and the apparatus during a finite time interval from time 0 to t. The unitary operator U describes the time evolution of the composite system, i.e.,

$$U = \exp[-it(H_{\rm I}\otimes 1 + 1\otimes H_{\rm II} + H_{\rm int})], \tag{3.2}$$

where $H_{\rm I}$ and $H_{\rm II}$ are Hamiltonians of the observed system I and the apparatus II, respectively, and $H_{\rm int}$ represents the interaction. Suppose that at the instant before the interaction the measured system is in the (unknown) state ρ. Then the composite system is in the state $\rho\otimes\sigma$ at time 0 and by the interaction it is in the state $U(\rho\otimes\sigma)U^*$ at time t. Thus the probability distribution $\mathrm{Prob}(X\in B;\rho)$ of the outcomes of this measurement must coincide with the probability distribution $\mathrm{Prob}(\tilde{X}\in B;t)$ of the observable $\tilde{X}$ at time t. Since $\mathrm{Prob}(X\in B;\rho) = \mathrm{Tr}[\rho X(B)]$ and $\mathrm{Prob}(\tilde{X}\in B;t) = \mathrm{Tr}[U(\rho\otimes\sigma)U^*\tilde{X}(B)]$, we should impose the requirement

$$\mathrm{Tr}[\rho X(B)] = \mathrm{Tr}[U(\rho\otimes\sigma)U^*\tilde{X}(B)] \tag{3.3}$$

for any B in $\mathcal{B}$, ρ in $\mathcal{S}(\mathcal{H})$. It is easy to see that the requirement (3.3) is equivalent to the requirement (3.1) in Definition 3.1.

We shall now show that the measuring process $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ determines a unique state change caused by this measurement. Suppose that a measuring process $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ of X is carried out in the initial state ρ of $\mathcal{H}$. Let $B\in\mathcal{B}$. Denote by ρ^B the state, at the instant after the measurement, of the subensemble of the measured system in which the outcomes of the measurement lie in B. In order to determine the state ρ^B, suppose that the observer were to measure the simultaneously measurable observables A in I and $\tilde{X}$ in II, where A is an arbitrary bounded observable with value space $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. Then we have the joint probability distribution of their values:

$$\mathrm{Prob}(A\in d\lambda,\tilde{X}\in d\omega) = \mathrm{Tr}[U(\rho\otimes\sigma)U^*(A(d\lambda)\otimes\tilde{X}(d\omega))]. \tag{3.4}$$

Thus, if $\mathrm{Prob}(\tilde{X}\in B)\neq 0$, we have also the conditional probability distribution of A conditioned by the value of $\tilde{X}$ lying in B,

$$\begin{aligned}\mathrm{Prob}(A\in d\lambda\,|\,\tilde{X}\in B) &= \mathrm{Prob}(A\in d\lambda,\tilde{X}\in B)/\mathrm{Prob}(\tilde{X}\in B)\\ &= \mathrm{Tr}[U(\rho\otimes\sigma)U^*(A(d\lambda)\otimes\tilde{X}(B))]/\mathrm{Tr}[\rho X(B)],\end{aligned} \tag{3.5}$$

and the conditional expectation $\mathrm{Ex}(A\,|\,\tilde{X}\in B)$ of A conditioned by the value of $\tilde{X}$ lying in B,

$$\begin{aligned}\mathrm{Ex}(A\,|\,\tilde{X}\in B) &= \int_{\mathbb{R}}\lambda\,\mathrm{Prob}(A\in d\lambda\,|\,\tilde{X}\in B)\\ &= \mathrm{Tr}[U(\rho\otimes\sigma)U^*(a\otimes\tilde{X}(B))]/\mathrm{Tr}[\rho X(B)],\end{aligned} \tag{3.6}$$

where $a = \int_{\mathbb{R}}\lambda A(d\lambda)$. On the other hand, by the probabilistic interpretation of the state ρ^B, the state ρ^B must satisfy the relation

$$\mathrm{Prob}(A\in d\lambda\,|\,\tilde{X}\in B) = \mathrm{Tr}[\rho^B A(d\lambda)] \tag{3.7}$$

or, equivalently,

$$\mathrm{Ex}(A\,|\,\tilde{X}\in B) = \mathrm{Tr}[\rho^B a]. \tag{3.8}$$

By the arbitrariness of A, we can determine the state ρ^B uniquely by Eqs. (3.6) and (3.8). That is, by Lemma 2.1, we have

$$\rho^B = \{1/\mathrm{Tr}[\rho X(B)]\}E_{\mathcal{K}}[U(\rho\otimes\sigma)U^*(1\otimes\tilde{X}(B))], \tag{3.9}$$

where $E_{\mathcal{K}}:\mathcal{T}(\mathcal{H}\otimes\mathcal{K})\to\mathcal{T}(\mathcal{H})$ is the partial trace over $\mathcal{K}$. In particular, we have

$$\rho^\Omega = E_{\mathcal{K}}[U(\rho\otimes\sigma)U^*]. \tag{3.10}$$

Therefore, we have determined the state change $\rho\to\rho^B$ caused by the measuring process $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ of the semiobservable X on $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$.

Let $M = \langle\mathcal{K},\tilde{X},\sigma,U\rangle$ be a measuring process of a semiobservable X. For any a in $\mathcal{L}(\mathcal{H})$, $\mathrm{Ex}^M(a|B;\rho)$ will denote the conditional expectation of the outcome of a measurement of a at that instant after the measuring process M under the condition that the measuring process M of X has been carried out in the initial state ρ on $\mathcal{H}$ and its outcome lies in $B\in\mathcal{B}$. Then from the above discussions, we have

$$\begin{aligned}\mathrm{Ex}^M(a|B;\rho) &= \mathrm{Tr}[\rho^B a]\\ &= \{1/\mathrm{Tr}[\rho X(B)]\}\\ &\quad\times\mathrm{Tr}[U(\rho\otimes\sigma)U^*(a\otimes\tilde{X}(B))].\end{aligned} \tag{3.11}$$

Conclusion: Every measuring process $M = \langle \mathscr{K}, \tilde{X}, \sigma, U \rangle$ of a semiobservable X determines a state change $\rho \to \rho^B$ caused by the measurement, where ρ^B is the state, at the instant after the measurement, of the subensemble of the measured system in which outcomes of the measurement in the initial state ρ lies in $B \in \mathscr{B}$.

4. COMPLETELY POSITIVE INSTRUMENTS

From the investigations of von Neumann's repeated measurements, Davies and Lewis[1] introduced a mathematical notion of instruments which represents statistical correlations of outcomes of successive measurements. For the theory of instruments, called operational quantum probability theory, we refer the reader to Refs. 1 and 4. In the present section, we shall provide some general results on instruments imposed complete positivity.

Our setting for operational quantum probability theory consists of a von Neumann algebra $\mathscr{M}$ on a Hilbert space $\mathscr{H}$ and a Borel space $(\Omega, \mathscr{B})$. A *state* ρ of $\mathscr{M}$ is a normal state on $\mathscr{M}$. Denote by $\mathscr{M}_*$ the predual of $\mathscr{M}$ and by $\Sigma(\mathscr{M})$ the space of all normal states on $\mathscr{M}$. A *semiobservable* X in $\mathscr{M}$ is a semiobservable on $\mathscr{H}$ whose range is contained in $\mathscr{M}$. A *subtransition map* T on $\mathscr{M}$ is a normal completely positive linear map $T: \mathscr{M} \to \mathscr{M}$ such that $0 \leqslant T(1) \leqslant 1$. A *transition map* T is a subtransition map such that $T(1) = 1$. We define the right action of a subtransition map T on $\mathscr{M}_*$ by the duality

$$\langle \rho, Ta \rangle = \langle \rho T, a \rangle, \tag{4.1}$$

for all a in $\mathscr{M}$, ρ in $\mathscr{M}_*$. A *CP instrument* $\mathscr{I}$ on $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$ is a subtransition map valued measure on $(\Omega, \mathscr{B})$ such that (i) for each countable family $\{ B_i \}$ of pairwise disjoint sets in $\mathscr{B}$,

$$\left\langle \rho, \mathscr{I}\left(\bigcup_i B_i\right) a \right\rangle = \sum_i \langle \rho, \mathscr{I}(B_i) a \rangle, \tag{4.2}$$

for all a in $\mathscr{M}$, ρ in $\mathscr{M}_*$ and that (ii) $\mathscr{I}(\Omega)1 = 1$. The condition (i) is equivalent to countable additivity of the right action in the strong operator topology on $\mathscr{L}(\mathscr{M}_*, \mathscr{M}_*)$. In what follows we shall also use the notation $\mathscr{I}(\cdot,\cdot)$ for a CP instrument $\mathscr{I}$ in such a way $\mathscr{I}(B,a) = \mathscr{I}(B)a$ for all B in $\mathscr{B}$, a in $\mathscr{M}$. By the same argument as in Ref. 1, Theorem 1, we can prove the following.

Proposition 4.1: For every CP instrument $\mathscr{I}$ on $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$ there is a unique semiobservable X in $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$ such that $X(B) = \mathscr{I}(B,1)$ for all B in $\mathscr{B}$. Every semiobservable is determined in such a way by at least one CP instrument.

Let $\mathscr{I}$ be a CP instrument. We say that a semiobservable X is the *associate semiobservable* of $\mathscr{I}$, if $X(B) = \mathscr{I}(B,1)$ for any B in $\mathscr{B}$ and that a transition map T is the *associate map* of $\mathscr{I}$ if $T(a) = \mathscr{I}(\Omega, a)$ for any a in $\mathscr{M}$. Let X be a semiobservable. A CP instrument $\mathscr{I}$ is called *X-compatible* if X is the associate semiobservable of $\mathscr{I}$. A transition map T is called *X-compatible* if the range of T is contained in $X(\mathscr{B})'$.

The following proposition is very useful in dealing with CP instruments which is a modification of the Stinespring theorem on completely positive maps.[18]

Proposition 4.2: For any CP instrument $\mathscr{I}$ of $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$ there is a Hilbert space $\mathscr{H}_0$, a spectral measure $E: \mathscr{B} \to \mathscr{L}(\mathscr{H}_0)$, a nondegenerate normal*-representation $\pi: \mathscr{M} \to \mathscr{L}(\mathscr{H}_0)$ and a linear isometry $V: \mathscr{H} \to \mathscr{H}_0$ satisfying

$$\mathscr{I}(B,a) = V^* E(B) \pi(a) V, \tag{4.3}$$

$$E(B)\pi(a) = \pi(a) E(B), \tag{4.4}$$

for any B in $\mathscr{B}$ and a in $\mathscr{M}$.

Proof: Denote by $B(\Omega)$ the space of all bounded $\mathscr{B}$-measurable functions on Ω. Consider the algebraic tensor product $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H}$. We define a sesquilinear form $(\cdot,\cdot)$ on $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H}$ as follows:

$$(\xi, \eta) = \sum_i \int_\Omega g_j(\omega)^* f_i(\omega) (\mathscr{I}(d\omega, b_j^* a_i) \xi_i, \eta_j),$$

for $\xi = \Sigma_i f_i \otimes a_i \otimes \xi_i$, $\eta = \Sigma_j g_j \otimes b_j \otimes \eta_j$ in $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H}$. Then we can prove that $(\xi, \xi) \geqslant 0$ by just a similar way as the proof of Ref. 18, Theorem 4, and thus $\xi \to \|\xi\| = (\xi, \xi)^{1/2}$ is a seminorm. Define actions π of $\mathscr{M}$ and E of $\mathscr{B}$ on $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H}$ as follows:

$$\pi(x)\xi = \sum_i f_i \otimes x a_i \otimes \xi_i,$$

$$E(B)\xi = \sum_i \chi_B f_i \otimes a_i \otimes \xi_i,$$

for x in $\mathscr{M}$, B in $\mathscr{B}$, and $\xi = \Sigma_i f_i \otimes a_i \otimes \xi_i$. Then we have that $\|\pi(x)\xi\| \leqslant \|x\| \, \|\xi\|$ and $\|E(B)\xi\| \leqslant \|\xi\|$. Thus the both actions are well defined also on the $\|\cdot\|$-completion $\mathscr{H}_0$ of the quotient space $B(\Omega) \otimes \mathscr{M} \otimes \mathscr{H} / \mathscr{N}$, where $\mathscr{N} = \{ \xi \mid \|\xi\| = 0 \}$. Define a map $V: \mathscr{H} \to \mathscr{H}_0$ as $V\phi = (1 \otimes 1 \otimes \phi) + \mathscr{N}$, for any ϕ in $\mathscr{H}$. Then the assertions can be checked in a routine manner (Ref. 18 and Ref. 19, p. 194). QED

A CP instrument $\mathscr{I}$ is called *decomposable* if it is of the form $\mathscr{I}(B,a) = X(B)T(a)$ for all B in $\mathscr{B}$, a in $\mathscr{M}$, where X is the associate semiobservable of $\mathscr{I}$ and T is the associate map of $\mathscr{I}$.

Proposition 4.3: A CP instrument $\mathscr{I}$ is decomposable if its associate semiobservable X is projection-valued or if its associate map T is homomorphic [i.e., $T(a^*a) = T(a)^*T(a)$ for all a in $\mathscr{M}$].

Proof: First suppose that T is homomorphic. We can suppose that $\mathscr{I}$ is of the form $\mathscr{I}(B,a) = V^* E(B) \pi(a) V$ as in Proposition 4.2. Since $T(a) = V^* \pi(a) V$ and $V^* V = 1$, we have

$$(\pi(a)V - VT(a))^*(\pi(a)V - VT(a)) = T(a^*a) - T(a)^*T(a) = 0.$$

Thus $\pi(a)V = VT(a)$ for all a in $\mathscr{M}$, and hence we obtain that $\mathscr{I}(B,a) = V^* E(B) \pi(a) V = V^* E(B) VT(a) = X(B)T(a)$ for any B in $\mathscr{B}$, a in $\mathscr{M}$. The proof for the case that X is projection-valued is similar. QED

Proposition 4.4: Let X be an observable in $\mathscr{M}$ with value space $(\Omega, \mathscr{B})$. Then there is a one-to-one correspondence between X-compatible CP instruments $\mathscr{I}$ on $\mathscr{M}$ and X-compatible transition maps T on $\mathscr{M}$, which is given by $\mathscr{I}(B,a) = X(B)T(a)$ for any B in $\mathscr{B}$, a in $\mathscr{M}$.

Proof: If a CP instrument $\mathscr{I}$ is decomposable, then its associate map T is X-compatible, since $X(B)T(a) = (X(B)T(a))^* = T(a)X(B)$ for any $a \geqslant 0$ in $\mathscr{M}$, B in $\mathscr{B}$.

Conversely, if T is an X-compatible transition map then it is easy to check that the relation $\mathcal{I}(B,a) = X(B)T(a)$, where $a \in \mathcal{M}$ and $B \in \mathcal{B}$, defines an X-compatible CP instrument. Thus the assertion follows immediately from Proposition 4.3. QED

5. CLASSIFICATION OF MEASURING PROCESSES

Let $\mathcal{H}$ be a Hilbert space and X be a semiobservable on $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. We say that two measuring processes M_1 and M_2 of X are *statistically equivalent* if

$$\mathrm{Ex}^{M_1}(a|B;\rho) = \mathrm{Ex}^{M_2}(a|B;\rho), \tag{5.1}$$

for any a in $\mathcal{L}(\mathcal{H})$, B in $\mathcal{B}$, ρ in $\mathcal{S}(\mathcal{H})$. Since every two statistically equivalent measuring processes give the same state change, it is desirable to classify these equivalence classes by more tractable mathematical objects concerned only with the observed system. In this section, we shall carry out such classification.

Let $M = \langle \mathcal{K},\tilde{X},\sigma,U \rangle$ be a measuring process of X. Consider the following relation:

$$\mathcal{I}(B)a = E_\sigma[U^*(a \otimes \tilde{X}(B))U], \tag{5.2}$$

for any B in $\mathcal{B}$, a in $\mathcal{L}(\mathcal{H})$. Then it is not hard to check that Eq. (5.2) defines an X-compatible CP instrument $\mathcal{I}$ on $\mathcal{L}(\mathcal{H})$. By Lemma 2.2, Eq. (5.2) is equivalent to

$$\rho\mathcal{I}(B) = E_{\mathcal{K}}[U(\rho \otimes \sigma)U^*(1 \otimes \tilde{X}(B))], \tag{5.3}$$

for all B in $\mathcal{B}$, ρ in $\mathcal{T}(\mathcal{H})$. By Eqs. (3.1) and (3.9), we have

$$X(B) = \mathcal{I}(B,1), \tag{5.4}$$

$$\rho^B = (1/\mathrm{Tr}[\rho\mathcal{I}(B)])\rho\mathcal{I}(B), \quad \text{whenever } \mathrm{Tr}[\rho X(B)] \neq 0, \tag{5.5}$$

for all ρ in $\mathcal{S}(\mathcal{H})$, B in $\mathcal{B}$. Thus the CP instrument $\mathcal{I}$ defined by Eq. (5.2) retains the all statistical data of the measuring process M, that is, the probability distribution of outcomes of the measurement and the state change caused by the measurement. The following theorem shows that every CP instrument on $\mathcal{L}(\mathcal{H})$ arises in this way.

Theorem 5.1: Let X be a semiobservable on $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. Then there is a one-to-one correspondence between statistical equivalence classes of measuring processes M of X and X-compatible CP instruments $\mathcal{I}$ on $\mathcal{L}(\mathcal{H})$, which is given by the relation

$$\mathrm{Tr}[\rho\mathcal{I}(B)]\mathrm{Ex}^M(a|B;\rho) = \mathrm{Tr}[\rho\mathcal{I}(B)a], \tag{5.6}$$

for all B in $\mathcal{B}$, ρ in $\mathcal{S}(\mathcal{H})$, a in $\mathcal{L}(\mathcal{H})$.

Proof: Let $M = \langle \mathcal{K},\tilde{X},\sigma,U \rangle$ be a measuring process of X. Then it is easy to see that the CP instrument $\mathcal{I}$ defined by Eq. (5.2) is a unique CP instrument which satisfies Eq. (5.6). It follows that the statistically equivalent measuring processes determine the same CP instrument by Eq. (5.2). Now it suffices to construct a measuring process of X which determines by Eq. (5.2) a given X-compatible CP instrument. Let $\mathcal{I}$ be an X-compatible CP instrument on $\mathcal{L}(\mathcal{H})$ with value space $(\Omega,\mathcal{B})$. Let $\mathcal{H}_0$, E, π, and V be such as obtained in Proposition 4.2 for the CP instrument $\mathcal{I}$. Since every nondegenerate normal $*$-representation of $\mathcal{L}(\mathcal{H})$ is unitarily equivalent to the multiple of the identity representation (Ref. 4, Lemma 9.2.2), there is a Hilbert space $\mathcal{H}_1$ such that $\mathcal{H}_0 = \mathcal{H} \otimes \mathcal{H}_1$ and that $\pi(a) = a \otimes 1$ for any a in $\mathcal{L}(\mathcal{H})$. Then by Eq. (4.3) and by the commutation theorem of von Neumann algebras, for any B in $\mathcal{B}$ there is a projection $E_1(B)$ in $\mathcal{L}(\mathcal{H}_1)$ such that $E(B) = 1 \otimes E_1(B)$. Obviously, the correspondence $E_1: B \to E_1(B)$ is a projection-valued measure from $\mathcal{B}$ to $\mathcal{L}(\mathcal{H}_1)$. By Eq. (4.3), we have

$$\mathcal{I}(B,a) = V^*(a \otimes E_1(B))V,$$

for any B in $\mathcal{B}$, a in $\mathcal{L}(\mathcal{H})$. Let η_0 be a unit vector in $\mathcal{H}_0$ and η_1 be a unit vector in $\mathcal{H}_1$. Define an isometry V_0 on $\mathcal{H} \otimes [\eta_1] \otimes [\eta_0]$ into $\mathcal{H} \otimes \mathcal{H}_1 \otimes \mathcal{H}_0$ by the relation

$$V_0(\xi \otimes \eta_1 \otimes \eta_0) = V\xi \otimes \eta_0,$$

for any ξ in $\mathcal{H}$. Then, since $\dim(\mathcal{H}_0) = \dim(\mathcal{H} \otimes \mathcal{H}_1)$, by the usual computations of cardinal numbers, it is easy to show that

$$\dim(\mathcal{H} \otimes \mathcal{H}_1 \otimes \mathcal{H}_0 - \mathcal{H} \otimes [\eta_1] \otimes [\eta_0]) = \dim(\mathcal{H} \otimes \mathcal{H}_1 \otimes \mathcal{H}_0 - V_0(\mathcal{H} \otimes [\eta_1] \otimes [\eta_0])).$$

It follows that there is a unitary operator U on $\mathcal{H} \otimes \mathcal{H}_1 \otimes \mathcal{H}_0$ which is an extension of V_0. Now let $\mathcal{K}$, σ, and $\tilde{X}$ be such that

$$\mathcal{K} = \mathcal{H}_1 \otimes \mathcal{H}_0, \quad \sigma = |\eta_1 \otimes \eta_0\rangle\langle\eta_1 \otimes \eta_0|,$$
$$\text{and} \quad \tilde{X}(B) = E_1(B) \otimes 1 \quad \text{on } \mathcal{H}_1 \otimes \mathcal{H}_0,$$

for any B in $\mathcal{B}$. Then we shall claim that $\langle \mathcal{K},\tilde{X},\sigma,U \rangle$ is a measuring process which determines the CP instrument $\mathcal{I}$ by Eqs. (5.2). For any a in $\mathcal{L}(\mathcal{H})$, ξ in $\mathcal{H}$, and B in $\mathcal{B}$, we have that

$$\begin{aligned}
(\mathcal{I}(B,a)\xi,\xi) &= (V^*(a \otimes E_1(B))V\xi,\xi) \\
&= ((a \otimes E_1(B))V\xi, V\xi) \\
&= ((a \otimes E_1(B))V\xi \otimes \eta_0, V\xi \otimes \eta_0) \\
&= ((a \otimes E_1(B) \otimes 1)U(\xi \otimes \eta_1 \otimes \eta_0), U(\xi \otimes \eta_1 \otimes \eta_0)) \\
&= (U^*(a \otimes \tilde{X}(B))U(\xi \otimes \eta_1 \otimes \eta_0), \xi \otimes \eta_1 \otimes \eta_0) \\
&= \mathrm{Tr}[U^*(a \otimes \tilde{X}(B))U(|\xi\rangle\langle\xi| \otimes \sigma)] \\
&= \mathrm{Tr}[|\xi\rangle\langle\xi|E_\sigma[U^*(a \otimes \tilde{X}(B))U]] \\
&= (E_\sigma[U^*(a \otimes \tilde{X}(B))U]\xi,\xi).
\end{aligned}$$

It follows that

$$\mathcal{I}(B,a) = E_\sigma[U^*(a \otimes \tilde{X}(B))U],$$

for any a in $\mathcal{L}(\mathcal{H})$ and B in $\mathcal{B}$. Therefore, $\langle \mathcal{K},\tilde{X},\sigma,U \rangle$ is a measuring process of X which determines $\mathcal{I}$ by Eq. (5.2). QED

We say that a measuring process M is a *realization* of a CP instrument $\mathcal{I}$ if M and $\mathcal{I}$ satisfies Eq. (5.6). The above theorem asserts that every CP instrument has its realization. In the conventional theory of quantum mechanics, it is always assumed that the Hilbert space is separable and the value space is a standard Borel space, i.e., a Borel space which is Borel isomorphic to a separable complete metric space.[20] Thus it is desirable that the realization is also with a separable Hilbert space in such circumstances. We say that realization $M = \langle \mathcal{K},\tilde{X},\sigma,U \rangle$ of a CP instrument $\mathcal{I}$ is *separable* if the Hilbert space $\mathcal{K}$ is separable.

Corollary 5.2: Let $\mathcal{I}$ be a CP instrument on $\mathcal{L}(\mathcal{H})$ with value space $(\Omega,\mathcal{B})$. If $\mathcal{H}$ is separable and $(\Omega,\mathcal{B})$ is a standard Borel space, then there is a separable realization of $\mathcal{I}$.

Proof (the notations are the same as in the proof of Theorem 5.1): It is easy to see that we can assume that $\mathscr{H}_0$ in Proposition 4.2 is spanned by $\{E(B)\pi(a)V\xi;\ B\in\mathscr{B},\ a\in\mathscr{L}(\mathscr{H})$, and $\xi\in\mathscr{H}\}$. Since $\mathscr{H}$ is separable, there is a countable family $\{a_n\}$ of a_n in $\mathscr{L}(\mathscr{H})$ which is dense in $\mathscr{L}(\mathscr{H})$ in the strong operator topology. Let $\{B_n\}$ be a countable generator of $\mathscr{B}$ and $\{\xi_n\}$ be a countable dense subset of $\mathscr{H}$. Then it is easy to see that the countable family $\{E(B_i)\times\pi(a_j)V\xi_k; i,j,k = 1,2,\cdots\}$ spans $\mathscr{H}_0$, so that $\mathscr{H}_0$ is separable. Since $\mathscr{H}\otimes\mathscr{K}=\mathscr{H}_0\otimes\mathscr{H}_0$, $\mathscr{K}$ is separable. QED

We say that a measuring process $M=\langle\mathscr{K},\tilde{X},\sigma,U\rangle$ is *pure* if σ is pure state, i.e., there is a unit vector ξ in $\mathscr{K}$ such that $\sigma=|\xi\rangle\langle\xi|$. In the conventional argument of quantum measurement, the assumption that the prepared state of the apparatus is pure has been justified in some contexts. The following is one of such justification from a most general point of view.

Corollary 5.3: Every measuring process is statistically equivalent to a pure measuring process.

Proof: The assertion is immediate from the construction of the measuring process in the proof of Theorem 5.1. QED

Let $M=\langle\mathscr{K},\tilde{X},|\eta\rangle\langle\eta|,U\rangle$ be a pure measuring process. Define an isometry $V:\mathscr{H}\to\mathscr{H}\otimes\mathscr{K}$ by $V\xi=U(\xi\otimes\eta)$ for all ξ in $\mathscr{H}$. Let $\mathscr{I}$ be the corresponding CP instrument. Then it is easy to see that

$$\mathscr{I}(B,a)=E_\sigma[U^*(a\otimes\tilde{X}(B))U]=V^*(a\otimes\tilde{X}(B))V,$$

for all a in $\mathscr{L}(\mathscr{H})$, B in $\mathscr{B}$.

The following result justifies our postulate, which is tacit in Eq. (2.2), that *semiobservables can be measured.*

Corollary 5.4: For any semiobservable X, there is a measuring process of X.

Proof: By proposition 4.1, for any semiobservable X, there is an X-compatible CP instrument $\mathscr{I}$. Then any realization of $\mathscr{I}$ obtained by Theorem 5.1 is a measuring process of X. QED

Consider the case that X is an observable. In this case the classification of measuring processes is surprisingly simpler, that is, the measuring processes of X are determined by their total state changes $\rho\to\rho^\Omega$.

Theorem 5.5: Let X be an observable on $\mathscr{H}$ with value space $(\Omega,\mathscr{B})$. Then there is a one-to-one correspondence between statistical equivalence classes of measuring processes M of X and X-compatible transition maps T on $\mathscr{L}(\mathscr{H})$, which is given by the relation

$$\mathrm{Tr}[\rho X(B)]\mathrm{Ex}^M(a|B;\rho)=\mathrm{Tr}[\rho X(B)T(a)],\qquad(5.7)$$

for any a in $\mathscr{L}(\mathscr{H})$, ρ in $\Sigma(\mathscr{H})$, B in $\mathscr{B}$.

Proof: The assertion follows immediately from Proposition 4.4 and Theorem 5.1. QED

6. REPEATABILITY

Consider von Neumann's repeatability hypothesis (Ref. 10, pp. 214, 335):

(M) If the physical quantity is measured twice in succession in a system, then we get the same value each time.

Let $M=\langle\mathscr{K},\tilde{X},\sigma,U\rangle$ be a measuring process of a semiobservable X. If X is discrete, then it is easy to see that (M) is equivalent to

(M′) $\mathrm{Ex}^M(X(\{\lambda\})|\{\mu\};\rho)=\delta_{\lambda,\mu}$

for all ρ in $\Sigma(\mathscr{H})$ and all λ,μ in Ω, whenever $\mathrm{Tr}[\rho X(\{\mu\})]\neq 0$. We say that a measuring process M of X is *weakly repeatable* if

(R) $\mathrm{Ex}^M(X(C)|B;\rho)=\mathrm{Tr}[\rho X(B\cap C)]/\mathrm{Tr}[\rho X(B)]$,

for any ρ in $\Sigma(\mathscr{H})$, B,C in $\mathscr{B}$, whenever $\mathrm{Tr}[\rho X(B)]\neq 0$. Then it is easy to see that if X is discrete the condition (M′) and (R) are equivalent. The condition (R) appeared first in Ref. 1 for instruments. We say that a CP instrument $\mathscr{I}$ is *weakly repeatable* if $\mathscr{I}(B)X(C)=X(B\cap C)$ for all B,C in $\mathscr{B}$, where X is the associate semiobservable of $\mathscr{I}$. It is easily seen that a measuring process M is weakly repeatable if and only if the corresponding CP instrument $\mathscr{I}$ is weakly repeatable. In Ref. 1, p. 247, it is conjectured that the existence of repeatable instruments for continuous observables is doubtful even in the case of standard quantum theory. In the present section, we shall prove this conjecture, that is, we shall prove that there is at least one X-compatible weakly repeatable CP instrument on $\mathscr{L}(\mathscr{H})$ if and only if X is discrete.

Let $\mathscr{M}$ be a von Neumann algebra on $\mathscr{H}$ and $(\Omega,\mathscr{B})$ be a Borel space. Let $\mathscr{I}$ be a weakly repeatable CP instrument on $\mathscr{M}$ with value space $(\Omega,\mathscr{B})$, X its associate semiobservable, and T its associate map. We can assume that $\mathscr{I}$ is of the form $\mathscr{I}(B,a)=V^*E(B)\pi(a)V$ for any B in $\mathscr{B}$, a in $\mathscr{M}$, as in Proposition 4.2.

Lemma 6.1: For any B,C in $\mathscr{B}$, a in $\mathscr{M}$, we have

(1) $T(X(B)^2)=X(B)$,

(2) $\mathscr{I}(B\cap C,a)=\mathscr{I}(C,aX(B))=\mathscr{I}(C,X(B)a)$,

(3) $\mathscr{I}(B,a)=T(aX(B))=T(X(B)a)$.

Proof: Since $\mathscr{I}(B,X(B))=X(B)$ by the weak repeatability of $\mathscr{I}$, a routine computation leads that

$$(\pi(X(B))V-E(B)V)^*(\pi(X(B))V-E(B)V)=T(X(B)^2)-X(B),\qquad(6.1)$$

for any B in $\mathscr{B}$. Thus we have $T(X(B)^2)\geqslant X(B)$. On the other hand, we have $X(B)^2\leqslant X(B)$, since $0\leqslant X(B)\leqslant 1$. By weak repeatability, $T(X(B))=X(B)$, so that $X(B)=T(X(B))\geqslant T(X(B)^2)$. Thus we have the relation (1). It follows that the left-hand side of Eq. (6.1) is 0, so that we have $\pi(X(B))V=E(B)V$ and $V^*\pi(X(B))=V^*E(B)$. Thus for any B,C in $\mathscr{B}$, a in $\mathscr{M}$, we have $\mathscr{I}(B\cap C,a)=V^*E(B\cap C)\pi(a)V=V^*E(B)E(C)\pi(a)V=V^*\pi(X(B))E(C)\pi(a)V=V^*E(C)\pi(X(B)a)V=\mathscr{I}(C,X(B)a)$. By the analogous way we can show that $\mathscr{I}(B\cap C,a)=\mathscr{I}(C,aX(B))$. Thus we obtain the relation (2). The relation (3) is obtained by putting $C=\Omega$ in (2). QED

Let p be the least projection in $X(\mathscr{B})''$ such that $T(p)=1$.

Lemma 6.2: For any x in $\mathscr{M}$, $T(x)=T(xp)=T(px)=T(pxp)$.

Proof: For any ξ,η in $\mathscr{H}$, we have

$$\begin{aligned}|(T(x-px)\xi,\eta)| &= |(V^*\pi(1-p)\pi(x)V\xi,\eta)|\\ &= |(\pi(x)V\xi,\pi(1-p)V\eta)|\leqslant\|\pi(x)V\xi\|\ \|\pi(1-p)V\eta\|\\ &= \|\pi(x)V\xi\|\ \|(V^*\pi(1-p)V\eta,\eta)\|^{1/2}\\ &= \|\pi(x)V\xi\|(T(1-p)\eta,\eta)^{1/2}=0.\end{aligned}$$

Thus we have $T(x)=T(px)$. The rest of the assertions are immediate. QED

Lemma 6.3: For every x in $X(\mathscr{B})''$ with $x \geqslant 0$, if $T(x) = 0$, then $pxp = 0$.

Proof: Let e be the range projection of x. Since e is a limit of polynomials of x not containing the constant term in the strong operator topology, we have $T(e) = 0$. Thus $1 - e \geqslant p$ so that $ep = pe = 0$. It follows that $pxp = 0$. QED

Define a positive operator valued measure P: $\mathscr{B} \to X(\mathscr{B})''$ by the relation $P(B) = pX(B)p$ for all B in $\mathscr{B}$.

Lemma 6.4: P is a projection valued measure such that $P(B) = pX(B) = X(B)p$ for any B in $\mathscr{B}$.

Proof: By Lemma 6.2, we have $T(P(B)) = T(pX(B)p) = T(X(B))$. By Lemmas 6.1 and 6.2, we have $T(P(B)^2) = T(pX(B)pX(B)p) = T(X(B)pX(B)) = \mathscr{I}(B,pX(B)) = \mathscr{I}(B \cap B,p) = \mathscr{I}(B,p) = T(pX(B)) = T(X(B))$. It follows that $T(P(B) - P(B)^2) = 0$. Since $P(B) - P(B)^2$ belongs to $X(\mathscr{B})''$, we have $P(B)^2 = P(B)$ by Lemma 6.3. Thus P is a projection-valued measure. We have $T((P(B) - X(B)p)^*(P(B) - X(B)p)) = 0$, by the routine computations. Thus, by Lemma 6.3, $P(B) = X(B)p$, since $P(B) - X(B)p$ is in $X(\mathscr{B})''$. By the positivity, we have $P(B) = pX(B)$. QED

Theorem 6.5: For any weakly repeatable CP instrument $\mathscr{I}$ on $\mathscr{M}$ with value space $(\Omega,\mathscr{B})$, there is a projection-valued measure P: $\mathscr{B} \to X(\mathscr{B})''$ such that

$$\mathscr{I}(B,a) = T(aP(B)) = T(P(B)a)$$

and that

$$P(B) = P(\Omega)X(B) = X(B)P(\Omega),$$

for any B in $\mathscr{B}$, a in $\mathscr{L}(\mathscr{H})$.

Proof: The assertion follows immediately from Lemmas 6.1 and 6.4. QED

We suppose for the rest of this section that the value space $(\Omega,\mathscr{B})$ is a standard Borel space and that the Hilbert space $\mathscr{H}$ is separable. We say that a positive operator valued measure P is *discrete* if there is a countable set $\Omega_0 \subseteq \Omega$ such that $P(\Omega \setminus \Omega_0) = 0$ and that a CP instrument is *discrete* if the associate semiobservable is discrete.

Theorem 6.6: Let $(\Omega,\mathscr{B})$ be a standard Borel space, and let $\mathscr{H}$ be a separable Hilbert space. Then every weakly repeatable CP instrument $\mathscr{I}$ on $\mathscr{L}(\mathscr{H})$ with value space $(\Omega,\mathscr{B})$ is discrete.

Proof: Let P be a projection-valued measure obtained in Theorem 6.5. By the relation $X(B) = \mathscr{I}(B,1) = T(P(B))$ for every B in $\mathscr{B}$, we have only to show that P is discrete. By Ref. 4, Lemma 4.4.1, there is a countable set B_0 such that $B \to P(B \cap B_0)$ is a discrete projection-valued measure with values in $\mathscr{L}(P(B_0)\mathscr{H})$ and $B \to P(B \setminus B_0)$ is a continuous projection-valued measure with values in $\mathscr{L}(P(\Omega \setminus B_0)\mathscr{H})$. Let Q be such that $Q = P(\Omega \setminus B_0)$. Then it suffices to prove that $Q = 0$. Let T be the associate map of $\mathscr{I}$ and T_0 be such that $T_0(a) = QT(a)Q$ for all a in $\mathscr{L}(Q\mathscr{H})$. Then $T_0(Q) = QT(Q)Q = QT(X(\Omega \setminus B_0))Q = QX(\Omega \setminus B_0)Q = Q$, and hence T_0 is a transition map on $\mathscr{L}(Q\mathscr{H})$. Thus there is a trace-preserving linear map S: $\mathscr{T}(Q\mathscr{H}) \to \mathscr{T}(Q\mathscr{H})$ such that $S^* = T_0$. For any a in $\mathscr{L}(Q\mathscr{H})$, B in $\mathscr{B}$, ρ in $\mathscr{T}(Q\mathscr{H})$, we have

$$\begin{aligned}\mathrm{Tr}[aP(B \setminus B_0)S(\rho)] &= \mathrm{Tr}[T_0(aP(B \setminus B_0))\rho] \\ &= \mathrm{Tr}[QT(aP(B \setminus B_0))Q\rho] = \mathrm{Tr}[QT(P(B \setminus B_0)a)Q\rho] \\ &= \mathrm{Tr}[T_0(P(B \setminus B_0)a)\rho] = \mathrm{Tr}[P(B \setminus B_0)aS(\rho)].\end{aligned}$$

It follows that $P(B \setminus B_0)S(\rho) = S(\rho)P(B \setminus B_0)$ for any B in $\mathscr{B}$, ρ in $\mathscr{T}(Q\mathscr{H})$. Since $B \to P(B \setminus B_0)$ is a continuous projection-valued measure, we can conclude that $S = 0$ (see, Ref. 4, Theorem 4.3.3), and hence $Q = T_0(Q) = 0$. QED

7. LOCALITY

Let $\mathscr{H}$ be a Hilbert space and $\mathscr{M}$ a von Neumann algebra on $\mathscr{H}$. Let X be an observable in $\mathscr{M}$ with value space $(\Omega,\mathscr{B})$. A transition map T on $\mathscr{M}$ is called *X-local* if $T(X(B)) = X(B)$ for any B in $\mathscr{B}$. It is easy to see that T is X-local if and only if $Tx = x$ for any x in $X(\mathscr{B})''$.

Let $\{x_1,\ldots,x_n\}$ be a mutually commutable family of self-adjoint operators on $\mathscr{H}$ corresponding to a family of simultaneously measurable observables of a quantum system. Suppose that X is the joint spectral measure of $\{x_1,\ldots,x_n\}$ on $\mathscr{H}$ with value space $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$. Recently, Mercer[9] proposed that the total state change $\rho \to \rho'$ caused by a simultaneous measurement of $x_1,\ldots,x_n$ should be described by an X-local transition map T on $\mathscr{L}(\mathscr{H})$ in such a way $\rho' = \rho T$ (see Ref. 9, p. 244). However, we should notice that the X-locality is not sufficient for describing state transformations caused by measurements. In fact, the identity transformation on $\mathscr{L}(\mathscr{H})$ is obviously an X-local transition map for any observable X, in spite of the fact that we cannot measure any nontrivial quantum observable unchanging every state of the system. Thus we have to impose some further requirements for eliminating such physically irrelevant X-local transition maps in order to describe a state change caused by the measurement of X. A moderate one of such requirements seems the existence of a measuring process for observables $x_1,\ldots,x_n$, whose state change is the given X-local transition map. The following result is an easy consequence of the results obtained in the previous sections, but shows that such requirement cannot be fulfilled unless all observables $x_1,\ldots,x_n$ are discrete.

Proposition 7.1: Let $\mathscr{M}$ be a von Neumann algebra on a Hilbert space $\mathscr{H}$ and X an observable in $\mathscr{M}$ with value space $(\Omega,\mathscr{B})$. There is a one-to-one correspondence between X-compatible X-local transition maps T on $\mathscr{M}$ and X-compatible weakly repeatable CP-instruments $\mathscr{I}$ on $\mathscr{M}$, which is given by

$$\mathscr{I}(B,a) = X(B)T(a), \tag{7.1}$$

for any B in $\mathscr{B}$, a in $\mathscr{M}$.

Proof: It is known in the proof of Ref. 1, Theorem 7, that a decomposable CP instrument $\mathscr{I}$ is weakly repeatable if and only if

$$T(X(B)) = X(B) \quad \text{and} \quad X(B \cap C) = X(B)X(C),$$

for any B,C in $\mathscr{B}$. Since every X-compatible CP instrument is decomposable, the assertion follows immediately from Proposition 4.4 QED

Theorem 7.2: Let X be an observable on a separable Hilbert space $\mathscr{H}$ whose value space is a standard Borel space and T be an X-local transition map on $\mathscr{L}(\mathscr{H})$. If there is a measuring process $M = \langle \mathscr{K},\tilde{X},\sigma,U \rangle$ of X such that $\rho^{\Omega} = \rho T$ for any ρ in $\mathscr{S}(\mathscr{H})$ [see Eq. (3.10)], then X is discrete.

Proof: It is obvious that T is the associate map of the CP instrument $\mathscr{I}$ determined by the measuring process M.

Thus, by Proposition 7.1, the CP instrument $\mathcal{I}$ is weakly repeatable and hence by Theorem 6.6 the corresponding observable X is discrete. QED

8. THE WIGNER–ARAKI–YANASE THEOREM

It was pointed out by Wigner[15] that the presence of a conservation law puts a limitation of the measurement of an operator which does not commute with the observed quantity. A proof of the above assertion was given by Araki and Yanase[16] in the conventional framework of measurement theory. In this section, we shall give another proof in our framework and under somewhat general assumptions. Our assertion is the following.

Theorem 8.1: Let X be an observable on a Hilbert space $\mathcal{H}$ with value space $(\Omega,\mathcal{B})$. Let $M=\langle\mathcal{K},\tilde{X},\sigma,U\rangle$ be a weakly repeatable measuring process of X. Suppose that there is L_1 in $\mathcal{L}(\mathcal{H})$ and L_2 in $\mathcal{L}(\mathcal{K})$ such that $[U,L]=0$, where $L=L_1\otimes 1+1\otimes L_2$. Then $L_1\in X(\mathcal{B})'$.

For the proof we use the following.

Lemma 8.2. Let $M=\langle\mathcal{K},\tilde{X},\sigma,U\rangle$ be a measuring process of an observable X on $\mathcal{H}$, and $\sigma=\Sigma_i\lambda_i|\eta_i\rangle\langle\eta_i|$ be the spectral decomposition of σ. Then for any i, $M_i=\langle\mathcal{K},\tilde{X},|\eta_i\rangle\langle\eta_i|,U\rangle$ is a pure measuring process of X such that

$$E_\sigma[U^*AU]=\sum_i\lambda_i E_{|\eta_i\rangle\langle\eta_i|}[U^*AU], \tag{8.1}$$

for any A in $\mathcal{L}(\mathcal{H}\otimes\mathcal{K})$. If M is weakly repeatable, then M_i is also weakly repeatable for every i.

Proof: The formula (8.1) is obtained from Lemma 2.3. Let $B\in\mathcal{B}$. Then

$$\begin{aligned}X(B)&=E_\sigma[U^*(1\otimes\tilde{X}(B))U]\\&=\sum_i\lambda_i E_{|\eta_i\rangle\langle\eta_i|}[U^*(1\otimes\tilde{X}(B))U].\end{aligned}$$

Since any projection is an extreme point of the positive part of the unit sphere of $\mathcal{L}(\mathcal{H})$, we have that

$$X(B)=E_{|\eta_i\rangle\langle\eta_i|}[U^*(1\otimes\tilde{X}(B))U],$$

for any i. Thus M_i is a measuring process of X. Since M_i is weakly repeatable if $X(B)=E_{|\eta_i\rangle\langle\eta_i|}[U^*(X(B)\otimes 1)U]$ for any B in $\mathcal{B}$ by Proposition 7.1, the assertion for the weak repeatability follows from the same reasoning. QED

Proof of Theorem 8.1: By Theorem 5.5, there is an X compatible transition map T such that $E_\sigma[U^*(a\otimes\tilde{X}(B))U]=X(B)T(a)$ for any B in $\mathcal{B}$, a in $\mathcal{L}(\mathcal{H})$. Then we have

$$\begin{aligned}&T(L_1)+E_\sigma[U^*(1\otimes L_2)U]\\&\quad=E_\sigma[U^*(L_1\otimes 1+1\otimes L_2)U]\\&\quad=E_\sigma[L_1\otimes 1+1\otimes L_2]\\&\quad=L_1+[\mathrm{Tr}(\sigma L_2)]1.\end{aligned}$$

Since T is X-compatible, $T(L_1)\in X(\mathcal{B})'$. Thus we have only to show that $E_\sigma[U^*(1\otimes L_2)U]\in X(\mathcal{B})'$. By Lemma 8.2, we can assume without any loss of generality that there is a unit vector η in $\mathcal{K}$ such that $\sigma=|\eta\rangle\langle\eta|$, so that there is an isometry $V:\mathcal{H}\to\mathcal{H}\otimes\mathcal{K}$ such that $E_\sigma[U^*AU]=V^*AV$ for all A in $\mathcal{L}(\mathcal{H}\otimes\mathcal{K})$, where $V\xi=U(\xi\otimes\eta)$ for any ξ in $\mathcal{H}$. Let $B\in\mathcal{B}$. Since the CP instrument $\mathcal{I}$ such that $\mathcal{I}(B,a)=V^*(a\otimes\tilde{X}(B))V$ is weakly repeatable, we have $V^*(X(B)\otimes 1)V=\mathcal{I}(\Omega,X(B))=X(B)$. Thus by the simple computations we have

$$((X(B)\otimes 1)V-VX(B))^*((X(B)\otimes 1)V-VX(B))=0,$$

and hence $(X(B)\otimes 1)V=VX(B)$ and $V^*(X(B)\otimes 1)=X(B)V^*$. It follows that

$$V^*(1\otimes L_2)VX(B)=V^*(X(B)\otimes L_2)V=X(B)V^*(1\otimes L_2)V.$$

Thus we conclude that $E_\sigma[U^*(1\otimes L_2)U]\in X(\mathcal{B})'$. QED

9. CONVENTIONAL MEASURING PROCESSES

In the conventional theory of quantum measurement, the only class of measuring processes studied at all seriously is as follows. Let $\mathcal{H}$ be a separable Hilbert space and X be a discrete observable on $\mathcal{H}$ with value space $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. Let $\{\xi_{ij}\}$ be a complete orthonormal set of eigenvectors of X where the eigenvalue of ξ_{ij} is λ_i. Let $\mathcal{K}$ be another separable Hilbert space with complete orthonormal vectors $\{\eta_i\}$. Let η be a unit vector of $\mathcal{K}$ and U be a unitary operator on $\mathcal{H}\otimes\mathcal{K}$ satisfying

$$U(\xi_{ij}\otimes\eta)=\xi_{ij}\otimes\eta_i \tag{9.1}$$

for any i,j. Then $M=\langle\mathcal{K},\tilde{X},|\eta\rangle\langle\eta|,U\rangle$ is a measuring process of X, where $\tilde{X}=\Sigma_i\lambda_i|\eta_i\rangle\langle\eta_i|$. In the sequel, we call this measuring process a *conventional* measuring process of X. The total state change corresponding to M is of the form

$$\rho\to\rho'=\sum_i P_i\rho P_i, \tag{9.2}$$

where $P_i=X(\{\lambda_i\})$, i.e., $P_i=\Sigma_j|\xi_{ij}\rangle\langle\xi_{ij}|$. In fact, for $\rho=\Sigma_{ijkl}\mu_{ijkl}|\xi_{ij}\rangle\langle\xi_{kl}|$ in $\mathcal{L}(\mathcal{H})$, we have

$$\begin{aligned}\rho'&=E_{\mathcal{K}}[U(\rho\otimes|\eta\rangle\langle\eta|)U^*]\\&=\sum_{ijkl}\mu_{ijkl}E_{\mathcal{K}}[|\xi_{ij}\otimes\eta_i\rangle\langle\xi_{kl}\otimes\eta_k|]\\&=\sum_{ijkl}\mu_{ijkl}(\eta_i,\eta_k)|\xi_{ij}\rangle\langle\xi_{kl}|\\&=\sum_i P_i\rho P_i\end{aligned}$$

[see Eq. (3.10)]. Conversely, every state change given by Eq. (9.2) is realized as the above measuring process M as shown by von Neumann (see Ref. 10, p. 442). By Eq. (9.2) the corresponding CP instrument $\mathcal{I}$ is of the form

$$\mathcal{I}(B,a)=\sum_{\lambda_i\in B}P_iaP_i, \tag{9.3}$$

for any B in $\mathcal{B}(\mathbb{R})$, a in $\mathcal{L}(\mathcal{H})$, and the corresponding transition map T is a conditional expectation onto $X(\mathcal{B}(\mathbb{R}))'$.

In the present section, we shall give a characterization of the above conventional measuring processes up to statistical equivalence. A similar problem is considered in Refs. 1 and 21 in different methods.

Definition 9.1: Let X be a semiobservable on $\mathcal{H}$ with value space $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. A measuring process M of X is called *standard* if it satisfies the following three conditions.

(WR) (Weak repeatability) M is weakly repeatable.

(MD) (Minimal disturbance condition) The set $\{\rho\in\mathcal{S}(\mathcal{H});\rho^{\mathrm{R}}\neq\rho\}$ is minimal in the set inclusion among all measuring process of X.

(ND) (Nondegeneracy condition) For any B in $\mathscr{B}(\mathbb{R})$ with $X(B)\neq 0$, there is some ρ in $\mathfrak{S}(\mathscr{H})$ such that $\mathrm{Tr}[\rho^{R}X(B)]\neq 0$.

Let M be a measuring process of X. Denote by $F(M)$ the set of all nondisturbed states, i.e., $F(M)=\{\rho\in\sigma(\mathscr{H});\rho^{R}=\rho\}$. Obviously, M satisfies (MD) if and only if for any measuring process M' of X, $F(M)\subseteq F(M')$ implies $F(M')\subseteq F(M)$.

Proposition 9.2: Let M be a conventional measuring process of a discrete observable X. Then M is standard.

Proof: It is well known that M is weakly repeatable. The condition (ND) is easy to check. Thus we shall prove that M satisfies the condition (MD). Let M' be a measuring process of X such that $F(M)\subseteq F(M')$. Denote by T and S the transition maps corresponding to M and M', respectively. Let $\rho\in\mathscr{T}(\mathscr{H})$ be such that $\rho S=\rho$. Then it suffices to show that $\rho T=\rho$. Since T is a conditional expectation onto $X(\mathscr{B}(\mathbf{R}))'$ and by the X-compatibility of S the range of S is contained in $X(\mathscr{B}(\mathbb{R}))'$, we have $TS=S$. Since $T^{2}=T$, we have $(\rho T)T=\rho T$ so that $\rho T\in F(M)$. Thus by the assumption that $F(M)\subseteq F(M')$, $\rho T\in F(M')$. It follows that $\rho T=\rho TS=\rho S=\rho$. This concludes the proof. QED

Theorem 9.3: Let $\mathscr{H}$ be a separable Hilbert space and X be a semiobservable on $\mathscr{H}$ with value space $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. Let M be a standard measuring process of X. Then X is a discrete observable, and M is statistically equivalent to a conventional measuring process of X.

Proof: Let $\mathscr{I}$ be the CP instrument corresponding to a standard measuring process M of X. Since $\mathscr{I}$ is weakly repeatable, by Theorem 6.6, X is discrete and, by Theorem 6.5, there is an orthogonal family $\{P_{\lambda};\lambda\in R\}$ of projections in $X(\mathscr{B}(R))''$ such that

$$\mathscr{I}(B,a)=T\Big(\sum_{\lambda\in B}P_{\lambda}aP_{\lambda}\Big),\tag{9.4}$$

for all B in $\mathscr{B}$, a in $\mathscr{L}(\mathscr{H})$. Let Q be a projection in $X(\mathscr{B}(\mathbb{R}))''$ such that $Q=1-\Sigma_{\lambda\in R}P_{\lambda}$. Then we have $T(Q)=0$. It follows from the condition (ND) that $Q=0$ so that $\Sigma_{\lambda\in R}P_{\lambda}=1$. Thus by Lemma 6.4 we have $X(B)=\Sigma_{\lambda\in B}P_{\lambda}$ for any B in $\mathscr{B}(\mathbb{R})$. It follows that X is an observable. Let M' be a conventional measuring process of X and $\mathscr{I}'$ be the corresponding CP instrument. Then

$$\mathscr{I}'(B,a)=\sum_{\lambda\in B}P_{\lambda}aP_{\lambda},\tag{9.5}$$

for any B in $\mathscr{B}$, a in $\mathscr{L}(\mathscr{H})$. Denote by T and S the corresponding transition maps of M and M', respectively. Since T is X-compatible, we have $T(a)=\Sigma_{\lambda\in R}P_{\lambda}T(a)=\Sigma_{\lambda\in R}P_{\lambda}T(a)P_{\lambda}=ST(a)$, for any a in $\mathscr{L}(\mathscr{H})$. On the other hand, by Eq. (9.4) we have $T=TS$. It follows that $T=ST=TS$. For any ρ in $\mathscr{T}(\mathscr{H})$, if $\rho T=\rho$, then $\rho S=\rho TS=\rho T=\rho$ and hence $F(M)\subseteq F(M')$. Thus by the condition (MD), $F(M')=F(M)$. Let $\rho\in\mathscr{T}(\mathscr{H})$. Then since $S^{2}=S,\rho S\in F(M')$, so that $\rho ST=\rho S$. It follows that $S=ST$. Thus we have $T=S$. Therefore, by Theorem 5.5, M is statistically equivalent to a conventional measuring process M' of X. QED

ACKNOWLEDGMENTS

The author wishes to thank Professor H. Umegaki for his useful comments and encouragement. He is also indebted to Professor M. M. Yanase and Professor H. Araki for the reading of the manuscript and enlightening comments, and he is grateful to Professor A. S. Holevo and Professor N. N. Čencov for the stimulating discussions.

[1]E. B. Davies and J. T. Lewis, "An operational approach to quantum probability," Commun. Math. Phys. **17**, 239–260 (1970).
[2]E. B. Davies, "Quantum stochastic processes," Commun. Math. Phys. **15**, 277–304 (1969).
[3]E. B. Davies, "On the repeated measurement of continuous observables in quantum mechanics," J. Funct. Anal. **6**, 318–346 (1970).
[4]E. B. Davies, *Quantum Theory of Open Systems* (Academic, London, 1976).
[5]A. S. Holevo, "Statistical decision theory for quantum systems," J. Multivar. Anal. **3**, 337–394 (1973).
[6]H. Cycon and K. -E. Hellwig, "Conditional expectations in generalized probability theory," J. Math. Phys. **18**, 1154–1161 (1977).
[7]M. D. Srinivas, "Foundations of quantum probability theory," J. Math. Phys. **16**, 1672–1685 (1975).
[8]M. D. Srinivas, "Collapse postulate for observables with continuous spectra," Commun. Math. Phys. **71**, 131–158 (1980).
[9]R. Mercer, "General quantum measurements: Local transition maps," Commun. Math. Phys. **84**, 239–250 (1982).
[10]J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton, U. P., Princeton, NJ, 1955).
[11]M. Nakamura and H. Umegaki, "On von Neumann's theory of measurements in quantum statistics," Math. Jpn. **7**, 151–157 (1962).
[12]H. Umegaki, "Conditional expectation in an operator algebra, I–II," Tohoku Math. J. **6**, 177–181 (1954); **8**, 86–100 (1956).
[13]W. B. Areveson, "Analyticity in operator algebras," Am. J. Math. **89**, 578–642 (1967).
[14]K. Kraus, "General state changes in quantum theory," Ann. Phys. **64**, 311–335 (1971).
[15]E. P. Wigner, "Die Messung Quantenmechanischer Operatoren," Z. Phys. **133**, 101–108 (1952).
[16]H. Araki and M. M. Yanase, "Measurements of quantum mechanical operators," Phys. Rev. **120**, 622–626 (1960).
[17]J. Tomiyama, "On the projection of norm one in W^{*}-algebra," Proc. Jpn. Acad. **33**, 608–612 (1957).
[18]W. F. Stinespring, "Positive functions on C^{*}-algebras," Proc. Am. Math. Soc. **6**, 211–216 (1955).
[19]M. Takesaki, *Theory of Operator Algebras I* (Springer-Verlag, New York, 1979).
[20]G. W. Mackey, "Borel structure in groups and their duals," Trans. Am. Math. Soc. **85**, 134–165 (1957).
[21]G. Lüders, "Über die Zustandsanderung durch den Messprozess," Ann. Physik **8**(6), 322–328 (1951).

The Darboux transformation and solvable double-well potential models for Schrödinger equations[a)]

W. M. Zheng[b)]
Center for Studies in Statistical Mechanics, University of Texas at Austin, Austin, Texas 78712

(Received 8 September 1982; accepted for publication 10 December 1982)

The Darboux transformation, a method used to transform a Schrödinger-type equation to a Schrödinger equation with a new potential, is discussed. An exactly solvable double-well potential model for the one-dimensional Schrödinger equation is obtained.

PACS numbers: 03.65.Ge, 02.30.Em, 31.90. + s

I. INTRODUCTION

Much effort has been made to look for exactly solvable models for the one-dimensional Schrödinger equations. Double-well potential problems occur in the quantum theory of molecules. Because the Fokker–Planck equation is closely related to the Schrödinger equation,[1] the solution of the above problem can be directly applied to the problem of diffusion in a bistable potential field. Recently, great attention has been put on constructing exactly solvable bistable models.

In general, there are four ways to devise potentials. The first is to use piecewise potentials[1,2] such as the double square well and the double oscillator, this being the most common method. Its main difficulty, however, is that to obtain eigenvalues from the matching conditions, one needs to solve transcendental equations, for which analytic solutions of eigenvalues are not available unless in some limiting cases expansion formulas can be applied to find approximate solutions for the low-lying eigenvalues. The second[3] is to construct potentials from the wave functions which are solutions to two or more Schrödinger equations with simple potentials at the same eigenvalue. Since the potentials are made to fit given wave functions, a set of different eigenfunctions cannot be obtained in this way. The third[4] is to solve the Schrödinger equation directly for specially chosen potentials; for example, the potential with three parameters, β, ξ, and a positive integer n:

$$V(x) = (\hbar^2\beta^2/2m)\left[\tfrac{1}{8}\xi^2 \cosh 4\beta x - (n+1)\xi \cosh 2\beta x - \tfrac{1}{8}\xi^2\right]. \tag{1}$$

For this potential, the low-lying eigenfunctions can be found analytically in a form of finite-term summation of simple functions. The fourth method is to transform the Schrödinger equations to be solved to known solvable differential equations. There are many examples of this given in textbooks[5]; so far it appears that no example dealing with a double-well potential has been solved in this way.

In this paper two systematic methods, the Darboux transformation[6] and a new one, will be presented for transforming known solvable Schrödinger-type equations to Schrödinger equations with new different potentials. As an example, a double-well potential model will be obtained from the Weber equation,[7] and other interesting applications of the transformations will be given.

[a)] Supported in part by the Robert A. Welch Foundation.

[b)] On leave of absence from the Institute of Theoretical Physics, Academia Sinica, Beijing, China.

II. THE DARBOUX TRANSFORMATION[8]

The Darboux theorem: If the general solution $\varphi = \varphi(x)$ of the equation

$$\frac{d^2\varphi}{dx^2} + [\epsilon - U(x)]\varphi = 0 \tag{2}$$

is known for all values of ϵ and for a particular value of $\epsilon = \epsilon_0$, the particular solution is $\varphi = \varphi_0(x)$. Then the general solution of the equation

$$\frac{d^2\psi}{dx^2} + [E - V(x)]\psi = 0 \tag{3a}$$

with

$$V(x) = \varphi_0(x)\frac{d^2}{dx^2}\left(\frac{1}{\varphi_0(x)}\right), \tag{3b}$$

$$E = \epsilon - \epsilon_0 \tag{3c}$$

for $E \neq 0$ is

$$\psi(x) = \varphi_0(x)(\varphi(x)/\varphi_0(x))' \tag{4a}$$

$$= \varphi'(x) - \frac{\varphi_0'(x)}{\varphi_0(x)}\varphi(x). \tag{4b}$$

The Darboux transformation (4a) was previously used to transform the Schrödinger equations with the potentials given by Eq. (1).[4] It should be emphasized that the Darboux transformation is very general in the sense that the original equation (2) need not be a physical Schrödinger equation.

As an example, we consider the Weber equation[7]

$$\frac{d^2y}{dx^2} - \left(\frac{x^2}{4} + a\right)y = 0. \tag{5}$$

For any given parameter a, this equation has the particular solution

$$y_1(a,x) = e^{-x^2/4}\,{}_1F_1(a/2 + \tfrac{1}{4};\tfrac{1}{2};x^2/2), \tag{6}$$

where ${}_1F_1(\alpha;\beta;x)$ is a Kummer function. Here we have

$$\epsilon_0 = 0, \quad U(x) = x^2/4 + a. \tag{7}$$

From the asymptotic behavior of the Kummer function, we know that the positive definite function $y_1(a,x)$ does not satisfy the natural boundary conditions, i.e., it does not vanish at infinity, so it is not a "physical" solution. Equation (5) is only a Schrödinger-type equation. According to Eq. (3b), the transformed potential is

$$V_a(x) = y_1(a,x)\frac{d^2}{dx^2}\left(\frac{1}{y_1(a,x)}\right)$$

 0022-2488/84/010088-03$02.50

$$= 2\left[\frac{y_1'(a,x)}{y_1(a,x)}\right]^2 - \left(\frac{x^2}{4} + a\right). \tag{8}$$

This potential has been considered in the discussion of the Fokker–Planck equation for diffusion of a Brownian particle with a particular initial δ-function distribution peaked at $x = 0$.[8]

The curvature of $V_a(x)$ at $x = 0$ is

$$V_a''(x)|_{x=0} = 4a - \tfrac{1}{2} \times \begin{cases} >0, & a > 1/2\sqrt{2} \quad \text{or} \quad a < -1/2\sqrt{2}, \\ =0, & a = \pm 1/2\sqrt{2}, \\ <0, & -1/2\sqrt{2} < a < 1/2\sqrt{2}. \end{cases} \tag{9}$$

The shapes of the symmetric functions $V_a(x)$ are shown for different values of a in Fig. 1. One can see that for $|a| = 0.5$, $V_a(x)$ is a single well; for $a = -0.25$, $V_a(x)$ has a double-well structure; for $a = 0.25$, the shape of the curve is relatively complex.

The generated Schrödinger equation for the transformed potential $V_a(x)$ is

$$\frac{d^2\psi}{dx^2} + \left\{E - 2\left[\frac{y_1'(a,x)}{y_1(a,x)}\right]^2 + \left(\frac{x^2}{4} + a\right)\right\}\psi = 0. \tag{10}$$

It is easy to verify that function $[y_1(a,x)]^{-1}$ satisfies Eq. (10) for $E = 0$. The function $[y_1(a,x)]^{-1}$ has no node (as long as a is not less than -0.5) and is a square-integrable function (see Sec. IV) satisfying the natural boundary conditions, so it is the ground state of Eq. (10). The higher eigenvalues and eigenfunctions can be obtained from Eqs. (3c), (4a), and (5):

$$E_n = n + a + \tfrac{1}{2}, \quad n = 0,1,2,\cdots, \tag{11a}$$

$$\psi_n(x) = y_1(a,x)\frac{d}{dx}\left(\frac{D_n(x)}{y_1(a,x)}\right), \tag{11b}$$

where the Weber function $D_n(x)$ can be expressed in terms of the Hermite polynomial $H_n(x)$[7]

$$D_n(x) = 2^{-n/2}e^{-x^2/4}H_n(x/\sqrt{2}). \tag{12}$$

We have thus found *all* the eigenvalues and eigenfunctions of Eq. (10). Furthermore, the normalization factor for $\psi_n(x)$ can be obtained analytically (see Sec. IV).

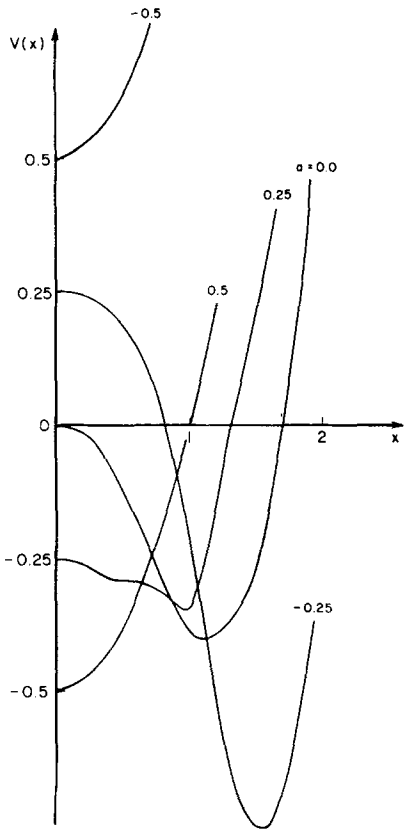

FIG. 1. Shape of $V(x)$.

III. A NEW TRANSFORMATION

The solutions to Eq. (3a) can also be given in another form:

$$\psi(x) = \frac{1}{\varphi_0(x)}\int^x \varphi(x)\varphi_0(x)\,dx. \tag{13}$$

By substituting $1/\varphi_0$ into Eq. (3a), one can directly verify that it is the solution to Eq. (3a) for eigenvalue $E = 0$. If we reinterpret Eq. (3a) as the original untransformed equation, then from the Darboux theorem we have the transformed potential

$$\tilde{V}(x) = \frac{1}{\varphi_0}\frac{d^2\varphi_0}{dx^2} = U(x) - \epsilon_0 \tag{14}$$

and the transformed equation

$$\frac{d^2\tilde{\psi}}{dx^2} + [(E + \epsilon_0) - U(x)]\tilde{\psi} = 0,$$

which is the same as Eq. (2) if one notices $E = \epsilon - \epsilon_0$. Thus, from Eq. (4a), we obtain

$$\tilde{\psi} = \varphi = \frac{1}{\varphi_0}(\psi\varphi_0)' \equiv \mathscr{W}\psi \tag{15a}$$

or

$$\psi = \mathscr{W}^{-1}\varphi = \frac{1}{\varphi_0}\int^x \varphi(x)\varphi_0(x)\,dx. \tag{15b}$$

To guarantee

$$\mathscr{W}^{-1}\mathscr{W} = \mathscr{W}\mathscr{W}^{-1} = \mathscr{I},$$

where $\mathscr{I}$ is the identity operator, we should choose the lower limit x_0 for integration in Eq. (15b) such that $\varphi(x_0) = 0$. However, the undetermined constant in the indefinite integral (13) is of no importance because the eigenvalue corresponding to $1/\varphi_0$ equals zero.

IV. DISCUSSION

(1) From the two forms of $\psi(x)$, Eqs. (4a) and (13), we have the general relation

$$\int_0^x \varphi_n(x)\varphi_0(x)\,dx = c\left[\varphi_0(x)\left(\frac{\varphi_n(x)}{\varphi_0(x)}\right)' - d\right]\varphi_0(x), \tag{16}$$

where

$$d=\varphi_0(0)\left(\frac{\varphi_n(x)}{\varphi_0(x)}\right)'_{x=0}$$

and c is a constant independent of x. By differentiating both sides of Eq. (16), we reobtain Eq. (2); thus the constant c is determined as $c=-1/E_n$.

(2) Calculation of the normalization factor for ψ_n defined by Eq. (4a) is as follows:

$$\begin{aligned}I_n &\equiv\int_{-\infty}^{\infty}\psi_n^2(x)\,dx\\ &=-2E_n\int_0^{\infty}\left[\varphi_0\left(\frac{\varphi_n}{\varphi_0}\right)'\right]\\ &\quad\times\left[\frac{1}{\varphi_0}\left(\int_0^x\varphi_n\varphi_0dx'+d\right)\right]dx\\ &=-2E_n\left\{\left[\left[\frac{\varphi_n}{\varphi_0}\left(\int_0^x\varphi_n\varphi_0dx'+d\right)\right]\right]_0^{\infty}-\int_0^{\infty}\varphi_n^2dx\right\}\\ &=2E_n\left[\frac{\varphi_n(0)}{\varphi_0(0)}d+\int_0^{\infty}\varphi_n^2dx\right].\end{aligned}\tag{17}$$

Thus for the example given in Sec. II we have

$$\begin{aligned}d&=y_1(a,0)\left(\frac{D_n(x)}{y_1(a,x)}\right)'_{x=0}\\ &=D'_n(x)|_{x=0}.\end{aligned}$$

From

$$D_n(0)\cdot D'_n(0)=0,$$

$$y_1(a,0)=1,\quad y'_1(a,0)=0,$$

and

$$\int_0^{\infty}D_n^2(x)dx=\tfrac{1}{2}n!(2\pi)^{1/2},$$

we obtain finally

$$I_n=(n+a+\tfrac{1}{2})n!(2\pi)^{1/2}.\tag{18}$$

(3) Calculation of the normalization factor for the ground state is as follows: For the ground state $E=0$, from Eq. (11), we have

$$n=-a-\tfrac{1}{2},$$

$$\frac{1}{y_1(a,x)}=ky_1(a,x)\frac{d}{dx}\left(\frac{D_{-a-1/2}(x)-D_{-a-1/2}(-x)}{2y_1(a,x)}\right),\tag{19}$$

where k is a constant. Because

$$y_1(a,0)=1,\quad D'_{\nu}(0)=-2^{\nu/2+1/2}\frac{\sqrt{\pi}}{\Gamma(-2/2)},\tag{20}$$

we have, from Eq. (19),

$$k=-2^{a/2-1/4}\Gamma(a/2+\tfrac{1}{4})/\sqrt{\pi}.\tag{21}$$

Therefore

$$\begin{aligned}\int_{-\infty}^{\infty}\frac{dx}{y_1^2(a,x)}&=2k\int_0^{\infty}d\left(\frac{D_{-a-1/2}(x)-D_{-a-1/2}(-x)}{2y_1(a,x)}\right)\\ &=2k\frac{D_{-a-1/2}(x)-D_{-a-1/2}(-x)}{2y_1(a,x)}\Bigg|_0^{\infty}\\ &=-k\lim_{x\to\infty}\left(\frac{D_{-a-1/2}(-x)}{y_1(a,x)}\right)\\ &=\frac{\sqrt{2}\Gamma(a/2+\frac{1}{4})}{\Gamma(a/2+\frac{3}{4})}.\end{aligned}\tag{22}$$

This result was derived previously in a quite different way.[9] To our knowledge, this integral is not found in tables.

(4) The exact solutions obtained by means of the transformations can be used to test approximate methods of solutions. For example, applying the WKB approximation to the energy levels below the top of the barrier in a symmetric double well,[10] one can find that at low transmission the energy levels appear in close pairs. The spectrum of our model is one in which all the energy levels higher than the lowest two are equally spaced. Thus the model is an example where the WKB approximation fails.

(5) Choosing a linear combination of $\alpha y_1(a,x)+\beta y_2(a,x)$ instead of $y_1(a,x)$ for $\varphi_0(x)$, one can construct an asymmetric potential similarly. The discussion will be made elsewhere.

(6) The methods can be applied to solve the Fokker–Planck equation[11] and other problems. In addition, the exactly solvable double-well potential model has some pedagogic value.

ACKNOWLEDGMENTS

The author would like to express deep gratitude to Max O. Hongler for introducing Ref. 4 and providing preprints of Refs. 8 and 9, as well as for the fruitful discussions which stimulated the present work. Thanks must be expressed to Professor I. Prigogine, Professor L. Reichl, and Professor W. Schieve for their hospitality at the Center for Studies in Statistical Mechanics of the University of Texas at Austin.

[1] N. G. van Kampen, J. Stat. Phys. **17**, 71 (1977).
[2] E. Merzbacher, *Quantum Mechanics* (Wiley, New York, 1970); E. W. Gettys and H. W. Gruber, Am. J. Phys. **43**, 625 (1975); M. Prakash, J. Phys. A **9**, 1847 (1976); J. Thomchick, J. P. McKelvey, and C. F. Elliott, Phys. Lett. A **66**, 86 (1978).
[3] R. G. Winter, Am. J. Phys. **45**, 569 (1977).
[4] M. Razavy, Am. J. Phys. **48**, 285 (1980).
[5] For instance, S. Flügge and H. Marschall, *Rechenmethoden der Quantentheorie* (Springer, Berlin, 1952).
[6] J. G. Darboux, C. R. Acad. Sci. Paris **94**, 1456 (1882); E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956), p. 132.
[7] M. Abramovitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
[8] M. O. Hongler, preprint.
[9] M. O. Hongler and N. D. Quach, preprint.
[10] D. ter Haar, *Problems in Quantum Mechanics* (Academic, New York, 1974), p. 144.
[11] M. O. Hongler and W. M. Zheng, J. Stat. Phys. **29**, 317 (1982).

Transmutation as a minimizing procedure

Robert Carroll
Mathematics Department, University of Illinois, Urbana, Illinois 61801

Stanley Dolzycki
Mathematics Department, Eastern Illinois University, Charleston, Illinois 61920

(Received 22 July 1983; accepted for publication 2 September 1983)

It is shown formally how transmutation kernels can be characterized via a minimizing procedure. The technique then can be extended to more general operators and transmutations.

PACS numbers: 03.65.Nk

1. INTRODUCTION

In Ref. 1 it is shown how Gel'fand–Levitan (GL) equations can be obtained by minimizing a certain quadratic functional $Q(t,K)$. The motiviation to consider $Q(t,K)$ came from a problem in optics[2] involving a feedback mechanism and statistical averaging but no motivation could be provided within scattering theory to consider $Q(t,K)$. Thus the process producing GL equations appeared to simply involve a mathematical trick which was naturally considered to be "unsatisfactory" in Ref. 1 and the "meaning" of such procedures seemed to be worth pursuing further. In the present article we will provide an interpretation of such minimizing processes in the context of transmutation theory which leads us eventually to minimize a quadratic functional essentially the same as $Q(t,K)$. This involves a characterization of transmutation kernels themselves in terms of a minimization procedure, and we sketch the development for a classical situation (a more extensive and general treatment for operators and transmutations as in Ref. 3 is clearly indicated and will appear later). Let us remark that there is a discrete version (which does not directly generalize) of a related minimization in the context of orthogonal polynomials, but without a connection to $Q(t,K)$ nor any explicit link to transmutation.[4] Although our characterization of transmutation kernels via a minimization is of interest in itself, and moreover provides "motivation" for some constructions in Ref. 1, there seem to be some deeper relations still beneath the surface. In particular, one knows that various connections between spectral measures, transmutation, autocorrelation functions, stochastic analysis, least squares optimization, etc., are all involved here.[4–8] Thus, hopefully this article will provide a contribution toward unifying some of this material as well.

2. BASIC CONSTRUCTIONS

In classical (half-line) inverse scattering theory in quantum mechanics,[9,10] one connects eigenfunctions of the Schrödinger operator $Q = D^2 - q(x)$ (q real here) with eigenfunctions of D^2 via certain (triangular) transmutation kernels of the form $\beta(y,x) = \delta(x-y) + K(y,x)$, where $K(y,x) = 0$ for $x > y$ (such K will be called causal here). Thus let $\varphi^Q_\lambda(x)$ [resp. $\theta^Q_\lambda(x)$] be solutions of

$$Qu = -\lambda^2 u \tag{2.1}$$

satisfying $\varphi^Q_\lambda(0) = 1$ and $D_x\varphi^Q_\lambda(0) = 0$ [resp. $\theta^Q_\lambda(0) = 0$ and $D_x\theta^Q_\lambda(0) = 1$]. We will write $s(\lambda,x)$ for φ^Q_λ or θ^Q_λ and think of connecting $s(\lambda,x)$ to $a(\lambda,x) = \cos\lambda x$ or $a(\lambda,x) = \sin\lambda x/\lambda$ by a formula

$$s(\lambda,y) = (1+K)a = a(\lambda,y) + \int_0^y K(y,x)a(\lambda,x)dx, \tag{2.2}$$

which we know to be valid for the GL kernel $K = K_0$. We can assume K_0 exists here and our procedure is designed to characterize it via minimization. For simplicity now let us think of $s = \theta^Q_\lambda$ and $a = \sin\lambda x/\lambda$, and remark that a systematic theory of transmutation operators $B: P \to Q$ can be developed for much more general differential operators P and Q (Ref. 3); the techniques of this article will be correspondingly extended at another time. Now one knows that associated to Q and the eigenfunctions $\theta^Q_\lambda = s$ is a spectral measure $d\omega = d\omega_Q$ which we assume here for convenience to be of the form $d\omega = \omega\, d\lambda$ (no bound states). Thus one can suppose, e.g.,

$$\int_0^\infty \theta^Q_\lambda(x)\theta^Q_\lambda(y)d\omega(\lambda) = \delta(x-y) \tag{2.3}$$

(acting on suitable functions) and we write $d\omega = d\sigma + 2\lambda^2 d\lambda/\pi$ with $\int_0^\infty a(\lambda,x)a(\lambda,y)d\sigma = \Omega(x,y)$. Thus

$$\begin{aligned}\mathfrak{A}(x,y) &= \int_0^\infty a(\lambda,x)a(\lambda,y)d\omega \\ &= \delta(x-y) + \Omega(x,y) = (1+\Omega)(x,y),\end{aligned} \tag{2.4}$$

where $a = \sin\lambda x/\lambda$ [we write 1 for the identity operator with kernel $\delta(x-y)$]. Now consider the expression (T arbitrary and fixed)

$$\Xi(T,K) = \int_0^T \int_0^\infty \{[(1+K)a(\lambda,\cdot)](y) - s(\lambda,y)\}^2 d\omega(\lambda)dy. \tag{2.5}$$

Note that when K is the GL kernel K_0 [which makes (2.2) correct], then formally $\Xi(T,K) = 0$. We can think here of Q, s, a, and $d\omega$ as given and the (causal) kernel $K(y,x)$ in (2.5) as unknown. It will be shown formally that:

Theorem 2.1: The kernel K is obtained by minimizing $\Xi(T,K)$ over a suitable class of admissible causal kernels satisfies the GL equation and represents the transmutation kernel K_0 connecting s and a via (2.2).

3. FORMAL ARGUMENTS

We proceed formally and refer to standard sources[3,9,10] for information about natural properties of $K(y,x)$, etc. Thus, from (2.5), for causal K,

 0022-2488/84/010091-03$02.50

$$\Xi(T,K)=\int_0^T\int_0^\infty\Big\{[a(\lambda,y)-s(\lambda,y)]^2+2a(\lambda,y)\int_0^y K(y,x)a(\lambda,x)dx-2s(\lambda,y)\int_0^y K(y,x)a(\lambda,x)dx+\int_0^y K(y,x)a(\lambda,x)dx\times\int_0^y K(y,\xi)a(\lambda,\xi)d\xi\Big\}d\omega(\lambda)dy. \tag{3.1}$$

Now one integrates in λ, using (2.4), and the convention $\int_0^T\Omega(y,y)dy=\operatorname{Tr}\Omega$, for example, to obtain (note that the trace Tr depends on T)

$$\Xi(T,K)=\hat{\Xi}(T)+2\operatorname{Tr}K+2\int_0^T\int_0^y K(y,x)\Omega(x,y)dx\,dy-2\int_0^T\int_0^y K(y,x)\tilde{\beta}(y,x)dx\,dy+\int_0^T\int_0^y\int_0^y K(y,x)K(y,\xi)\{\delta(x-\xi)+\Omega(x,\xi)\}d\xi\,dx\,dy, \tag{3.2}$$

where we have written $\hat{\Xi}(T)=\int_0^T\{a(\lambda,y)-s(\lambda,y)\}^2d\omega\,dy$ which we know makes sense [in fact $\hat{\Xi}(T)=\int_0^T\int_0^\infty(K_0a)^2d\omega\,dy=\int_0^T\int_0^y\int_0^y K_0(y,x)K_0(y,\xi)\{\delta(x-\xi)+\Omega(x,\xi)\}d\xi\,dx\,dy=\operatorname{Tr}\{K_0(1+\Omega)K_0^*\}$—see calculations below]. Here the term $\tilde{\beta}(y,x)=\langle s(\lambda,y),a(\lambda,x)\rangle_\omega$ is a standard object in general transmutation theory[3] which appears in extended GL equations [e.g., $\langle\beta(y,t),\mathfrak{A}(t,x)\rangle=\tilde{\beta}(y,x)$] and in particular $\tilde{\beta}(y,x)=0$ for $x<y$ (i.e., it is anticausal) with a $\delta(x-y)$ term arising along the diagonal.[11] Thus the $\tilde{\beta}$ term contributes $-2\int_0^T K(y,y)dy=-2\operatorname{Tr}K$ to (3.2). We can write now

$$K\Omega g(y)=\int_0^y K(y,x)\int_0^\infty\Omega(x,s)g(s)ds\,dx=\int_0^\infty g(s)\Big\{\int_0^y K(y,x)\Omega(x,s)dx\Big\}ds \tag{3.3}$$

(for suitable g) so that $\operatorname{Tr}K\Omega=\int_0^T\{\int_0^y K(y,x)\Omega(x,y)dx\}dy$. Similarly $\ker K^*=K(\cdot,x)$ on $[x,\infty)$ since $\int_0^\infty g(y)\int_0^y K(y,x)h(x)dx\,dy=\int_0^\infty h(x)\int_x^\infty g(y)K(y,x)dy\,dx$, and hence

$$KK^*g(y)=\int_0^y K(y,x)\int_x^\infty K(\xi,x)g(\xi)d\xi\,dx=\int_0^\infty g(\xi)\int_0^{\min(y,\xi)}K(y,x)K(\xi,x)dx\,d\xi. \tag{3.4}$$

Consequently $\operatorname{Tr}KK^*=\int_0^T\{\int_0^y K(y,x)K(y,x)dx\}dy$. Finally we have

$$K\Omega K^*g(y)=\int_0^y K(y,x)\int_0^\infty\Omega(x,s)\times\int_s^\infty K(\xi,s)g(\xi)d\xi\,ds\,dx=\int_0^y K(y,x)\int_0^\infty g(\xi)\int_0^\xi\Omega(x,s)K(\xi,s)ds\,d\xi\,dx=\int_0^\infty g(\xi)\Big\{\int_0^y K(y,x)\int_0^\xi\Omega(x,s)K(\xi,s)ds\,dx\Big\}d\xi. \tag{3.5}$$

It follows that $\operatorname{Tr}K\Omega K^*=\int_0^T\{\int_0^y K(y,x)\int_0^y\Omega(x,s)K(y,s)ds\,dx\}dy$. Now go back to (3.2) and insert the information just derived from Eqs. (3.3)–(3.5) plus the $\tilde{\beta}$ contribution, to obtain

Lemma 3.1: The expression $\Xi(T,K)$ defined in (2.5) can be written

$$\Xi(T,K)=\hat{\Xi}(T)+\operatorname{Tr}\{K(1+\Omega)K^*+K\Omega+\Omega K^*\}. \tag{3.6}$$

Proof: One obtains from (3.2), $\Xi(T,K)=\hat{\Xi}(T)+\operatorname{Tr}\{2K\Omega+KK^*+K\Omega K^*\}$. But $K(1+\Omega)K^*=KK^*+K\Omega K^*$ with $\operatorname{Tr}K\Omega=\operatorname{Tr}\Omega K^*$ (note $\Omega^*=\Omega$). Q.E.D.

Written in the form (3.6), $\Xi(T,K)$ is essentially in the same form as the expression $Q(t,K)$ (or D) in Refs. 1 and 2. We now formally examine a variational argument to minimize $\Xi=\Xi(T,K)$. Thus [note $\Xi\geqslant 0$ from (2.5)] we know there is a minimizing $K=K_0$ in some additive class $\mathfrak{K}$ of admissible (causal) kernels. Then consider $K=K_0+\epsilon L$ in $\Xi(T,K)=\hat{\Xi}(T)+\Xi_K(T)$ [$\hat{\Xi}(T)$ is independent of K] for $L\in\mathfrak{K}$ and ϵ a real number. Formally we set $D_\epsilon\Xi_k(T)|_{\epsilon=0}=0$. This leads to $\operatorname{Tr}\{L(1+\Omega)K_0^*\}+\operatorname{Tr}\{K_0(1+\Omega)L^*\}+\operatorname{Tr}L\Omega+\operatorname{Tr}\Omega L^*=2\operatorname{Tr}\{[K_0(1+\Omega)+\Omega]L^*\}=0$ for $L\in\mathfrak{K}$. If we write now $\Lambda=K_0(1+\Omega)+\Omega$ with kernel $\Lambda(y,x)$, then evidently $\ker\Lambda L^*=\int_0^{\min(s,y)}\Lambda(y,x)L(s,x)dx$ [cf. (3.4)] and $\operatorname{Tr}\Lambda L^*=\int_0^T\{\int_0^y\Lambda(y,x)L(y,x)dx\}dy$. The statement that $\operatorname{Tr}\Lambda L^*=0$ for all $L\in\mathfrak{K}$ will be true if $\Lambda(y,x)=0$ for $x<y$, and heuristically we conclude here the converse.

Theorem 3.2: The (unique) minimizing kernel K_0 satisfies the GL equation $K_0(y,x)+\Omega(y,x)+\int_0^y K_0(y,\xi)\Omega(\xi,x)d\xi=0$ for $x<y$.

One knows that the GL equation has a unique solution and this is the transmutation kernel of (2.2).[3] Hence Theorem 2.1 is verified formally.

Remark 3.3: Let us note also the following calculation which will specify (again) the minimum Ξ_0 of $\Xi(T,K)$ achieved at the GL kernel K_0. Thus given the GL equation in Theorem 3.2 we can say $K_0+\Omega+K_0\Omega=B^*$, where B is a causal operator. It follows easily that $1+B^*=(1+K_0)(1+\Omega)$, and thus

$$(1+B^*)(1+K_0^*)=(1+K_0)(1+\Omega)(1+K_0^*) \tag{3.7}$$

which is formally self-adjoint. But the left side of (3.7) is $1+$ an anticausal operator so both sides of (3.7) must be 1 (cf. Ref. 1). Hence [recall $\hat{\Xi}(T)=\operatorname{Tr}\{K_0(1+\Omega)K_0^*\}$], $\Xi_0=\min\Xi(T,K)=\hat{\Xi}(T)+\min\Xi_K(T)=\operatorname{Tr}\{2K_0(1+\Omega)K_0^*+K_0\Omega+\Omega K_0^*\}=\operatorname{Tr}\{2(1+K_0)(1+\Omega)(1+K_0^*)-2(1+\Omega)-2K_0-2K_0^*-K_0\Omega-\Omega K_0^*\}=-\operatorname{Tr}\{2\Omega+2K_0+2K_0^*+K_0\Omega+\Omega K_0^*\}=-\operatorname{Tr}\{B+B^*+K_0+K_0^*\}=\operatorname{Tr}\{B^*K_0^*+K_0B\}=0$ (since K_0 and B are causal—cf. Ref. 1). This is the desired conclusion.[12]

[1] F. Dyson, in *Studies in Math. Physics* (Princeton, U.P., Princeton, NJ, 1976), pp. 151–167.
[2] F. Dyson, J. Opt. Soc. Am. **65**, 551–558 (1975).

[3]R. Carroll, *Transmutation, Scattering Theory, and Special Functions* (North-Holland, Amsterdam, 1982).
[4]K. Case, Advances in Math. Suppl. Studies, **3**, 25–43 (1978).
[5]H. Dym and H. McKean, *Gaussian Processes, Function Theory, and the Inverse Spectral Problem* (Academic, New York, 1976).
[6]R. Carroll and F. Santosa, in *Proceedings of the Conference on Inverse Scattering*, University of Tulsa, 1983 (in press).
[7]R. Carroll and F. Santosa, "Spectral measures and autocorrelation via transmutation," C. R. Roy. Soc. Canada (in press).
[8]T. Kailath, IEEE Trans. Inf. Theory, **IT-20**, 145–181 (1974).
[9]K. Chadan and P. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer, New York, 1977).
[10]L. Faddeev, Uspehi Mat. Nauk **14**, 57–119 (1959).
[11]V. Marcenko, *Sturm-Liouville Operators and Their Applications*, (Izd. Nauk. Dumka, Kiev, 1977)—see also Ref. 3.
[12]This calculation suggests (as is indeed the case) that the characterization of K_0 by minimization does not require the trace argument [i.e., the last integral in (2.3)]; the details will appear elsewhere.

Inverse scattering in dimension two[a)]

Margaret Cheney
Department of Mathematics, Stanford University, Stanford, California 94305

(Received 17 May 1983; accepted for publication 5 August 1983)

The inverse scattering problem is solved for the two-dimensional time-independent Schrödinger equation. That is, the potential is reconstructed from the scattering amplitude, which is assumed to be known for all energies and angles.

PACS numbers: 03.65.Nk

INTRODUCTION

Our goal here is to solve the inverse scattering problem for the Schrödinger equation in two dimensions. That is, we recover the potential from the scattering data, which we take to be the entire scattering amplitude as a function of the energy and two angles.

Actually, there are a number of aspects to the inverse scattering problem: uniqueness, reconstruction, construction, and characterization. The uniqueness problem deals with the question, "Does the scattering amplitude uniquely determine the potential?" The reconstruction problem is the problem of constructing a potential from scattering data that are known to come from an underlying potential, whereas the construction problem is to construct the potential without this knowledge. And finally, the characterization problem is to determine what scattering data actually arise and to correlate properties of the potential with properties of the data.

In the one-dimensional case, solutions to all these questions via the Gel'fand–Levitan and Marchenko methods are well known. Moreover, in the 25 years since their discovery, one-dimensional inverse scattering techniques have been found to have important applications not only to particle physics but also to geophysics and to certain classes of nonlinear differential equations, the so-called soliton equations, which themselves describe a wide range of phenomena.

The popularity of one-dimensional inverse scattering has inspired much interest in the construction of higher-dimensional inversion theories; nevertheless, the uniqueness question was for many years the only one of the higher-dimensional inversion questions that was answered satisfactorily: although in the one-dimensional case additional bound state information is needed for uniqueness, in three dimensions the scattering data alone do indeed determine the potential uniquely. The other three inversion questions, however, are so much more difficult than their one-dimensional counterparts that for 25 years attempts to solve even the simplest one, the reconstruction problem, met with only partial success.

The first of these reconstruction attempts was made by Kay and Moses,[1,2] whose generalization of the Gel'fand–Levitan method accomplished inversion in a class of potentials which includes those that are nonlocal (i.e., are not multiplication operators) in the angular variables. This class, however, was never shown to include the local potentials. Another attempt, made by Faddeev[3] and Newton,[4] depended on a new, direction-dependent Green's function which had been constructed by Faddeev.[5,6] This Faddeev–Newton method, however, was awkward and cumbersome, and was hampered by a number of unanswered questions concerning exceptional points. A third attempt at multidimensional inverse scattering was made by Prosser,[7–9] who attacked all three of the remaining inversion problems using essentially an iterative scheme that applies only to weak potentials and to scattering data that are small in a certain norm. Recently, Morawetz[10] has found a generalization to higher dimensions of the Deift–Trubowitz one-dimensional method.[11] Her scheme, which is also iterative, has yet to be shown to converge for any specific class of potentials. Then, beginning in 1980, Newton published a series of papers[12–14] containing successful and elegant generalizations of both the Gel'fand–Levitan and Marchenko methods to three dimensions. Both his methods solve the reconstruction problem; his Marchenko method, in addition, solves the construction problem and gives a partial solution to the characterization problem. In this paper, we shall adapt Newton's generalized Marchenko method to dimension two.

Newton's ideas could, in fact, be applied to inverse scattering in any dimension provided that the relevant estimates hold; the success of Newton's inverse scattering techniques in two dimensions thus depends on estimates that can be considered part of the direct scattering problem.

The first five sections therefore contain the necessary results concerning direct scattering. Section 1 sets up the problem and contains basic facts and definitions for scattering in two dimensions. Also contained in Sec. 1 is a result on the behavior of the wave function for large energy. Section 2 contains the investigation of the wave function's small energy behavior.

Knowledge of the energy dependence of the wave function is crucial to our method of inverse scattering. In fact, the behavior at zero and infinity, which is heavily dimension-dependent, is the reason that later estimates must have proofs quite different from those of the corresponding estimates of Newton.[12,13] The behavior in two dimensions differs from that in three dimensions in its faster decay at infinity and in the presence of zero-energy singularities that appear in the derivatives.

The properties of symmetry and analytic continuation, however, are exactly the same as in three dimensions. These properties are recorded in Sec. 3.

[a)] This is based on the author's Ph.D. thesis, "Quantum Mechanical Scattering and Inverse Scattering in Two Dimensions," Indiana University, 1982.

 0022-2488/84/010094-14$02.50

Another ingredient essential to inverse scattering is a good deal of spectral theory. Fortunately many of the needed results have already been proved by Agmon[15] and are merely quoted in Sec. 4. These include not only the unitarity of the S matrix but also the eigenfunction expansion theorem, which is used in Sec. 5 to prove that the scattering operator maps incoming to outgoing wave functions. This relation, when combined with the analyticity properties of the wave function, forms a Riemann–Hilbert problem or a Wiener–Hopf factorization problem. This is the key to the Marchenko method of inversion.

We arrive at Sec. 6 having proved all the estimates necessary for the generalized Marchenko method of inverse scattering. The inverse scattering results, therefore, are all contained in this section; in fact the reader interested only in the results might read just Sec. 6, referring to Sec. 1 for notation. Section 6 is intended merely to give the reader a taste of the inverse scattering theory that is more fully developed in Newton's series of papers and which is generally dimension-independent. Nevertheless, in Sec. 6, the uniqueness theorem is proved, the Marchenko equation is derived, and the potential is extracted from the solution of the Marchenko equation. Thus the results of Sec. 6 solve only the reconstruction problem; the reader interested in construction should refer to Newton's work.[13,14]

Notation

In what follows we denote by $\|\cdot\|_p$ the usual L^p norm; if confusion is possible, we will add as a superscript the variable with respect to which the L^p norm is being taken.

The symbol $\|\cdot\|_{m,p}$ denotes the norm of the Sobolev space $W^{m,p}$, the space of functions with m derivatives in L^p. We shall write $H^2 = W^{2,2}$.

The symbols $\theta,\theta',\theta'',\phi$, etc., in most places denote unit vectors, although occasionally they will be used as simple angles in carrying out integrations. Where confusion is possible, the unit vectors will be decorated with hats, e.g., $\hat{\theta}$.

1. PRELIMINARIES

Two-particle scattering in the center of mass system is governed by the time-independent Schrödinger equation

$$-\Delta\psi(k,x) + V(x)\psi(k,x) = k^2\psi(k,x).$$

Here $x\in R^2$, the potential $V(x)$ is real-valued, and k is a positive scalar.

Scattering solutions are defined by the Lippmann–Schwinger equation

$$\psi(k,\theta,x) = \exp(ik\theta\cdot x) + \int G(k,|x-y|)V(y)\psi(k,\theta,y)\,d^2y, \tag{1.1}$$

where θ denotes a unit vector in R^2 and the function G is a fundamental solution of $\Delta + k^2$. We take G to be

$$G(k,r) = -(i/4)H_0^{(1)}(kr),$$

where H_0 is the zero-order Hankel function and $r = |x|$.

In order to apply Fredholm theory, we multiply the Lippman–Schwinger equation by $|V(x)|^{1/2}$ and make the following definitions:

$$\xi(k,\theta,x) = |V(x)|^{1/2}\psi(k,\theta,x),$$
$$\xi^0(k,\theta,x) = |V(x)|^{1/2}\exp(ik\theta\cdot x),$$
$$V_{1/2}(y) = V(y)|V(y)|^{-1/2},$$
$$K(k)f(x) = \int |V(x)|^{1/2}G(k,|x-y|)V_{1/2}(y)f(y)\,d^2y.$$

With this notation, the Lippmann–Schwinger equation becomes

$$\xi(k,\theta,x) = \xi^0(k,\theta,x) + K(k)\xi(k,\theta,x). \tag{1.2}$$

For k bounded away from zero, we recall[16] the following result concerning the operator $K(k)$:

Proposition 1.1: Suppose $V\in L^2$ with $\int\int |V(x)V(y)|\,|x-y|^{-1}\,d^2x\,d^2y = M < \infty$. Then for each $k_0 > 0$ the estimate $\|K(k)\|_{\text{H.S.}} \leqslant ck^{-1/2}$ holds for $k > k_0$, where c depends only on k_0 and on V.

Henceforth we will usually assume that V belongs to $L^1\cap L^2$ because[16] this assumption allows us to apply Fredholm theory to (1.2); we obtain a unique solution $\xi(k,\theta,x)$ provided the operator K does not have the eigenvalue 1. Note that for k large enough, the operator norm of $K(k)$ is less than 1, which certainly implies that (1.2) is uniquely solvable (by iteration, in fact).

We recall[16] that for V belonging to $L^1\cap L^2$ with $\int |V(x)|\,|x|^4\,d^2x < \infty$, the large x behavior of scattering states is given by

$$\psi(k,\theta,x) = \exp(ik\theta\cdot x) + \exp(-3\pi i/4)(8\pi)^{-1/2}A(k,\hat{x},\theta)\times\exp(ik|x|)(k|x|)^{-1/2} + h(k,\theta,x), \tag{1.3}$$

where $\hat{x} = x/|x|$,

$$A(k,\theta,\theta') = \int \exp(-ik\theta\cdot x)V(x)\psi(k,\theta',x)\,d^2x, \tag{1.4}$$

and

$h(k,\theta,x)\in L^2(x)$ uniformly in θ.

The quantity $A(k,\theta,\theta')$ is called the *scattering amplitude*; it essentially gives us the large x behavior of the wave function. We let the scattering amplitude act on $L^2(S^1)$ via $(A(k)f)(\theta') = \int_{S^1} A(k,\theta,\theta')f(\theta')\,d\theta$; the operator $A(k)$ is then bounded[16] and linear on $L^2(S^1)$. We also define the *scattering operator* or *S matrix* $S(k)$ on $L^2(S^1)$ by

$$S(k) = I - i(4\pi)^{-1}(\text{sgn } k)A(k).$$

In later sections we will also need the following information on the large k behavior of ψ.

Lemma 1.2: Let $V\in L^2\cap W^{2,1}$ and suppose that for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with

$$\int_0^\infty F(r)r\,dr \leqslant c\|V\|_{2,1},$$

and for some $\epsilon > 0$, $F(r) < Mr^{-1+\epsilon}$ near $r = 0$. Let k_0 be so large that, for $k > k_0$, $\|K(k)\| \leqslant a < 1$. Then for $k > k_0$, we have the estimate

$$|\psi(k,\theta,x) - \exp(ik\theta\cdot x)| \leqslant c\,k^{-(1+\epsilon/2)/2},$$

where c depends only on k_0 and V.

Proof: See Appendix A.

2. BEHAVIOR AT $k = 0$

Since the kernal of the operator $K(k)$ contains a Hankel function with a logarithmic divergence at the origin, one might expect the operator $K(k)$ and the wave function $\psi(k)$ to diverge logarithmically in some sense at the origin. However, as we shall see, the logarithmic divergence of $K(k)$ is due entirely to a rank-1 piece, and this prevents $\psi(k)$ from diverging at $k = 0$.

We recall[16] that properties of the Hankel function allow us to write $K(k) = L(k) + P \log k$, where

$$L(k)f(x) = \frac{-i}{4}\int |V(x)|^{1/2} \times\left(H_0^{(1)}(k|x-y|) - \frac{2i}{\pi}\log k\right) \times V_{1/2}(y)f(y)\,d^2y,$$

$$Pf(x) = (2\pi)^{-1}|V(x)|^{1/2}(V_{1/2},f).$$

If V is in L^1 with $\int|V(x)|\,|x|\,d^2x$ and $\int\int|V(x)V(y)|\,|\log|x-y|\,|^2\,d^2x\,d^2y$ finite, then L is a Hilbert–Schmidt operator and is well behaved at $k = 0$.

To investigate the behavior of ψ for k near zero, we will need the followimg lemma and its corollary:

Lemma 2.1: Suppose $V \in L^1$ with $\int|x|^{2\alpha}|V(x)|\,d^2x$ finite for some $0 < \alpha < 1$. Then $\|(\exp(ik\theta\cdot x) - 1)|V|^{1/2}\|_2 < ck^\alpha$ for k near zero.

Proof: Note that $\exp(ik\theta\cdot x) - 1 = (k\theta\cdot x)^\alpha h_\alpha(k\theta\cdot x)$, where $h_\alpha(it) = (\exp it - 1)t^{-\alpha}$. The function h is bounded because it is continuous and decays to zero for both large and small t. Thus

$$\|(\exp(ik\theta\cdot x) - 1)|V|^{1/2}\|_2^2 = \int(k\theta\cdot x)^{2\alpha}h_\alpha^2(ik\theta\cdot x)|V(x)|\,d^2x \leqslant k^{2\alpha}c\int|x|^{2\alpha}|V(x)|\,d^2x. \quad \text{QED}$$

Corollary 2.2: Suppose $V \in L^1 \cap L^2$ with $\int|x|^{2\alpha}|V(x)|\,d^2x$ finite for some $0 < \alpha < 1$, and suppose $(I - L(0))^{-1}$ exists. Then for $\xi_0(k) = \exp(ik\theta\cdot x)|V(x)|^{1/2}$ and for k near zero,

$$\|(I - L(k))^{-1}\xi^0(k) - (I - L(k))^{-1}|V|^{1/2}\|_2 \leqslant ck^\alpha.$$

Proposition 2.3: Let $V \in L^1 \cap L^2$ with $\int|x|\,|V(x)|\,d^2x < \infty$, and suppose $(I - L(0))^{-1}$ exists. Then $\xi(k)$ satisfies

$$\xi(k) = (I - L(k))^{-1}\xi^0(k) + \frac{(V_{1/2},(I - L(k))^{-1}\xi^0(k))\log k}{2\pi - (V_{1/2},(I - L(k))^{-1}|V|^{1/2})\log k} \times (I - L(k))^{-1}|V|^{1/2}. \tag{2.1}$$

The L^2 norm is bounded by

$$\|\xi(k)\|_2 \leqslant \begin{cases} c(\log k)^{-1} & \text{if } a_0 \neq 0, \\ c & \text{if } a_0 = 0, \end{cases}$$

where $a_0 = (V_{1/2},(I - L(0))^{-1}|V|^{1/2})$.

Proof: We shall solve the equation $(I - K(k))\xi = \xi^0$ assuming that $(I - L(k))^{-1}$ exists in a neighborhood of $k = 0$. Rewriting the equation in terms of the operators P and L, we have

$$(I - L(k))\xi - (\log k)P\xi = \xi^0. \tag{2.2}$$

Since P is a rank-1 operator, it will turn out that $P\xi = a|V(x)|^{1/2}$, where the constant a is given by

$$a = (2\pi)^{-1}(V_{1/2},\xi). \tag{2.3}$$

We will determine a at the end of our calculation. In the meantime, writing $P\xi = a|V|^{1/2}$, we can solve Eq. (2.2):

$$\xi = (I - L(k))^{-1}[\xi^0 + a(\log k)|V|^{1/2}]. \tag{2.4}$$

It now remains to determine the value of a. To do this, we use (2.3) and (2.4):

$$\begin{aligned} a &= (2\pi)^{-1}(V_{1/2},(I - L(k))^{-1}[\xi^0 + a(\log k)|V|^{1/2}]) \\ &= (2\pi)^{-1}(V_{1/2},(I - L(k)^{-1}\xi^0) \\ &\quad + a(2\pi)^{-1}(\log k)(V_{1/2},(I - L(k))^{-1}|V|^{1/2}). \end{aligned}$$

Solving this linear equation for a gives

$$a = \frac{(V_{1/2},(I - L(k))^{-1}\xi^0)}{2\pi - (\log k)(V_{1/2},(I - L(k))^{-1}|V|^{1/2})}.$$

Substitution of this value for a back into our expression for ξ, (2.4), gives us (2.1).

We now compute the limit as $k \to 0$ of (2.1). We write $\xi^0 = |V|^{1/2} + (\xi^0 - |V|^{1/2})$; by Corollary 2.2, the inner product in the numerator of (2.1) is $(V_{1/2},(I - L(k))^{-1}|V|^{1/2})$ plus something that decays like k^α as $k \to 0$. In the limit as $k \to 0$, we may therefore replace the ξ^0 by $|V|^{1/2}$. We recall[16] that differentiability of $(I - L(k))^{-1}$ allows us to write $(V_{1/2},(I - L(k))^{-1}|V|^{1/2}) = a_0 + a_1k$, where

$$a_0 = (V_{1/2},(I - L(0))^{-1}|V|^{1/2})$$

and a_1 is bounded for small k. With this in mind we compute

$$\xi(0) = \begin{cases} (I - L(0))^{-1}|V|^{1/2} & \text{if } a_0 = 0, \\ 0 & \text{if } a_0 \neq 0. \end{cases}$$

We will also need a bound for $\|\xi(k)\|_2$. Equation (2.1) yields

$$\xi(k) = (I - L(k))^{-1}(\xi^0(k) - |V|^{1/2}) + (I - L(k))^{-1}|V|^{1/2} \times\left(1 + \frac{[(V_{1/2},(I - L(k))^{-1}|V|^{1/2}) + (V_{1/2},(I - L(k))^{-1}(\xi^0 - |V|^{1/2}))]\log k}{2\pi - (V_{1/2},(I - L(k))^{-1}|V|^{1/2})\log k}\right).$$

Corollary 2.2 then gives

$$\|\xi(k)\|_2 \leqslant ck + \|(I - L(k))^{-1}\|\,\|V\|_1\frac{2\pi + ck^\alpha\log k}{2\pi + (a_0 + a_1k)\log k} \leqslant \begin{cases} c(\log k)^{-1} & \text{if } a_0 \neq 0 \\ c & \text{if } a_0 = 0. \end{cases} \quad \text{QED}$$

3. SYMMETRY AND ANALYTIC CONTINUATION

So far the wave function ψ has been defined only for positive k—the speed of the incoming particle. However, the Lippman–Schwinger equation makes sense for other values of k as well.

Invariance of the fundamental solution and of the plane wave under simultaneous complex conjugation and substitution of $-k$ for k shows that the wave function satisfies

$$\overline{\psi(-k,\theta,x)} = \psi(k,\theta,x).$$

This equation defines the wave function for negative k.

Similarly there is a relation between the incoming and outgoing waves

$$\psi^-(k,\theta,x) = \psi(-k,-\theta,x).$$

We will also need the reciprocity theorem, which is an expression of time reversal invariance of the scattering process.

Proposition 3.1 (Reciprocity Theorem): Let $V \in L^1 \cap L^2$. Then $A(k,\theta,\theta') = A(k,-\theta',-\theta)$.

Proof: See Appendix B.

Next we turn to the analyticity properties of the wave function as a function of k, which we now consider as a complex variable. Since we obtain the wave function only by means of the Lippman–Schwinger equation, we must analytically continue the integral equation.

First we note that the operator GV is Hilbert–Schmidt in the open upper half k-plane.

Proposition 3.2: Let $V \in L^2$. Then for $\operatorname{Im} k > 0$ the operator $G(k)V$ given by $G(k)Vf(x) = (-i/4)\int H_0(k|x-y|)V(y) f(y)\, d^2y$ is Hilbert–Schmidt, and $\|G(k)V\|_{\text{H.S.}} \leqslant c|k|^{-1}$.

Proof:

$$\begin{aligned}\|GV\|^2_{\text{H.S.}} &= c\int\int |H_0(k|x-y|)V(y)|^2\, d^2x\, d^2y \\ &= c\|V\|_2^2 \int |H_0(k|z|)|^2\, d^2z \\ &= c\|V\|_2^2 |k|^{-2} \int |H_0(|z'|k/|k|)|^2\, d^2z' \\ &< c|k|^{-2}. \qquad \text{QED}\end{aligned}$$

Similarly we have:

Proposition 3.3: Let $V \in L^2$. Then the operator $K(k)$ (defined in Sec. 1) is Hilbert–Schmidt for $\operatorname{Im} k \geqslant 0$, $k \neq 0$.

Proof: The proof is similar to that of Proposition 3.2. QED

However, the inhomogeneity in Eq. (1.2) is not in L^2 for $\operatorname{Im} k > 0$; we multiply the equation by $\exp(-ik\theta\cdot x)$ to obtain

$$\chi(k,\theta,x) = |V(x)|^{1/2} + \mathscr{K}(k)\chi(k,\theta,x),$$

where

$$\chi(k,\theta,x) = |V(x)|^{1/2}\psi(k,\theta,x)\exp(ik\theta\cdot x)$$

and where $\mathscr{K}(k)$ depends on θ:

$$\mathscr{K}(k)f(x) = \frac{-i}{4}\int |V(x)|^{1/2} H_0^{(1)}(k|x-y|)V_{1/2}(y) \times \exp(-ik\theta\cdot(x-y))f(y)\, d^2y.$$

Proposition 3.4: Let $V \in L^2$ with

$$\int\int \frac{|V(x)V(y)|}{|x-y|}\, d^2x\, d^2y < \infty.$$

Then the operator $\mathscr{K}(k)$ defined above is Hilbert–Schmidt for $\operatorname{Im} k \geqslant 0$, $k \neq 0$, and satisfies $\|\mathscr{K}(k)\|_{\text{H.S.}} \leqslant c|k|^{-1/2}$.

Proof: We apply the definition of the Hilbert–Schmidt norm to the operator $\mathscr{K}$:

$$\|\mathscr{K}(k)\|^2_{\text{H.S.}} = c\int\int |V(x)V(y)|\, |H_0^{(1)}(k|x-y|)|^2 \times \exp(2\operatorname{Im} k\theta\cdot(x-y))\, d^2x\, d^2y = c(I_1 + I_2),$$

where I_1 and I_2 are the integrals over the sets $|k|\,|x-y| < 1$ and $|k|\,|x-y| > 1$, respectively.

First we consider I_1. We use the small-argument behavior of the Hankel function to bound I_1 by

$$I_1 \leqslant \int\int_{|k|\,|x-y|<1} |V(x)V(y)||\log k|x-y||^2 \times \exp(2\operatorname{Im} k\theta\cdot(x-y))\, d^2x\, d^2y.$$

Next we let $z = x - y$ and note that for $|kz| < 1$, $\operatorname{Im} k\theta\cdot z \leqslant |kz| < 1$. Thus we have

$$\begin{aligned}I_1 &\leqslant c\int\int_{|kz|<1} |V(z+y)V(y)|\, |\log k|z||^2\, d^2z\, d^2y \\ &< c\|V\|_2^2 \int_{|kz|<1} |\log k|z||^2\, d^2z < c\|V\|_2^2 |k|^{-2}.\end{aligned}$$

We now turn to I_2. We use the large-argument behavior of the Hankel function to bound I_2 by

$$I_2 \leqslant \int\int_{|k|\,|x-y|>1} |V(x)V(y)|\exp(-2\operatorname{Im} k(|x-y| + \theta\cdot(x-y)))(|k|x-y||)^{-1}\, d^2x\, d^2y.$$

Note that the coefficient of $-2\operatorname{Im} k$ in the exponent is always positive; thus the exponential is bounded by 1. Use of this fact gives us

$$I_2 \leqslant |k|^{-1}\int\int \frac{|V(x)V(y)|}{|x-y|}\, d^2x\, d^2y < c|k|^{-1}. \qquad \text{QED}$$

Corollary 3.5: Let $V \in L^1 \cap L^2$. Then, for each θ, $\chi(k,\theta,x) = |V(x)|^{1/2}\psi(k,\theta,x)\exp(ik\theta\cdot x)$ is a meromorphic L^2-valued function of k for $\operatorname{Im} k > 0$.

Remark 3.6: A similar argument shows that, for $V \in L^1 \cap L^2$ and for each θ, $\chi^-(k,\theta,x) = |V(x)|^{1/2}\psi^-(k,\theta,x)\exp(-ik\theta\cdot x)$ is a meromorphic L^2-valued function of k for $\operatorname{Im} k < 0$.

4. AGMON'S SPECTRAL THEORY RESULTS

In this section we shall quote various results of Agmon[15] that will be used in the next section.

Let $L^{2,s}(R^2)$ denote the space of complex-valued functions $u(x)$ on R^2 with $(1+|x|^2)^{s/2}u(x) \in L^2(R^2)$, and let the weighted Sobolev spaces $H^{m,s}$ consist of $L^{2,s}$ functions with the first m derivatives also in $L^{2,s}$.

Agmon proves the following three theorems:

Theorem 4.1: Let $H = -\Delta + V$, where $V \in L^2_{\text{loc}}$ with $V(x) = O(|x|^{-1-\epsilon})$ as $|x| \to \infty$. Consider the resolvent $(H-E)^{-1}$ as an analytic operator-valued function on

$\mathbf{C}\backslash\sigma(\mathrm{H})$ with values in $B(L^{2,s},H^{2,-s})$ for any $S>\frac{1}{2}$. Then for real $E\neq 0$, the limits

$$\lim_{\epsilon\to 0}(H-E\pm i\epsilon)^{-1}$$

exist in the uniform operator topology of $B(L^{2,s},H^{2,-s})$.

Theorem 4.2: Let $H=-\Delta+V$, where $V\in L^2_{\mathrm{loc}}$ with $V(x)=O(|x|^{-3/2-\epsilon})$ as $|x|\to\infty$. Then there exist two families $\phi_\pm(k,\theta,x)$ of generalized eigenfunctions of H such that for every fixed k and θ, $\phi_\pm(k,\theta,x)$ as a function of x belongs to $C(R^2)\cap H^2_{\mathrm{loc}}(R^2)$ and satisfies the Schrödinger equation. Furthermore, for almost all $\theta\in S^1$, $\phi_\pm$ satisfies

$$\phi_\pm(k,\theta,x)-\exp(ik\theta\cdot x)$$

$$=-\lim_{\epsilon\to 0}(H-k^2\pm i\epsilon)^{-1}(V(x)\exp(ik\theta\cdot x)). \tag{4.1}$$

The eigenfunctions ϕ are continuous in k,θ, and x.

Theorem 4.3: Let $H=-\Delta+V$ where $V\in L^2_{\mathrm{loc}}$ with $V(x)=0(|x|^{-3/2-\epsilon})$ as $|x|\to\infty$, and let $\phi_\pm$ be the above family of generalized eigenfunctions. Let $P_{(a^2,b^2)}$ for $a>0$ denote the usual spectral projection. Then for any $f\in L^2$,

$$(P_{(a^2,b^2)}f)(x)=(2\pi)^{-2}\int_a^b\int_{S^1}\phi_\pm(k,\theta,x)$$
$$\times\int\overline{\phi_\pm(k,\theta,y)}f(y)\,d^2y\,d\theta\,k\,dk.$$

We must now relate Agmon's generalized eigenfunctions $\phi_\pm$ to our wave functions $\psi_\pm$. We first obtain a relation between the full and free resolvents by multiplying the relation $-\Delta+V+E=(-\Delta+E)+V$ on the left by $(-\Delta+E)^{-1}$ and on the right by $(-\Delta+V+E)^{-1}$. This gives us the relation

$$(-\Delta+E)^{-1}=(-\Delta+V+E)^{-1}+(-\Delta+E)^{-1}$$
$$\times V(-\Delta+V+E)^{-1}.$$

Multiplication on the left by $(I+(-\Delta+E)^{-1}V)^{-1}$ gives us

$$(-\Delta+V+E)^{-1}=(I+(-\Delta+E)^{-1}V)^{-1}$$
$$\times(-\Delta+E)^{-1}.$$

In Agmon's notation this is

$$-(H-k^2\pm i\epsilon^2)^{-1}=(I-G(\mp k+i\epsilon)V)^{-1}$$
$$\times G(\mp k+i\epsilon).$$

Upon composition with the multiplication operator $V_{1/2}$, this is

$$-(H-k^2\pm i\epsilon^2)^{-1}V_{1/2}$$
$$=(I-GV)^{-1}|V|^{-1/2}|V|^{1/2}GV_{1/2}$$
$$=|V|^{-1/2}(I-K(\mp k+i\epsilon))^{-1}K(\mp k+i\epsilon). \tag{4.2}$$

The formula (4.1) can then be expressed as (4.2) applied to ξ^0,

$$\phi_\pm(k,\theta,x)$$
$$=\exp(ik\theta\cdot x)-\lim_{\epsilon\to 0}(H-k^2\pm i\epsilon^2)^{-1}(V(x)\exp(ik\theta\cdot x))$$
$$=(|V(x)|^{-1/2}I+|V(x)|^{-1/2}(I-K^\mp)^{-1}K^\mp)\xi^0$$
$$=|V(x)|^{-1/2}(I+(I-K^\mp)^{-1}K^\mp)\xi^0$$
$$=|V|^{1/2}(I-K^\mp)^{-1}\xi^0$$
$$=|V|^{1/2}\xi^\mp(k,\theta,x)$$
$$=\psi^\mp(k,\theta,x).$$

5. THE SCATTERING OPERATOR

In this section we investigate some of the properties of the scattering operator.

The Marchenko method of inverse scattering rests on the following theorem (see Ref. 17).

Theorem 5.1: Let $V\in L^2_{\mathrm{loc}}$ with $V(x)=O(|x|^{-2-\epsilon})$ as $|x|\to\infty$. Let the scattering amplitude act on $L^2(S^1)$ via

$$A(k)f(\theta)=\int_{S^1}A(k,\theta',\theta)f(\theta')\,d\theta',$$

and let the scattering operator $S(k)$ be defined as an operator on $L^2(S^1)$ by

$$S(k)=I-i(4\pi)^{-1}\operatorname{sgn}k\,A(k). \tag{5.1}$$

Then

$$S(k)\psi^-(k,\theta,x)=\psi^+(k,\theta,x). \tag{5.2}$$

Remarks: The factor sgn k in (5.1) is needed to make (5.2) hold for negative k.

We shall show that the equality in (5.2) holds in the sense of $H^{2,-s}$ for some $s>\frac{1}{2}$; however, we recall (Theorem 4.2) that ψ^+ and ψ^- are continuous in x. Equation (5.2) therefore holds for each x.

Proof: We use Theorem 4.2 to write out the expression

$$\psi^+(k,\theta,x)-\psi^-(k,\theta,x)=\lim_{\epsilon\to 0}((H-k^2+i\epsilon)^{-1}$$
$$-(H-k^2-i\epsilon)^{-1})(V(x)\exp(ik\theta\cdot x)),$$

where the limit is in the $H^{2,-s}$ norm for some $x>\frac{1}{2}$. By Stone's formula,[18] the jump in the resolvent is the spectral projection, which in turn is given by Theorem 4.3:

$$\lim_{\epsilon\to 0}(2\pi i)^{-1}\int_{k_0}^{k_0+\delta}((H-k^2+i\epsilon)^{-1}$$
$$-(H-k^2-i\epsilon)^{-1})2k\,dk\,(V(x)\exp(ik\theta\cdot x))$$
$$=-P_{(k_0^2,(k_0+\delta)^2)}(V(x)\exp(ik\theta\cdot x))$$
$$=-(2\pi)^{-2}\int_{k_0}^{k_0+\delta}\int_{S^1}\psi^-(k,\theta',x)\int\overline{\psi^-(k,\theta',y)}V(y)$$
$$\times\exp(ik\theta\cdot y)\,d^2y\,d\theta'\,k\,dk.$$

Because of the symmetry properties of the wave function and scattering amplitude set forth in Sec. 3, the y integral is precisely $A(k,\theta',\theta)$. To remove the integration over k, we next multiply by $1/\delta$ and take the limit as δ approaches zero. Provided that the δ and ϵ limits are interchangeable, continuity in k gives us

$$\psi^+(k,\theta,x)-\psi^-(k,\theta,x)$$
$$=-i(4\pi)^{-1}\int_{S^1}\psi^-(k,\theta',x)A(k,\theta',\theta)\,d\theta'.$$

This proves the theorem, provided we can show that we may interchange the δ and ϵ limits. In other words, we must show that the following expression approaches zero as δ goes to zero:

$$I(\delta)$$
$$= \left\| \lim_{\epsilon\to 0}\left(\frac{2}{\delta}\right)\int_{k_0}^{k_0+\delta} ((H-k^2-i\epsilon)^{-1}-(H-k^2+i\epsilon)^{-1}) \times (V(x)\exp(ik\theta\cdot x))k\,dk \right.$$
$$\left. - 2k\lim_{\epsilon\to 0}((H-k^2-i\epsilon)^{-1}-(H-k^2+i\epsilon)^{-1}) \times (V(x)\exp(ik\theta\cdot x))\right\|_{2,2,-s},$$

where the norm is the $H^{2,-s}$ norm. We write the second term as the integral of a constant vector times $1/\delta$. Then continuity of the norm allows us to bring the ϵ limit outside; since the k integral is a limit of sums, the triangle inequality tells us that we can only increase $I(\delta)$ by bringing the norm inside the integral. We shall consider only I_1, the $-\epsilon$ term; the $+\epsilon$ term is similar. We have

$$\begin{aligned} I_1(\delta) &\leqslant \frac{2}{\delta}\lim_{\epsilon\to 0}\int \{\|((H-k^2-i\epsilon)^{-1} \\ &\quad -(H-k_0^2-i\epsilon)^{-1})(V(x)\exp(ik\theta\cdot x))\|_{2,2,-s} \\ &\quad +\|(H-k_0^2-i\epsilon)^{-1}(V(x)(\exp(ik\theta\cdot x) \\ &\quad -\exp(ik_0\theta\cdot x)))\|_{2,2,-s}\}k\,dk \\ &\leqslant 2\lim_{\epsilon\to 0}\max_{(k_0,k_0+\delta)}(\|(H-k^2-i\epsilon)^{-1} \\ &\quad -(H-k_0^2-i\epsilon)^{-1}\|\,\|V\|_{0,2,s} \\ &\quad +\|(H-k_0^2-i\epsilon)^{-1}\|\,\|V(x)(\exp(ik\theta\cdot x) \\ &\quad -\exp(ik_0\theta\cdot x))\|_{0,2,-s}), \end{aligned}$$

which goes to zero by continuity of the resolvent. [The hypothesis $V(x)=0(|x|^{-2-\epsilon})$ insures that $\|V(x)(\exp(ik\theta\cdot x) - \exp(ik_0\theta\cdot x))\|_{0,2,s}\to 0$.] QED

The following estimate on the scattering operator will allow us to define a Fourier transform in the next section:

Proposition 5.2: Let $V\in W^{2,1}$, and suppose that, for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with $\int_0^\infty F(r)r\,dr\leqslant c\|V\|_{2,1}$ and $F(r)<Mr^{-1+\epsilon}$ near 0 for some $\frac{1}{2}>\epsilon>0$. Then

$$\int_{-\infty}^{\infty}(\|(S(k)-I)f\|_2^{(\theta)})^2\,dk\leqslant c(\|f\|_2^{(\theta)})^2.$$

Proof: Turn to Appendix C.

6. INVERSE SCATTERING

This section is devoted to methods discovered by Newton[12] of extracting the potential $V(x)$ from the scattering amplitude $A(k,\theta,\theta')$.

We note first that the scattering amplitude does indeed uniquely determine the potential:

Theorem 6.1 (Uniqueness): Suppose the scattering amplitude $A(k,\theta,\theta')$ is constructed from a potential $V(x)$ belonging to $L^1\cap L^2$. Then the Fourier transform $\hat{V}$ can be recovered by means of the formula

$$\hat{V}(x)=\lim_{\substack{k\to\infty\\ k(\theta-\theta')=x}} A(k,\theta,\theta'). \tag{6.1}$$

This limit, an ordinary pointwise limit, is uniform in the sense that the difference

$$\hat{V}(k(\theta-\theta'))-A(k,\theta,\theta') \tag{6.2}$$

goes to zero uniformly in both angles as k goes to infinity.

Proof: We write out the definition of each term of (6.2) and apply the Schwarz inequality:

$$\begin{aligned} &|\hat{V}(k(\theta-\theta'))-A(k,\theta,\theta')| \\ &\quad=\left|\int \exp(ik\theta\cdot x)V(x)(\exp(ik\theta'\cdot x)-\psi(k,\theta',x))\,d^2x\right| \\ &\quad\leqslant\|V\|_1\|\,|V|^{1/2}(\psi(k,\theta',x)-\exp(ik\theta'\cdot x))\|_2. \end{aligned}$$

In the notation of Eq. (1.2), the second factor of this last expression is $\xi-\xi^0$. We write (1.2) as $\xi=(I-K)^{-1}\xi^0 = \xi^0+K(I-K)^{-1}\xi^0$. This allows us to bound (6.2) by

$$\begin{aligned} &|\hat{V}(k(\theta-\theta'))-A(k,\theta,\theta')| \\ &\quad\leqslant\|V\|_1\,\|K(k)\|\,\|(I-K(k))^{-1}\|\,\|V\|_1. \end{aligned} \tag{6.3}$$

By Proposition 1.1, the right side of (6.3) goes to zero as k becomes infinite. QED

Remark: Formula (6.1) is the well-known *Born approximation*. It gives a simple solution of the inverse scattering problem provided that the scattering amplitude is known for all k. In fact, this method of inversion depends exclusively on the high energy scattering data, which in practice may be known only approximately. There is, therefore, reason to investigate other inversion techniques, especially those whose dependence on high energy data might be less severe. One such technique is given in the following theorem.

Remark 6.2 (Notation): We shall use the following notation. We define the operator Q: $L^2(S^1)\to L^2(S^1)$ by $Qf(\theta)=f(-\theta)$. We use $\mathcal{F}$ for the vector-valued Fourier transform in k,

$$\mathcal{F}_k f(\alpha)=(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp(-ik\alpha)f(k)\,dk,$$

where for f belonging to $L^2(S^1)$, the limit inherent in the integral is taken in the norm topology.

We write $\beta(k,\theta,x)=\psi(k,\theta,x)\exp(-ik\theta\cdot x)$ and $\eta(\alpha,\theta,x)=\mathcal{F}_k(\beta(k,\theta,x)-1)$; we note that Lemma 1.2 implies that $\beta-1$ is a square-integrable $L^2(S^1)$-valued function of k, and that therefore η is a square-integrable $L^2(S^1)$-valued function of α.

We define the operator $\mathcal{S}(k)$: $L^2(S^1)\to L^2(S^1)$ by $\mathcal{S}(k)=\exp(ik\theta\cdot x)\,\overline{S(k)}\exp(-ik\theta\cdot x)$, where the exponentials act as multiplication operators and $\overline{S(k)}$ denotes the integral operator whose kernal is the complex conjugate of that of $S(k)$. For any f belonging to $L^2(S^1)$, we write $G(\alpha)f=\mathcal{F}_k^{-1}((\mathcal{S}(k)-I)f)$; by Proposition 5.2, $G(\alpha)f$ is a square-integrable $L^2(S^1)$-valued function of α. We note that G depends on x; this dependence will, however, be suppressed in what follows. Explicitly, $G(\alpha)$ is given by

$$\begin{aligned} &G(\alpha)f(\theta) \\ &\quad=(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp(ik(\alpha+\theta\cdot x))i(4\pi)^{-1}(\operatorname{sgn}k) \\ &\qquad\times\int_{S^1}\overline{A(k,\theta',\theta)}\exp(-ik\theta'\cdot x)f(\theta')\,d\theta'\,dk. \end{aligned} \tag{6.4}$$

Theorem 6.3 (The Marchenko Equation): Suppose $V\in W^{2,1}$ with $V(x)=O(|x|^{-2-\epsilon})$ as $|x|\to\infty$, and suppose that, for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all

bounded by a decreasing positive radial function $F(|x|)$ with $\int_0^\infty F(r)r\,dr<c\|V\|_{2,1}$ and $F(r)<Mr^{-1+\epsilon}$ near zero for some $0<\epsilon<\frac{1}{2}$. Suppose further that $-\Delta+V$ has no bound or half-bound states. Then in the notation of the above remark, the following equation in $L^2(R^+\times S^1)$ holds for positive α and for fixed x

$$\eta(\alpha,\theta,x)=\int_0^\infty G(\alpha+\beta)Q\eta(\beta,\theta,x)\,d\beta+G(\alpha)1. \quad (6.5)$$

Remark: The above hypotheses on V allow, for example, radial potentials with logarithmic singularities at x_0. Absence of bound states can be ascertained by means of the two-dimensional Levinson theorem.[16] Inversion in the presence of bound states can be accomplished with the use of further dimension-independent techniques developed by Newton.[12–14]

Proof: Theorem 5.1 gives us the relation

$$S(k)\psi^-(k,\theta,x)=\psi^+(k,\theta,x); \quad (6.6)$$

Sec. 3 allows us to eliminate ψ^- from (6.6):

$$S(k)Q\psi^+(-k,\theta,x)=\psi^+(k,\theta,x). \quad (6.7)$$

Note that $\psi^+(k)$ is analytic in the upper half-plane in k while $\psi^+(-k)$ is analytic in the lower one; (6.7) is therefore a Wiener–Hopf factorization problem or a Riemann–Hilbert problem. We shall solve the problem by using the Fourier transform to convert it into an integral equation.

In (6.7), we first put $-k$ in place of k and use the fact that $S(-k)=\overline{S(k)}$:

$$\overline{S(k)}Q\psi(k,\theta,x)=\psi(-k,\theta,x);$$

then multiplication by $\exp(ik\theta\cdot x)$ gives, in the notation of Remark 6.2,

$$\mathscr{S}(k)Q\beta(k,\theta,x)=\beta(-k,\theta,x). \quad (6.8)$$

In order to apply the Fourier transform to (6.8), we must subtract off the asymptotic values:

$$\beta(-k)-1=(\mathscr{S}(k)-I)Q(\beta(k)-1)+Q(\beta(k)-1)+(\mathscr{S}(k)-I)Q1. \quad (6.9)$$

Application of the inverse Fourier transform to (6.9) now gives

$$\eta(\alpha,\theta,x)=\int_{-\infty}^{\infty}G(\alpha-\beta)Q\eta(-\beta,\theta,x)\,d\beta+Q\eta(-\alpha,\theta,x)+G(\alpha)1. \quad (6.10)$$

Note that analyticity of $\beta(k)-1$ in the upper half k-plane implies that $\eta(\alpha,\theta,x)$ is zero for negative α. Consideration of positive α only in (6.10) gives us the Marchenko equation (6.5).

Theorem 6.4 (Compactness): Let $V\in W^{3,1}$ with $\int|x|^i|V(x)|\,d^2x<\infty$ for $i=1,2,3,4$, and suppose that for some x_0, $|V(x-x_0)|$, $|\nabla\nabla(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with $\int F(r)\,dr$ and $\int F(r)r^{3/2}\,dr$ finite. Suppose also that $(I-L(0))^{-1}$ exists. Then the integral operator $\mathscr{G}$ occuring in the Marchenko equation is a Hilbert–Schmidt operator on $L^2(R^+\times S^1)$.

Proof: The proof will be given in a later paper.

Remark 5.6: Newton has shown[13] that the spectrum of $\mathscr{G}$ is in fact contained in the interval $[-1,1]$. Thus, if $\mathscr{G}$ has neither the eigenvalue 1 nor -1, then $\mathscr{G}$ is a contraction and the Marchenko equation can be solved by iteration.

The above theorem allows us to apply Fredholm theory to the Marchenko equation, and, if the spectrum of $\mathscr{G}$ does not contain the point one, to obtain a solution $\eta(\alpha,\theta,x)$ belonging to $L^2(R^+\times S^1)$ for each x. We could then invert the Fourier transform to obtain the wave function, which could then be used in the formula

$$V(x)=[(\Delta+k^2)\psi(k,\theta,x)]/\psi(k,\theta,x).$$

However, the following formal calculation gives a simpler method of recovering the potential.

We use $\psi(k,\theta,x)=\beta(k,\theta,x)\exp(ik\theta\cdot x)$ in the Schrödinger equation, and find that the function $\beta(k,\theta,x)$ satisfies the equation

$$(\Delta+2ik\theta\cdot\nabla)\beta=V. \quad (6.11)$$

Into (6.11) we substitute

$$\beta(k,\theta,x)=1+\mathscr{F}_\alpha^{-1}(\eta(\alpha,\theta,x)),$$

obtaining

$$\int_0^\infty(\Delta-V(x)+2ik\theta\cdot\nabla)\eta(\alpha,\theta,x)\exp(ik\alpha)\,d\alpha-V(x)=0. \quad (6.12)$$

Formal integration by parts of the third term of (6.12) leads to

$$V(x)+2\theta\cdot\nabla_x\eta(0,\theta,x)+\int_0^\infty\exp(ik\alpha)\left[\Delta-V(x)-\frac{\partial}{\partial\alpha}\theta\cdot\nabla\right]\eta(\alpha,\theta,x)\,d\alpha=0.$$

For smooth η, the integral will go to zero for large k and leave us with

$$[\Delta_x-V(x)-(\partial/\partial\alpha)\theta\cdot\nabla_x]\eta=0,$$
$$V(x)=-2\theta\cdot\nabla_x\eta(0,\theta,x). \quad (6.13)$$

Equation (6.13) is known as the *miracle*. It is related to the characterization problem as follows. If the scattering amplitude with which we begin is known to come from a potential satisfying the hypotheses of Theorem 6.3, then the right side of (6.13) is guaranteed to be independent of θ. However, if we begin with an inadmissible scattering amplitude (i.e., one that does not correspond to a potential), then the miracle will not be satisfied (i.e., the right side of (6.13) will depend on θ). By counting variables, it is easy to see that most randomly chosen scattering amplitudes will not lead to a miraculous solution of (6.5). This is because the scattering amplitude, a function of three variables, is being used to determine the potential, which is a function of only two variables. At present, this miracle is the only known characterization of admissible scattering amplitudes.

ACKNOWLEDGMENTS

I am grateful to my thesis advisor, Roger G. Newton, for suggesting the problem and for discussing it with me on many occasions. I would also like to thank Joe Keller for reading the manuscript and making a number of helpful

comments. The work was supported in part by the Air Force Office of Scientific Research, the National Science Foundation, the Office of Naval Research, and the Army Research Office.

APPENDIX A: LARGE k BEHAVIOR OF ψ

Lemma 1.2: Let $V\in W^{2,1}\cap L^2$, and suppose that for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with $\int_0^\infty F(r)r\,dr\leqslant c\|V\|_{2,1}$ and $F(r)<Mr^{-1+\epsilon}$ near $r=0$, for some $\epsilon>0$. Let $k_0>0$ be so large that for $k>k_0$, $\|K(k)\|\leqslant a<1$. Then, for $k>k_0$, $|\psi(k,\theta,x)-\exp(ik\theta\cdot x)|\leqslant ck^{-(1+\epsilon/2)/2}$, where c depends only on V.

Proof: The wave function ψ is defined by Eq. (1.1). Provided that k is not an exceptional point, this equation has a solution with $\xi=|V|^{1/2}\psi\in L^2$. We split the integral in (1.1) into pieces corresponding to small and large arguments of the Hankel function:

$$\int H_0^{(1)}(k|x-y|)V(y)\psi(k,\theta,y)\,d^2y=I_1+I_2+I_3+I_4,$$

where

$$I_1=\frac{-2i}{\pi}\int_{|x-y|<k^{-1}}\log(k|x-y|)V(y)\psi(k,\theta,y)\,d^2y,$$

$$I_2=\int_{|x-y|<k^{-1}}\left[H_0^{(1)}(k|x-y|)+\frac{2i}{\pi}\log(k|x-y|)\right]V(y)\psi(k,\theta,y)\,d^2y,$$

$$I_3=\int_{|x-y|>k^{-1}}2^{1/2}(\pi k|x-y|)^{-1/2}\times\exp[ik|x-y|-\tfrac{1}{4}i\pi]V(y)\psi(k,\theta,y)\,d^2y,$$

$$I_4=\int_{|x-y|>k^{-1}}[H_0^{(1)}(k|x-y|)-2^{1/2}(\pi k|x-y|)^{-1/2}\times\exp(ik|x-y|-\tfrac{1}{4}i\pi)]V(y)\psi(k,\theta,y)\,d^2y.$$

Application of the Schwarz inequality to I_1 gives

$$|I_1|\leqslant\frac{2}{\pi}\left(\int_{|x-y|<k^{-1}}|\log k|x-y||^2|V(y)|d^2y\right)^{1/2}\|\xi\|_2. \tag{A1}$$

For $k>k_0$, the second factor of (A1) is bounded by

$$\|\xi\|_2\leqslant(1+\|K\|+\|K\|^2+\cdots)\|\xi^0\|_2\leqslant(1-a)^{-1}\|V\|_1, \tag{A2}$$

where the notation is as in Eq. (1.2). In the first factor of (A1), we let $x-y=r\phi$ with $\phi=(x-y)/|x-y|$ and $r=x-y$:

$$\int_{S^1}\int_0^{k^{-1}}|\log kr|^2|V(x-r\phi)|f\,dr\,d\phi\leqslant 2\pi\int_0^{k^{-1}}(kr)^{-1-\epsilon/2}F(|r-|x+x_0||)r\,dr\leqslant 2\pi k^{-1-\epsilon/2}\int_0^{k^{-1}}F(|r-|x+x_0||)r^{-\epsilon/2}\,dr. \tag{A3}$$

The integral converges if it converges when the singularities coincide; therefore (A3) is bounded by

$$ck^{-1-\epsilon/2}\int_0^{k^{-1}}r^{-1+\epsilon/2}\,dr\leqslant ck^{-1-\epsilon}.$$

Thus we have $|I_1|\leqslant c\|V\|_1k^{-(1+\epsilon)/2}$.

We treat I_2 the same way and obtain the same bound: $|I_2|\leqslant c\|V\|_1k^{-(1+\epsilon)/2}$.

Next we consider I_3. We replace $|V|^{1/2}\psi$ by

$$|V(y)|^{1/2}\psi(k,\theta,y)=|V(y)|^{1/2}\exp(ik\theta\cdot y)+K(k)[|V(y)|^{1/2}\psi(k,\theta,y)].$$

This splits I_3 into $I_3=I_5+I_6$, where

$$I_5=\int_{|x-y|>k^{-1}}2^{1/2}(\pi k|x-y|)^{-1/2}\times\exp[ik|x-y|-\tfrac{1}{4}i\pi+ik\theta\cdot y]V(y)d^2y, \tag{A4}$$

$$I_6=\int_{|x-y|>k^{-1}}2^{1/2}(\pi k|x-y|)^{-1/2}\exp[ik|x-y|-i\pi/4]\times V_{1/2}(y)K(k)[|V(y)|^{1/2}\psi(k,\theta,y)]\,d^2y. \tag{A5}$$

First we consider I_5. Letting $z=x-y$ in (A4) gives

$$I_5=\left(\frac{2}{\pi}\right)^{1/2}\exp\left(\frac{-i\pi}{4}\right)\times\int_{|z|>k^{-1}}\exp[ik|z|+ik\theta\cdot(x-z)]\times(k|z|)^{-1/2}V(x-z)\,d^2z.$$

With z written in polar coordinates as $z=r\hat{\phi}$, I_5 becomes

$$I_5=ck^{-1/2}\exp\left(\frac{-i\pi}{4}+ik\hat{\theta}\cdot x\right)\int_{k^{-1}}^\infty r^{1/2}\times\exp(ikr)\int_{S^1}\exp(-ikr\cos\phi)V(x-r\hat{\phi})\,d\hat{\phi}\,dr, \tag{A6}$$

where the unit vectors are now adorned with hats and ϕ is the angle between the vectors $\hat{\phi}$ and $\hat{\theta}$. We can now apply the stationary phase approximation (Appendix D) to the angular integral:

$$\int_{S^1}\exp(-ikr\cos\phi)V(x-r\hat{\phi})\,d\hat{\phi}=M(kr)^{-1/2}(aV(x-r\hat{\theta})+bV(x+r\hat{\theta}))+R,$$

where

$$|R|\leqslant M(kr)^{-1}\max_{\hat{\phi}\in S^1}\{|V(x-\hat{\phi})|,|\nabla V(x-r\hat{\phi})|,|\Delta V(x-r\hat{\phi})|\}.$$

We note that over the range of integration in (A6), we have $(kr)^{-1}<1$. This allows us to combine the leading term and remainder term:

$$|I_5|\leqslant ck^{-1}\int_{k^{-1}}^\infty F(|r-|x+x_0||)\,dr\leqslant ck^{-1}\|V\|_{2,1}.$$

Next we consider I_6. We apply the Schwarz inequality

$$|I_6|\leqslant\left(\frac{c}{k}\int_{|x-y|>k^{-1}}\frac{|V(y)|}{|x-y|}d^2y\right)^{1/2}\|K(k)(|V|^{1/2}\psi)\|_2\leqslant ck^{-1/2}\left(\int\frac{|V(x-z)|}{|z|}d^2z\right)^{1/2}\|K(k)\|\,\|\xi\|_2\leqslant ck^{-1},$$

where we have used the estimate $\|K(k)\|\leqslant ck^{-1/2}$.

Finally we consider I_4. Application of the Schwarz inequality and use of information about the asymptotic behavior of $H_0^{(1)}$ gives

$$|I_4| \leqslant \left(\int_{|x-y|>k^{-1}} c(k|x-y|)^{-3} |V(y)|\, d^2y \right)^{1/2} \| \, |V|^{1/2}\psi \|_2. \tag{A7}$$

Since $k|x-y|>1$, some of the factors of $k|x-y|$ in the denominator can be replaced by 1; we also use inequality (A2) to estimate the second factor of (A7).

$$\begin{aligned} |I_4| &< \Big(ck^{-1-\epsilon/2} \\ &\quad \times \int_{S^1} \int_{k^{-1}} |V(x+r\phi)| r^{-1-\epsilon/2} r\, dr\, d\phi \Big)^{1/2} \frac{\|V\|_1}{1-a} \\ &< ck^{-(1+\epsilon/2)/2} \left(\int_{k^{-1}}^{\infty} F(|r-|x+x_0||) r^{-\epsilon/2}\, dr \right)^{1/2} \|V\|_1 \\ &< c\|V\|_1 k^{-(1+\epsilon/2)/2}. \end{aligned}$$

APPENDIX B: THE RECIPROCITY THEOREM

Proposition 3.1: Let V belong to $L^1 \cap L^2$. Then $A(k,\theta,\theta') = A(k,-\theta',-\theta)$.

Proof: We recall that the scattering amplitude is given by $A(k,\theta,\theta') = \int \overline{\exp(ik\theta\cdot x)} V(x) \psi(k,\theta',x)\, d^2x$. We now use the Lippman–Schwinger equation (1.1) to write the exponential in terms of the wave functions:

$$\begin{aligned} A(k,\theta,\theta') &= \int \overline{\psi^-(k,\theta,x)} V(x) \psi^+(k,\theta',x)\, d^2x \\ &\quad - \iint \overline{G^-(k,|x-y|)} V(y) \overline{\psi^-(k,\theta,y)}\, d^2y \\ &\quad \times V(x)\psi^+(k,\theta',x)\, d^2x. \end{aligned}$$

Next we use the symmetry properties mentioned at the beginning of Sec. 3:

$$\begin{aligned} A(k,\theta,\theta') &= \int \psi^+(k,-\theta,x) V(x) \psi^+(k,\theta',x)\, d^2x \\ &\quad - \int \psi^+(k,-\theta,y) V(y) \int G^+(k,|x-y|) \\ &\quad \times V(x)\psi^+(k,\theta',x)\, d^2x\, d^2y. \end{aligned}$$

Again we use the Lippmann–Schwinger equation to obtain an exponential: $A(k,\theta,\theta') = \int \psi^+(k,-\theta,x) V(x)$ $\times \exp(ik\theta'\cdot x)\, d^2x = A(k,-\theta',-\theta)$. The interchange of x and y integration in the third step is justified by absolute convergence of the iterated integral:

$$\begin{aligned} &\iint |G^-(k,|x-y|) V(y) \psi^-(k,\theta,y)|\, d^2y \\ &\quad \times |V(x)\psi^+(k,\theta',x)|\, d^2x \\ &< \int |(K\xi)(k,\theta,x)|\, |\xi(k,\theta',x)|\, d^2x \\ &< \|K\|\, \|\xi\|_2^2. \end{aligned}$$

QED

APPENDIX C: STRONG SQUARE INTEGRABILITY OF $S-I$

Proposition 5.2: Let $V \in W^{2,1}$, and suppose that for some x_0, $|V(x-x_0)|$, $|\nabla V(x-x_0)|$, and $|\Delta V(x-x_0)|$ are all bounded by a decreasing positive radial function $F(|x|)$ with

$$\int_0^{\infty} F(r) r\, dr < c\|V\|_{2,1}$$

and

$$F(r) < Mr^{-1+\epsilon} \quad \text{near zero,}$$

where $0<\epsilon<\frac{1}{2}$. Then

$$\int_{-\infty}^{\infty} (\|(S(k)-I)f\|_2^{(\theta)})^2\, dk < c(\|f\|_2^{(\theta)})^2. \tag{C1}$$

[The superscript θ reminds us that this is the $L^2(S^1)$ norm.]

Sketch of Proof (Details may be found in Cheney[a]): Let us fix $k_0 > \frac{1}{2}$, and split the left side of (C1) into small-k and large-k pieces.

The small-k piece is easy: the results of Sec. 2 show that $\|S(k)-I\|$ is bounded for $|k|<k_0$, which implies that

$$\int_{-k_0}^{k_0} \|(S(k)-I)f\|_2^2\, dk \leqslant c\|f\|_2^2.$$

Now we consider $|k|>k_0$. The difficulty we face is to extract from the integrand enough negative powers of k to make the k integral converge. In order to obtain explicit formulas, we write out the first few terms of the Born series:

$$\begin{aligned} |V|^{1/2}\psi &= (I-K)^{-1}(|V|^{1/2}\exp(ik\theta'\cdot x)) \\ &= (I+K+(I-K)^{-1}K^2)(|V|^{1/2}\exp(ik\theta'\cdot x)). \end{aligned} \tag{C2}$$

This allows us to write the kernel of $S(k)-I$ as

$$\begin{aligned} &(S(k)-I)(\theta',\theta) \\ &\quad = -i(4\pi)^{-1} \int \exp(-ik\theta\cdot x) V(x) \psi(k,\theta',x)\, d^2x \\ &\quad = D_1 + D_2 + D_3, \end{aligned} \tag{C3}$$

where

$$\begin{aligned} D_1(k,\theta',\theta) &= -i(4\pi)^{-1} \int \exp(-ik\theta\cdot x) V(x) \\ &\quad \times \exp(ik\theta'\cdot x)\, d^2x, \end{aligned} \tag{C4}$$

$$\begin{aligned} D_2(k,\theta',\theta) &= -i(4\pi)^{-1} \int \exp(-ik\theta\cdot x) V_{1/2}(x) \\ &\quad \times \frac{i}{4} \int |V(x)|^{1/2} \\ &\quad \times H_0(k|x-y|) V_{1/2}(y) \\ &\quad \times |V(y)|^{1/2} \exp(ik\theta'\cdot y)\, d^2y\, d^2x, \end{aligned} \tag{C5}$$

$$\begin{aligned} D_3(k,\theta',\theta) &= -i(4\pi)^{-1} \int \exp(-ik\theta\cdot x) V_{1/2}(x) \\ &\quad \times (I-K)^{-1}K^2 \\ &\quad \times (|V|^{1/2}\exp(ik\theta'\cdot x))\, d^2x. \end{aligned} \tag{C6}$$

Because we know from Sec. 1 that $\|K\|$ behaves like $|k|^{-1/2}$ for large k, it is fairly easy to see that

$$\int_{|k|>k_0} \|D_3 f\|_2^2\, dk \leqslant c\|f\|_2^2.$$

The terms corresponding to D_1 and D_2, however, require more work.

First we consider the part of the (C1) integral corresponding to D_1

$$16\pi^2\int_{|k|>k_0}\left|\left|\int_{S^1} D_1(k,\theta,\theta')f(\theta')\,d\theta'\right|\right|_2^2 dk$$
$$=\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int \exp(-ik(\theta'-\theta)\cdot x)$$
$$\times V(x)\,d^2x f(\theta')\,d\theta'\int_{S^1}\int \exp(ik(\theta''-\theta)\cdot y)$$
$$\times V(y)d^2y\,\overline{f(\theta'')}\,d\theta''\,d\theta\,dk. \tag{C7}$$

The absolute convergence of the θ integral allows us to do the θ integration first:

$$\int_{S^1}\exp(ik\theta\cdot(y-x))\,d\theta = 2\pi J_0(|k|\,|x-y|). \tag{C8}$$

Next we let $z=x-y$ in (C8) and use the asymptotic expansion for J_0 to split up (C7) into pieces corresponding to $|z|<|k|^{-1}$ and $|z|>|k|^{-1}$, respectively.

The piece corresponding to $|z|<|k|^{-1}$ is fairly easy because J_0 is bounded near the origin. We obtain the necessary k decay by using the inequality $1<|kz|^{-1}$ and by noting that the domain of z integration shrinks as k grows.

The piece of (C7) corresponding to $|z|>|k|^{-1}$ is harder to estimate. We shall consider in detail only the leading term of the J_0 asymptotic expansion; the remainder term already contains a factor of $(|kz|)^{-3/2}$ and is therefore easier to estimate. We write the leading term as

$$F=2\pi\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int_{|z|>|k|^{-1}}\int V(z+y)$$
$$\times 2^{1/2}(\pi|kz|)^{-1/2}\cos(|kz|-\tfrac{1}{4}\pi)$$
$$\times\exp(ik\theta'\cdot z)\,d^2z\,V(y)\exp(ik(\theta''-\theta')\cdot y)$$
$$\times d^2y f(\theta')\,d\theta'\,\overline{f(\theta'')}\,d\theta''\,dk. \tag{C9}$$

We let $z=r\hat{\phi}$ in the innermost integral of (C9); the z integral is then

$$\int_{|k|^{-1}}^{\infty}\int_{S^1} V(r\hat{\phi}+y)2^{1/2}(\pi|k|r)^{-1/2}$$
$$\times\cos(|k|r-\tfrac{1}{4}\pi)\exp(ik\cos\phi)\,d\hat{\phi}\,r\,dr,$$

where the unit vectors are now adorned with hats and ϕ is the angle between the vectors $\hat{\phi}$ and $\hat{\theta}'$.

Use of the stationary phase approximation (Lemma D.1) on the ϕ integral gives

$$\int_{S^1} V(r\hat{\phi}+y)\exp(ik\cos\phi)\,d\hat{\phi} = R+U(k,r,\hat{\theta}',y),$$

where

$$U(k,r,\hat{\theta}',y)=M_1(|k|r)^{-1/2}(aV(r\hat{\theta}'+y)+bV(-r\hat{\theta}'+y)) \tag{C10}$$

and

$$|R|\leqslant M_2(|k|r)^{-1}$$
$$\times\max_{\hat{\phi}\in S^1}\{|V(r\hat{\phi}+y)|,|\nabla V(r\hat{\phi}+y)|,|\Delta V(r\hat{\phi}+y)|\}. \tag{C11}$$

This application of the stationary phase approximation to (C9) gives

$$F=2\pi\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int_{|k|^{-1}}^{\infty} V(y)$$
$$\times\exp(ik(\theta''-\theta')\cdot y)2^{1/2}(\pi rk)^{-1/2}$$
$$\times\cos(|k|r-\tfrac{1}{4}\pi)(U(k,r,\theta',y)+R)$$
$$\times r\,dr\,d^2y f(\theta')\,d\theta'\,\overline{f(\theta'')}\,d\theta''\,dk, \tag{C12}$$

where we have once again dropped the hats on unit vectors. Next we split up the y integral in (C12): $F=F_1+F_2$, where

$$F_1=(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int_{|y|<|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty} V(y)$$
$$\times\exp(ik(\theta''-\theta')\cdot y)\cos(|k|r-\tfrac{1}{4}\pi)$$
$$\times(|k|r)^{-1/2}(U(k,r,\theta',y)+R)r\,dr\,d^2y f(\theta')$$
$$d\theta'\,\overline{f(\theta'')}\,d\theta''\,dk, \tag{C13}$$

$$F_2=(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S^1}\int_{|y|>|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty}$$
$$\text{(same integrand)}. \tag{C14}$$

To estimate F_1, we use the bounds (C10) and (C11) in (C13) to obtain decay of $|k|^{-1}$. We obtain additional decay by using the hypotheses on the potential and by noting that the domain of y integration shrinks as k grows.

Next we consider F_2 [Eq. (C14)]. We split F_2 into pieces corresponding to integration over different parts of S^1. We write $S^1=S_<\cup S_>$, where $S_<$ corresponds to $|\theta'-\theta''|<|k|^{-1+2\epsilon}$ and $S_>$ corresponds to $|\theta'-\theta''|>|k|^{-1+2\epsilon}$. Thus $F_2=C_1+C_2$, where

$$C_1=(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S_<}\int_{|y|>|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty} V(y)$$
$$\times\exp(ik(\theta''-\theta')\cdot y)$$
$$\times(|k|r)^{-1/2}\cos(|k|r-\tfrac{1}{4}\pi)(U(k,r,\theta',y)+R)$$
$$\times r\,dr\,d^2y f(\theta')\,d\theta'\,\overline{f(\theta'')}\,d\theta''\,dk, \tag{C15}$$

$$C_2=(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S_>}\int_{|y|>|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty}$$
$$\text{(same integrand)}. \tag{C16}$$

First we consider C_1: (C10) and (C11) applied to (C15) give us

$$|C_1|\leqslant(8\pi)^{1/2}\int_{|k|>k_0}\int_{S^1}\int_{S_<}\int_{|y|>|k|^{-\epsilon}}\int_{|k|^{-1}}^{\infty}(|k|r)^{-1/2}|V(y)|$$
$$\times|(2\pi)^{1/2}(|k|r)^{-1/2}F(|r\theta'+x_0+y|)$$
$$+(2\pi)^{1/2}(|k|r)^{-1/2}F(|-r\theta'+x_0+y|)+4M_2$$
$$\times(|k|r)^{-1}F(|r\phi_0+x_0+y|)|r\,dr\,d^2y|f(\theta')|$$
$$\times d\theta'\,|f(\theta'')|\,d\theta''\,dk. \tag{C17}$$

Over the range of integration $r>|k|^{-1}$, we can bound $(|k|r)^{-1}$ by $(|k|r)^{-1/2}$. We also use the fact that $|\pm r\theta'+x_0+y|$ and $|r\phi_0+x_0+y|$ can be bounded below by $||x_0+y|-r|$ to simplify (C17); we obtain

$$|C_1| < c \int_{|k|>k_0} |k|^{-1} \int_{S^1} \int_{S_<} \int |V(y)| \times \int_{|k|^{-1}}^{\infty} F(||x_0+y|-r|)\, dr\, d^2y \times |f(\theta')|\, d\theta' |f(\theta'')|\, d\theta''\, dk.$$

Carrying out the r and y integrations gives us

$$|C_1| \leqslant \int_{|k|>k_0} |k|^{-1} \int_{S^1} \int_{S_<} \|V\|_1 \times |f(\theta')| d\theta' |f(\theta'')|\, d\theta''\, dk.$$

We next apply the Schwarz inequality to the θ' integral, obtaining

$$|C_1| \leqslant c \int_{|k|>k_0} |k|^{-1} \int_{S^1} \|f\|_2 \left(\int_{S_<} d\theta' \right)^{1/2} \times |f(\theta'')|\, d\theta''\, dk. \tag{C18}$$

The θ' integral of (C18) is the measure of the angle subtending the chord of length $|k|^{-1+2\epsilon}$ between the unit vectors θ' and θ''. It is not hard to show that the measure of the angle also behaves like $|k|^{-1+2\epsilon}$ for large k. This gives us the additional k-decay we need in order to show $|C_1| \leqslant c\|f\|_2^2$.

We now turn our attention to C_2 [Eq. (C16)]. The right side of (C16) contains two pieces, one corresponding to U and the other to R. By (C11), the term corresponding to R already contains a factor of $(|k|r)^{-3/2}$; in order to make both the r integral and the k integral converge, we replace $(|k|r)^{-3/2}$ by $(|k|r)^{-1-\epsilon/2}$. This trick disposes of the remainder term, and we are left with the term corresponding to U. This term we write as

$$C_3 = (8\pi)^{1/2} \int_{|k|>k_0} \int_{S^1} f(\theta') \times \int_{S_>} f(\theta'') \int_{|y|>|k|^{-\epsilon}} \int_{|k|^{-1}}^{\infty} V(y) \times \exp(ik(\theta''-\theta')\cdot y) \times (|k|r)^{1/2} \cos(|k|r - \pi/4) M_1(|k|r)^{-1/2} \times [aV(-r\theta + y) + bV(r\theta' + y)] \times r\, dr\, d^2y\, d\theta'\, d\theta''\, dk. \tag{C19}$$

We note that the y integral of (C19) can be done first because the inner two integrals (r and y) converge absolutely. The y integral of (C19) can then be evaluated by letting $y = s\phi$ and applying the stationary phase approximation to the ϕ integral as follows. For notational convenience we define

$$W(s\phi, r\theta') = V(s\phi)[aV(-r\theta' + s\phi) + bV(r\theta' + s\phi)]$$

and

$$\tilde{\theta} = (\theta' - \theta'')/|\theta' - \theta''|.$$

Then the y integral is

$$\int_{|k|^{-\epsilon}}^{\infty} \int_{S^1} W(s\phi, r\theta') \exp(iks|\theta' - \theta''| \cos\phi)\, d\phi\, s\, ds. \tag{C20}$$

Application of the stationary phase approximation (Lemma D.1) to (C20) gives us

$$\int_{|k|^{-\epsilon}}^{\infty} \{M_1(|k|s|\theta' - \theta''|)^{1/2} \times [aW(s\tilde{\theta}, r\theta') + bW(-s\tilde{\theta}, r\theta')] + R'\} s\, ds,$$

where

$$|R'| \leqslant M_2(|k|s|\theta' - \theta''|)^{-1} \times \max_{\phi \in S^1} \{|W(s\phi, r\theta')|, |\nabla W(s\phi, r\theta')|, |\Delta W(s\phi, r\theta')|\}.$$

We use this in (C19) to obtain

$$|C_3| \leqslant c \int_{|k|>k_0} |k|^{-1} \int_{S^1} |f(\theta')| \times \int_{S_>} |f(\theta'')| \times \int_{|k|^{-\epsilon}}^{\infty} \int_{|k|^{-1}}^{\infty} \{M_1(|k|s|\theta' - \theta''|)^{-1/2} \times |aW(s\tilde{\theta}, r\theta') + bW(-s\tilde{\theta}, r\theta')| + |R'|\} \times dr\, s\, ds\, d\theta'\, d\theta''\, dk. \tag{C21}$$

In (C21) we use the assumptions on the potential

$$|W(s\tilde{\theta}, r\theta')| \leqslant F(|s - |x_0||) F(|\,|x_0| - s| - r|).$$

A similar bound holds for ∇W and ΔW. This allows us to estimate the right side of (C21) by

$$|C_3| \leqslant c \int_{|k|>k_0} |k|^{-1} \int_{S^1} |f(\theta'')| \times \int_{S_>} |f(\theta')| \int_{|k|^{-\epsilon}}^{\infty} \times \int_{|k|^{-1}}^{\infty} [(|k|s|\theta' - \theta''|)^{-1/2} + (|k|s|\theta' - \theta''|)^{-1}] \times F(|s - |x_0|\,|) F(|\,|x_0| - s| - r|) \times dr\, s\, ds\, d\theta'\, d\theta''\, dk. \tag{C22}$$

We now carry out the r integration and use the fact that over the range of integration, we have $|k|s|\theta' - \theta''| > |k|^{\epsilon}$. We can therefore bound the right side of (C22) by

$$|C_3| \leqslant c \int_{|k|>k_0} |k|^{-1} \int_{S^1} |f(\theta'')| \int_{S_>} |f(\theta')| \times \int_{|k|^{-\epsilon}}^{\infty} F(s - |x_0|) |k|^{-\epsilon/2} \times s\, ds\, d\theta'\, d\theta''\, dk \leqslant c\|f\|_2^2.$$

We have now shown that the right side of (C7) is bounded by $c\|f\|_2^2$; in other words, we have disposed of the D_1 term. Next we must consider the D_2 term.

We write out the piece of the (C1) integral corresponding to D_2 [(C5)]

$$16\pi^2 \int_{|k|>k_0} \left| \left| \int_{S^1} D_2(k, \theta, \theta') f(\theta')\, d\theta' \right| \right|_2^2 dk = \int_{|k|>k_0} \int_{S^1} \int_{S^1} \int \int \frac{i}{4} V(x) H_0(|k|\,|x-y|) V(y) \times \exp(-ik[\theta'\cdot y - \theta\cdot x])\, y\, d^2x\, f(\theta')\, d\theta' \times \int_{S^1} \int \int (-i/4) V(z) \overline{H_0(|k|\,|z-w|)} V(w) \times \exp(-ik[\theta\cdot z - \theta''\cdot w])\, d^2w\, d^2z\, f(\theta'')\, d\theta''\, d\theta\, dk. \tag{C23}$$

In the right side of (C23), we make the substitutions $y' = x - y$ and $w' = z - w$, and note that the θ integral is absolutely convergent. The θ integral can therefore be done first:

$$\int_{S^1} \exp(ik\theta\cdot(z-x))\, d\theta = 2\pi J_0(|k|\,|z-x|),$$

and so (C23) is

$$\begin{aligned}
&16\pi^2 \int_{|k|>k_0} \|D_2 f\|^2\, dk \\
&\quad = \frac{\pi}{8}\int_{|k|>k_0}\int_{S^1} f(\theta')\int\int\int_{S^1} f(\theta'') \\
&\quad \times \int\int J_0(|k(z-x)|)V(x)H_0(|ky'|)V(x-y') \\
&\quad \times \exp[-ik\theta'\cdot(x-y')]\, d^2y'\, d^2x\, d\theta'\, V(z) \\
&\quad \times \overline{H_0(|kw'|)}V(z-w')\exp[-ik\theta''\cdot(z-w')] \\
&\quad \times d^2w'\, d^2z\, d\theta''\, dk. \qquad \text{(C24)}
\end{aligned}$$

We shall obtain the sought-after k-decay from the spatial integrals. We therefore estimate the y' (or w') integral of (C24) first; we write

$$\int H_0(|ky'|)V(x-y')\, d^2y' = I_1 + I_2,$$

where I_1 and I_2 correspond to integration over the sets $|ky'| < 1$ and $|ky'| > 1$, respectively.

Use of the small-argument behavior of H_0 to estimate I_1 gives

$$|I_1| \leqslant c\int_{|ky'|<1} |\log|ky'|\, V(x-y')|\, d^2y'. \qquad \text{(C25)}$$

We apply Hölder's inequality to (C25), obtaining

$$\begin{aligned}
|I_1| &\leqslant c\left(\int_{|y|<|k|^{-1}} |\log|ky'|\,|^{(1+\epsilon)/\epsilon}\, d^2y\right)^{\epsilon/(1+\epsilon)} \\
&\quad \times \left(\int_{|y'|<|k|^{-1}} |V(x-y')|^{1+\epsilon}\, d^2y'\right)^{(1+\epsilon)^{-1}} \\
&\leqslant c\left(\int_0^{|k|^{-1}} F(|\,|x+x_0|-|y'|\,|)^{1+\epsilon}|y'|d\,|y'|\right)^{(1+\epsilon)^{-1}}.
\end{aligned} \qquad \text{(C26)}$$

To (C26) we apply Lemma D.2:

$$\begin{aligned}
|I_1| &\leqslant c\left(2|k|^{-1}\int_0^{|k|^{-1}} F(r)^{1+\epsilon}\, dr\right)^{(1+\epsilon)^{-1}} \\
&\leqslant c(|k|^{-1-\epsilon^2})^{(1+\epsilon)^{-1}}.
\end{aligned}$$

Use of the large-argument asymptotic behavior of H_0 to estimate I_2 shows

$$\begin{aligned}
|I_2| &\leqslant c\int_{|y'|>|k|^{-1}} |ky'|^{-1/2}|V(x-y')|\, d^2y' \\
&\leqslant c|k|^{-1/2}\int_{|k|^{-1}}^{\infty} F(|x-y+x_0|)|y|^{1/2}\, d\,|y| \\
&\leqslant c|k|^{-1/2}.
\end{aligned}$$

Thus the y' integral of (C24) can be estimated for large k by

$$\left|\int H_0(|ky'|)V(x-y')\, d^2y'\right| \leqslant c|k|^{-1/2}.$$

This shows that the right side of (C24) is bounded by

$$\begin{aligned}
&c\int_{|k|>k_0} |k|^{-1}\|f\|_1^2 \\
&\quad \times \int\int |J_0(|k(z-x)|)|\,|V(x)|\,|V(z)|\, d^2x\, d^2z\, dk. \qquad \text{(C27)}
\end{aligned}$$

It remains to do the x and z integrals of (C27); to do this end, we let $z' = z - x$, and split the z' integral into pieces corresponding to integration over $|kz'| < 1$ and $|kz'| > 1$, respectively. We obtain extra k-decay in the small-argument piece because the domain of integration shrinks as $k \to \infty$. Decay is obtained in the large-argument integral from the $|kz'|^{-1/2}$ behavior of J_0 at infinity. QED

APPENDIX D: TECHNICAL LEMMAS

Lemma D.1 (Stationary phase approximation): Let $Q \in C^2(R^2)$. Then

$$\begin{aligned}
&\int_{S^1} Q(r\hat{\phi})\exp(ikr\cos\phi)\, d\phi \\
&\quad = M_1(|k|\,|r|)^{-1/2}(aQ(r\hat{\phi})\,|_{\phi=0} + bQ(r\hat{\phi})\,|_{\phi=\pi}) + R, \qquad \text{(D1)}
\end{aligned}$$

where

$$|R| \leqslant M_2(|k|\,|r|)^{-1}\max_{\hat{\phi}\in S^1}\{|Q(r\hat{\phi})|, |\nabla Q(r\hat{\phi})|, |\Delta Q(r\hat{\phi})|\}; \qquad \text{(D2)}$$

here $\hat{\phi} = (\cos\phi, \sin\phi)$, a and b are constants of modulus 1 and the M_i are positive constants independent of Q.

Proof: The proof follows Erdelyi.[19] Our first task is to split up the integral

$$I = \int_{S^1} Q(r\hat{\phi})\exp(ikr\cos\phi)\, d\phi \qquad \text{(D3)}$$

so that we consider only one stationary point at a time. To this end, we write $I = I_1 + I_2$, where I_1 is the integral over $[0,\pi]$, I_2 the integral over $[\pi,2\pi]$. First we consider I_1, which we split into $I_1 = A + B$, where

$$A = \int_0^{\pi} Q(r\hat{\phi})\exp(ikr\cos\phi)\eta(\phi)\, d\phi, \qquad \text{(D4)}$$

$$B = \int_0^{\pi} Q(r\hat{\phi})\exp(ikr\cos\phi)[1-\eta(\phi)]\, d\phi, \qquad \text{(D5)}$$

and where η is an infinitely differentiable cutoff function with

$$\begin{aligned}
\eta(\phi) &= 1 \quad \text{for } 0 \leqslant \phi \leqslant \pi/4, \\
&= 0 \quad \text{for } 3\pi/4 \leqslant \phi \leqslant \pi.
\end{aligned}$$

We consider A first. In (D4) we make the change of variable $t^2 = 1 - \cos\phi$;

$$\begin{aligned}
A &= \exp(ikr)\int_0^{2^{1/2}} Q(r\hat{\phi}) \\
&\quad \times \exp(-ikrt^2)\tilde{\eta}(t)2t\,[1-(1-t^2)^2]^{-1/2}\, dt, \qquad \text{(D6)}
\end{aligned}$$

where $\tilde{\eta}(t) = -\eta(\arccos(1-t^2))$. Integration by parts of (D6) [differentiation of $2Q(r\hat{\phi})\tilde{\eta}(t)(2-t^2)^{-1/2}$ and integration of $\exp(-ikrt^2)$] gives

$$\begin{aligned}
A &= (ikr)\Big[2Q(r\hat{\phi})\tilde{\eta}(t)h_1(t)(2-t^2)^{-1/2}|_0^{2^{1/2}} \\
&\quad - \int_0^{2^{1/2}} \frac{\partial}{\partial t}(2Q(r\hat{\phi})\tilde{\eta}(t)(2-t^2)^{-1/2})h_1(t)\, dt\Big], \qquad \text{(D7)}
\end{aligned}$$

where

$$h_1(t) = -\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right) \times \int_0^\infty \exp\left[-ikr\left(t+\sigma\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right)\right)^2\right] d\sigma.$$

To compute the first term of (D7), we need to evaluate h_1 at zero:

$$h_1(0) = -\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right)\int_0^\infty \exp[-|k|r\sigma^2]\,d\sigma = -\pi^{1/2}(|k|r)^{-1/2}\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right).$$

We substitute this expression into (D7) and recall that $\tilde\eta(2^{1/2}) = 0$. Equation (D7) is then

$$A = (2\pi)^{1/2}(|k|r)^{-1/2}\exp(ikr) \times \exp(-i\,\mathrm{sgn}\,k\,\pi/4)Q(r\hat\phi)|_{\phi=0} + R_1,$$

where

$$R_1 = -\exp(ikr) \times \int_0^{2^{1/2}} \frac{\partial}{\partial t}[2Q(r\hat\phi)\tilde\eta(t)(2-t^2)^{-1/2}]h_1(t)\,dt.$$

We have now found the leading term of (D1); our next task is to obtain the correct decay for the remainder. To this end, we integrate R_1 by parts; this gives us

$$R_1 = -\exp(ikr) \times \frac{\partial}{\partial t}[2Q(r\hat\phi)\tilde\eta(t)(2-t^2)^{-1/2}]h_2(t)|_0^{2^{1/2}} + R_2, \qquad \text{(D8)}$$

where

$$R_2 = \exp(ikr) \times \int_0^{2^{1/2}} \frac{\partial^2}{\partial t^2}[2Q(r\hat\phi)\tilde\eta(t)(2-t^2)^{-1/2}]h_2(t)\,dt \qquad \text{(D9)}$$

and where h_2, given by

$$h_2(t) = -\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right) \times \int_0^\infty \sigma\exp\left[-ikr\left(t+\sigma\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right)\right)^2\right] d\sigma, \qquad \text{(D10)}$$

is the primitive of h_1 satisfying

$$h_2(0) = -\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right)\int_0^\infty \sigma\exp(-|k|r\sigma^2)\,d\sigma = -(|k|r)^{-1}\exp\left(-i\,\mathrm{sgn}\,k\,\frac{\pi}{4}\right). \qquad \text{(D11)}$$

To estimate $h_2(t)$ for $t>0$, we note that along the path of integration, the quantity

$$\begin{aligned} &-ikr(t+\sigma\exp(-i\,\mathrm{sgn}\,k\,\pi/4))^2 + |k|r\sigma^2 \\ &\quad = -ikr[t^2 + 2t\sigma\exp(-i\,\mathrm{sgn}\,k\,\pi/4) \\ &\qquad + (\mathrm{sgn}\,k)i\sigma^2 - i\,\mathrm{sgn}\,k\,\sigma^2] \\ &\quad = -ikrt\,[t+2\sigma\exp(-i\,\mathrm{sgn}\,k\,\pi/4)] \end{aligned}$$

has negative real part; thus

$$\exp[-ikr(t+\exp(-i\,\mathrm{sgn}\,k\,\pi/4)\sigma)^2] \leqslant \exp(-|k|r\sigma^2).$$

With this estimate, we have

$$|h_2(t)| < \int_0^\infty \sigma\exp(-|k|r\sigma^2)\,d\sigma = (|k|r)^{-1}.$$

With this information, a bound on R_1 can be obtained as follows. We write $\mu(t) = \tilde\eta(t)(2-t^2)^{-1/2}$. Then we compute the derivatives appearing in (D8) and (D9) [$' = (d/dt)$]:

$$\frac{\partial}{\partial t}[Q(r\hat\phi)\mu(t)] = Q(r\hat\phi)\mu'(t) + \nabla Q(r\hat\phi)\cdot\hat\phi\,'\mu(t)$$

and

$$\begin{aligned} &\frac{\partial^2}{\partial t^2}[Q(r\hat\phi)\mu(t)] \\ &\quad = Q(r\hat\phi)\mu''(t) + 2\nabla Q(r\hat\phi)\cdot\hat\phi\,'\mu'(t) \\ &\qquad + \nabla Q(r\hat\phi)\cdot\hat\phi\,''\mu(t) + \nabla Q(r\hat\phi)\|\hat\phi\,'\|^2\mu(t). \end{aligned}$$

Let

$$M_2 = 6\max_{0<t<(1+2^{-1/2})^{1/2}}\{|\mu'(t)|,|\mu''(t)|,|\hat\phi\,'\mu(t)|,|2\hat\phi\,'\mu'(t)|, |\hat\phi\,''\mu(t)|\ \|\hat\phi\,'\|^2|\mu(t)|\}.$$

Then

$$|R_1| \leqslant M_2(|k|r)^{-1} \times \max_{0<t<(1+2^{-1/2})^{1/2}}\{|Q(r\hat\phi)|,|\Delta Q(r\hat\phi)|,|\Delta Q(r\hat\phi)|\}.$$

This concludes the estimate for A; now for B, the change of variables $t^2 = \cos\phi + 1$ gives a similar estimate; and in I_2 the change of variables $\beta = \phi - \pi$ converts I_2 to an integral of the form I_1.

Lemma D.2: Let $F(r)$ be a positive nonincreasing function on $[0,b]$, $b>0$. Then for $a>0$ and $\alpha>0$,

$$\int_0^b F(|a-r|)r^\alpha\,dr \leqslant 2b^\alpha\int_0^b F(r)\,dr. \qquad \text{(D12)}$$

Proof: In the left side of (D12) we note that $r \leqslant b$, and then we use the definition of absolute value

$$\begin{aligned} I &= \int_0^b F(|a-r|)r^\alpha\,dr \leqslant b^\alpha\int_0^b F(|a-r|)\,dr \\ &= b^\alpha\int_0^{\min(a,b)} F(a-r)\,dr + b^\alpha\int_{\min(a,b)}^b F(r-a)\,dr. \end{aligned}$$

In the first integral let $s = a - r$; in the second let $s = r - a$. Then

$$I \leqslant b^\alpha\int_{a-\min(a,b)}^a F(s)\,ds + b^\alpha\int_{\min(a,b)-a}^{b-a} F(s)\,ds.$$

Case $a<b$:

$$I \leqslant b^\alpha\int_0^a F(s)\,ds + b^\alpha\int_0^{b-a} F(s)\,ds \leqslant 2b^\alpha\int_0^b F(s)\,ds.$$

Case $b<a$:

$$I \leqslant b^\alpha\int_{a-b}^a F(s)\,ds \leqslant b^\alpha\int_0^b F(s)\,ds.$$

[1] I. Kay and H. E. Moses, Nuovo Cimento **22**, 689 (1961).
[2] I. Kay and H. E. Moses, Comm. Pure Appl. Math. **14**, 435 (1961).
[3] L. D. Faddeev, Itogi Nauk Tekh. Sov. Probl. Mat. **3**, 93 (1974) [J. Sov. Math. **5**, 334 (1976)].

[4]R. G. Newton, *Scattering Theory in Mathematical Physics*, edited by J. A. Lavita and J.-P. Marchand (Reidel, Dordrecht, 1974).
[5]L. D. Faddeev, Dokl. Akad. Nauk SSSR **165**, 514 (1965) [Sov. Phys. Dokl. **10**, 1033 (1966)].
[6]L. D. Faddeev, Dokl. Akad. Nauk SSSR **167**, 69 (1966) [Sov. Phys. Dokl. **11**, 209 (1966)].
[7]R. T. Prosser, J. Math. Phys. **10**, 1819 (1969).
[8]R. T. Prosser, J. Math. Phys. **17**, 1775 (1976).
[9]R. T. Prosser, J. Math. Phys. **21**, 2635 (1980).
[10]C. Morawetz, Comp. Math. Appls. **7**, 319 (1981).
[11]P. Deift and E. Trubowitz, Comm. Pure. Appl. Math. **32**, 121 (1979).
[12]R. G. Newton, J. Math. Phys. **21**, 1698 (1980).
[13]R. G. Newton, J. Math. Phys. **22**, 2191 (1981).
[14]R. G. Newton, J. Math. Phys. **23**, 594 (1982).
[15]S. Agmon, Ann. Scuola Norm. Sup. Pisa, Ser. IV, **2**, 151 (1975).
[16]M. Cheney, "Two-dimensional scattering: the number of bound states from scattering data," J. Math. Phys. (in press).
[17]R. G. Newton, *Scattering Theory of Waves and Particles*, (Springer, New York, 1982), 2nd ed., p. 286.
[18]M. Reed and B. Simon, *Methods of Modern Mathematical Physics. I. Functional Analysis* (Academic, New York, 1972), p. 237.
[19]A. Erdelyi, *Asymptotic Expansions* (Dover, New York, 1965).

Eigenvalues and eigenfunctions associated with the Gel'fand–Levitan equation

Harry E. Moses[a)]
Center for Atmospheric Research, University of Lowell, Lowell, Massachusetts 01854

Reese T. Prosser
Department of Mathematics, Dartmouth College, Hanover, New Hampshire 03755

(Received 8 June 1983; accepted for publication 10 August 1983)

It is shown here that the solutions of the Gel'fand–Levitan equation for inverse potential scattering on the line may be expressed in terms of the eigenvalues and eigenfunctions of certain associated operators of trace class. The details are sketched for the case of rational reflection coefficients, and carried out for the simplest class of examples.

PACS numbers: 03.80. + r, 03.65.Nk

1. INTRODUCTION

The Gel'fand–Levitan equation plays a central role in solving inverse scattering problems in one dimension.[1] In the case where the problem involves a *scattering potential* $V(x)$ defined for $-\infty < x < +\infty$, for example, we know that $V(x)$ may be recovered from the *reflection coefficient* $r(k)$, defined for $-\infty < k < +\infty$, as follows: set

$$R(x,y) = \hat{r}(x+y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ikx} r(k) e^{-iky}\, dk, \quad (1)$$

and then solve for $K(x,y)$ the Gel'fand–Levitan equation

$$K(x,y) + R(x,y) + \int_{-\infty}^{x} K(x,z)R(z,y)dz = 0. \quad (2)$$

Then the potential $V(x)$ appears as

$$V(x) = 2\frac{d}{dx}K(x,x). \quad (3)$$

(See Ref. 2 for a general discussion of this procedure.)

In order to study the behavior of the solutions of (2), it is useful to consider the associated equation, to be solved for $K(x,y,w)$:

$$K(x,y,w) + R(x,y) + \int_{-\infty}^{w} K(x,z,w)R(z,y)dz = 0. \quad (4)$$

Evidently $K(x,y,x) = K(x,y)$. Now (4) may be expressed in operator form with w as a parameter:

$$K(w) + R + K(w)P(w)R = 0. \quad (5)$$

Here R, $K(w)$, and $P(w)$ are integral operators with kernels $R(x,y)$, $K(x,y,w)$, and $P(x,y,w)$, with

$$P(x,y,w) = \theta(w-x)\delta(x-y). \quad (6)$$

Here $\theta(z)$ is the Heaviside function, and $\delta(z)$ its derivative.

Now (4) yields

$$K(w)(I + P(w)R) = -R, \quad (7)$$

and hence, whenever $(I + P(w)R)$ is invertible,

$$K(w) = -R(I + P(w)R)^{-1}. \quad (8)$$

Now suppose that the reflection coefficient $r(k)$ is such that its Fourier transform $\hat{r}(z)$ is smooth and integrable. Then it follows that the operator $P(w)R$ is of *trace class* for each w, and

$$\operatorname{tr} P(w)R = \int_{-\infty}^{w} \hat{r}(2z)dz. \quad (9)$$

One can then define the *Fredholm determinant* $\Delta(w)$ of the operator $(I + P(w)R)$ by (cf. Ref. 3, p. 255ff)

$$\Delta(w) = \det(I + P(w)R) = \exp \operatorname{tr} \log(I + P(w)R). \quad (10)$$

Evidently

$$\log \Delta(w) = \operatorname{tr} \log(I + P(w)R) \quad (11)$$

and so

$$\frac{d}{dw}\log \Delta(w) = \frac{\Delta'(w)}{\Delta(w)} = \operatorname{tr} P'(w)R(I + P(w)R)^{-1} = -\operatorname{tr} P'(w)K(w). \quad (12)$$

Here we have used (8). But $P'(w)K(w)$ has kernel $\delta(w-x)K(x,y,w)$, so

$$-\operatorname{tr} P'(w)K(w) = -\int_{-\infty}^{w} \delta(w-x)K(x,x,w)dx = -K(w,w,w) = -K(w,w). \quad (13)$$

Hence by (3)

$$V(w) = 2\frac{d}{dw}K(w,w) = -2\frac{d^2}{dw^2}\log \Delta(w). \quad (14)$$

This formula, which gives V directly in terms of R, first appears in Ref. 4, and has since been rediscovered by several authors, including us.[5] In one sense, this formula by-passes the Gel'fand–Levitan equation, since it gives V directly in terms of R, and once V is known everything about the scattering problem is known, at least in principle.

In another sense (14) is no better than (4), since the calculation of the determinant $\Delta(w)$ of $(I + P(w)R)$ is not usually an easy matter in practice. One possible approach is to calcu-

a)Research Sponsored in part by AFOSR Grant No. 81-0253A.

 0022-2488/84/010108-05$02.50

late the *eigenvalues* $\lambda_n(w)$ of the operator $P(w)R$ and use them to calculate $\Delta(w)$:

$$\Delta(w) = \prod_{n=1}^{\infty} (1 + \lambda_n(w)). \tag{15}$$

We indicate here how this might be done in the case where the reflection coefficient $r(k)$ is a rational function of k. (This case has already been treated by other methods in Refs. 6 and 7.)

Accordingly, we assume now that $r(k)$ has the form

$$r(k) = p(-ik)/q(-ik), \tag{16}$$

where p and q are polynomials with real coefficients, chosen so that degree $p <$ degree q, and so that $r(k)$ is regular in the upper half k-plane. If $r(k)$ is to be a reflection coefficient, then we should require that $|r(k)| \leqslant 1$ and $r(0) = -1$, but these requirements will play no role in solving (4).

It follows from our assumptions that $R(x, y) = \hat{r}(x + y)$ vanishes if $x + y < 0$, and satisfies an ordinary differential equation if $x + y > 0$, of the form

$$q(D)R(x, y) = p(D)\delta(x + y), \tag{17}$$

where $D = d/dx$.

The eigenvalues $\lambda_n(w)$ of the trace-class operator $P(w)R$ are discrete and the corresponding eigenfunctions $\phi_n(w)$ satisfy

$$P(w)R\phi_n(w) = \lambda_n(w)\phi_n(w). \tag{18}$$

It follows that $\phi_n(w) = P(w)\phi_n(w)$ and, hence, that

$$R(w)\phi_n(w) \equiv P(w)RP(w)\phi_n(w) = \lambda_n(w)\phi_n(w). \tag{19}$$

Moreover, it is easy to verify from (1) that if $|r(k)| \leqslant M$, then the operator R^2 is positive and satisfies $0 \leqslant R^2 \leqslant M^2 I$. It follows that the same is true of $R(w)^2$. Hence we have

$$0 \leqslant \lambda_n^2(w) \leqslant M^2. \tag{20}$$

Since $R(w)$ is of trace class, we also have, after a suitable rearrangement,

$$M^2 \geqslant \lambda_1^2 \geqslant \lambda_2^2 \geqslant \cdots \geqslant \lambda_n^2 \downarrow 0, \tag{21}$$

$$\sum_{n=1}^{\infty} \lambda_n(w) = \operatorname{tr}(P(w)RP(w)) = \operatorname{tr}(R(w)), \tag{22}$$

$$\prod_{n=1}^{\infty} (1 + \lambda_n(w)) = \Delta(w). \tag{23}$$

Now Eq. (19) may be written, for $-\infty < x \leqslant w$,

$$\int_{-\infty}^{w} R(x + y)\phi_n(y, w)dy = \lambda_n(w)\phi_n(x, w). \tag{24}$$

Applying (17) to (24), we get, for $-w \leqslant x \leqslant w$,

$$\lambda_n(w)q(D)\phi_n(x, w) = \int_{-\infty}^{w} q(D)R(x + y)\phi_n(y, w)dy$$
$$= p(D)\phi_n(-x, w). \tag{25}$$

It follows that, for $-w \leqslant x \leqslant w$,

$$\lambda_n^2(w)q(-D)q(D)\phi_n(x, w) = \lambda_n(w)p(D)q(-D)\phi_n(-x, w)$$
$$= p(D)p(-D)\phi_n(x, w). \tag{26}$$

Thus we see that the eigenfunctions $\phi_n(x, w)$ of the operator $R(w) = P(w)RP(w)$ satisfy an ordinary differential equation of even order with constant coefficients. By inserting the known form of the solutions of this equation back into (24), we may determine the integration constants and the admissible values of $\lambda_n(w)$. Specifically, the solutions of (26) all have the form

$$\phi_n(x, w) = \sum_{j=1}^{m} (A_j e^{ik_j x} + B_j e^{-ik_j x}), \tag{27}$$

where the $\pm k_j$ are the $2m$ solutions of the equation

$$r(-k)r(k) = \lambda_n^2(w). \tag{28}$$

Here m is the degree of the polynomial q. Note that if k_j is a solution of this equation, then so is $-k_j$, and so is $\bar{k}_j$. We assume here that these solutions are all distinct, and that $\operatorname{Im}(+k_j) \geqslant 0$.

Now if we insert (27) back into (24), do the integration, and equate coefficients of the various resulting exponentials, we get $2m - 1$ equations relating the A_j and B_j, and one equation determining the admissible values of $\lambda_n(w)$ for given w. Details are presented in the next section.

Once the eigenvalues $\lambda_n(w)$ and eigenfunctions $\phi_n(x, w)$ of the operator $R(w) = P(w)RP(w)$ are known, then we can calculate the determinant $\Delta(w)$ by (15). Moreover, we can also calculate the kernel of $K(w)$, since if the eigenfunctions $\phi_n(w)$ are normalized by

$$\|\phi_n(w)\|_2 = 1, \tag{29}$$

then we have, for $-\infty < x, y \leqslant w$,

$$R(x, y, w) = \sum_{n=1}^{\infty} \lambda_n(w)\phi_n(x, w)\phi_n(y, w), \tag{30}$$

and so, by (8), for $-\infty < x, y \leqslant w$,

$$K(x, y, w) = \sum_{n=1}^{\infty} \frac{-\lambda_n(w)}{1 + \lambda_n(w)} \phi_n(x, w)\phi_n(y, w) \tag{31}$$

and

$$K(w, w, w) = \sum_{n=1}^{\infty} \frac{-\lambda_n(w)}{1 + \lambda_n(w)} \phi_n(w, w)^2. \tag{32}$$

But from (12) and (13) we have

$$K(w, w, w) = -\frac{d}{dw} \log \Delta(w)$$
$$= -\sum_{n=1}^{\infty} \frac{\lambda_n'(w)}{1 + \lambda_n(w)}. \tag{33}$$

Comparing (32) and (33), we see that when $x = w$, we have

$$\phi_n(w, w)^2 = \lambda_n'(w)/\lambda_n(w). \tag{34}$$

On the other hand, since $R(x + y) = 0$ if $x + y < 0$, we see from (24) that, when $x = -w$, we have

$$\phi_n(-w, w) = 0. \tag{35}$$

Thus we see that the eigenfunction $\phi_n(x, w)$ of $R(w)$ vanishes unless $|x| \leqslant w$, and then is a real exponential polynomial which vanishes at $x = -w$ and takes the value $(\lambda_n'(w)/\lambda_n(w))^{1/2}$ at $x = +w$.

It is not clear to us yet what role these eigenvalues and eigenfunctions may play in a further study of the Gel'fand–Levitan equation, nor what physical significance, if any, may be attached to them. We note here only that Eq. (4) admits an iterative solution

$$K(w) = -R(w) + R(w)^2 - R(w)^3 + \cdots \tag{36}$$

which converges in operator norm, according to the Fredholm theory, if and only if the eigenvalues $\lambda_n(w)$ of $R(w)$ all satisfy

$$|\lambda_n(w)| < 1. \tag{37}$$

This condition provides a natural obstacle to the convergence of any iterative procedure. In the physically interesting case $|r(k)| < 1$, and so (20) implies (37). We conclude that in this case the iterative solution (36) actually converges geometrically in operator norm to the operator $K(w)$.

It may also be possible to develop effective approximate solutions to the Gel'fand–Levitan equation by using a finite number of the eigenvalues and eigenfunctions as normal modes, to be computed numerically, e.g., by a suitable variational principle.

2. CALCULATIONS

Now we assume that $r(k)$ is rational, of the form (16), and rewrite it as

$$r(k) = \sum_{i=1}^{m} \frac{a_i}{k - b_i}. \tag{38}$$

Here a_i and b_i are complex constants, with $\operatorname{Im} b_i < 0$. We assume that the b_i are all distinct. It follows from (1) that

$$R(x+y) = \theta(x+y) \sum_{i=1}^{m} (-ia_i) e^{-ib_i(x+y)}. \tag{39}$$

We now insert the forms (27) and (39) into Eq. (21) and equate the coefficients of the exponential terms $e^{\pm ik_j x}$. In this way we find

$$\left(\sum_{i=1}^{m} \frac{a_i}{k_j - b_i}\right) A_j = r(k_j) A_j = -\lambda B_j, \tag{40}$$

$$\left(\sum_{i=1}^{m} \frac{a_i}{-k_j - b_i}\right) B_j = r(-k_j) B_j = -\lambda A_j, \tag{41}$$

$$\sum_{j=1}^{m} \left(\frac{A_j}{i(k_j - b_i)} e^{i(k_j - b_i)w} - \frac{B_j}{i(k_j + b_i)} e^{-i(k_j + b_i)w} \right) = 0. \tag{42}$$

To satisfy (40) and (41), we put

$$s(k_j) = (r(-k_j))^{1/2}, \tag{43}$$

$$t(k_j) = (r(+k_j))^{1/2}, \tag{44}$$

where the square roots are chosen so that $s(k_j)t(k_j) = -\lambda$. Then we put

$$A_j = s(k_j) C_j, \tag{45}$$

$$B_j = t(k_j) C_j, \tag{46}$$

with C_j to be determined, and note that (40) and (41) are satisfied for any choice of C_j.

Now (42) takes the form

$$\sum_{j=1}^{m} A_{ij} C_j = 0, \tag{47}$$

where the matrix A_{ij} is given by

$$A_{ij} = A_{ij}(w, \lambda) = \frac{s(k_j)}{i(k_j - b_i)} e^{i(k_j - b_i)w} - \frac{t(k_j)}{i(k_j + b_i)} e^{-i(k_j + b_i)w} \tag{48}$$

Note that the k_j, and hence the A_{ij}, depend on λ. Equation (47), and hence (42), admits a nontrivial solution if and only if

$$\det(A_{ij}(w, \lambda)) = 0. \tag{49}$$

This is the case only for certain values λ_n of λ; these values λ_n are then the eigenvalues, and the corresponding functions ϕ_n are the eigenfunctions of (24). In this way the eigenvalue problem for (24) reduces to the problem of solving (49).

It is instructive to apply this same procedure to obtain a solution $K(x, y, w)$ for the integral equation (4). An argument similar to that leading to (26) shows that if $y < x$, then $K(x, y, w)$ satisfies a differential equation in y of the form

$$q(-D) q(D) K(x, y, w) = p(D) p(-D) K(x, y, w). \tag{50}$$

Here $D = \partial/\partial y$. Hence $K(x, y, w)$ must have the form

$$K(x, y, w) = \sum_{j=1}^{m} A_j(x, w) e^{ik_j y} + B_j(x, w) e^{-ik_j y}, \tag{51}$$

where the k_j are now solutions of

$$r(-k) r(k) = 1. \tag{52}$$

This is just (28) with $\lambda^2 = 1$. If we insert (50) and (39) into (4) and equate coefficients of $e^{\pm ik_j y}$, we find

$$\left(\sum_{i=1}^{m} \frac{a_i}{k_j - b_i}\right) A_j = r(k_j) A_j = B_j, \tag{53}$$

$$\left(\sum_{i=1}^{m} \frac{a_i}{-k_j - b_i}\right) B_j = r(-k_j) B_j = A_j, \tag{54}$$

$$\sum_{j=1}^{m} \left(\frac{A_j}{i(k_j - b_i)} e^{i(k_j - b_i)w} - \frac{B_j}{i(k_j + b_i)} e^{-i(k_j + b_i)w} \right) = -e^{-ib_i x}. \tag{55}$$

Note that (53) and (54) are just (40) and (41) with $\lambda = -1$, and (42) is the homogeneous form of (55). Hence, with the choices (45) and (46) (with $\lambda = -1$) for A_j and B_j, we know that (53) and (54) are satisfied, and (55) becomes

$$\sum_{j=1}^{m} A_{ij} C_j = -e^{-ib_i x} \tag{56}$$

with the matrix $A_{ij} = A_{ij}(w, -1)$ given again by (48), with $\lambda = -1$. When $|r(k)| < 1$, we know [cf. (20)] that $\lambda = -1$ cannot be an eigenvalue of $R(w)$, and hence that $\det A_{ij}(w, -1)$ cannot vanish. Hence $A_{ij}(w, -1)$ must be invertible. Setting

$$B_{jk}(w) = (A^{-1}(w, 1))_{jk}, \tag{57}$$

we have

$$C_j(x, w) = \sum_{k=1}^{m} B_{jk}(w) e^{-ib_k x}, \tag{58}$$

and so

$$K(x, y, w) = -\sum_{j,k=1}^{m} B_{jk}(w) e^{-ib_k x} (s(k_j) e^{ik_j y} + t(k_j) e^{-ik_j y}). \tag{59}$$

Since

$$s(k_j) e^{ik_j y} + t(k_j) e^{-ik_j y} = A'_{kj}(y) e^{ib_k y}, \tag{60}$$

where $A'_{kj}(y) = dA_{kj}(y, -1)/dy$, we may rewrite (59) as

$$K(x, y, w) = -\sum_{j,k=1}^{m} B_{jk}(w) A'_{kj}(y) e^{-ib_k(x-y)}. \tag{61}$$

In particular, when $x = y = w$, (61) becomes

$$K(w,w,w) = -\frac{d}{dw}\mathrm{tr}\log A(w,-1). \tag{62}$$

Comparing (62) with (33), we see that

$$\Delta(w) = \mathrm{const}\times\det A(w,-1). \tag{63}$$

The constant in (63) need not be 1, as our example in the next section shows, but it plays no role in determining $K(w,w)$ or $V(w)$.

3. EXAMPLES

Here we work through the simplest class of examples. We assume that $m = 1$ in (38) and set $a_1 = i\alpha$, $b_1 = -i\beta$, so that

$$r(k) = \frac{i\alpha}{k+i\beta} = \frac{\alpha}{\beta - ik}, \tag{64}$$

with α,β real constants, $\alpha,\beta > 0$. (The potentials for these reflection coefficients have been obtained using Gel'fand–Levitan methods for $-\beta \leqslant \alpha < \beta$ in Refs. 8 and 9 and for $\alpha = \beta$ in Ref. 10. Note that the case $\alpha = \beta$ is a pathological case in which two distinct potentials can be found which have the same reflection coefficient.[10]) Then we have from (1)

$$R(x+y) = \theta(x+y)\alpha e^{-\beta(x+y)}, \tag{65}$$

and the eigenvalue equation (24) becomes

$$\alpha\int_{-x}^{w} e^{-\beta(x+y)}\phi(y,w)dy = \lambda\phi(x,w). \tag{66}$$

One may verify that (26) holds:

$$\begin{aligned}\lambda^2 q(-D)q(+D)\phi(x,w) &= \lambda^2(\beta^2 - D^2)\phi(x,w)\\ &= p(D)p(-D)\phi(x,w) = \alpha^2\phi(x,w),\end{aligned} \tag{67}$$

from which it follows that $\phi(x,w)$ must have the form (setting $k_1 = \kappa$)

$$\phi(x,w) = Ae^{i\kappa x} + Be^{-i\kappa x}, \tag{68}$$

with $\pm\kappa$ chosen so that

$$r(\kappa)r(-\kappa) = \alpha^2/\kappa^2 + \beta^2 = \lambda^2. \tag{69}$$

We assume first that $\lambda^2 < \alpha^2/\beta^2$, so that $\pm\kappa$ are real. Inserting (68) into (66), integrating, and equating the coefficients of $e^{\pm i\kappa x}$, we find

$$(i\alpha/(\kappa + i\beta))A = -\lambda B, \tag{70}$$

$$(i\alpha/(-\kappa + i\beta))B = -\lambda A. \tag{71}$$

Setting

$$s(\kappa) = (i\alpha/(-\kappa + i\beta))^{1/2}, \tag{72}$$

$$t(\kappa) = (i\alpha/(\kappa + i\beta))^{1/2}, \tag{73}$$

$$A = s(\kappa)C, \tag{74}$$

$$B = t(\kappa)C, \tag{75}$$

we get

$$\phi(x,w) = C(s(\kappa)e^{i\kappa x} + t(\kappa)e^{-i\kappa x}). \tag{76}$$

The matrix $A_{ij}(w,\lambda)$ in this case reduces to a single entry

$$A_{11}(w,\lambda) = \frac{s(\kappa)e^{(i\kappa-\beta)w}}{(i\kappa-\beta)} - \frac{t(\kappa)e^{(-i\kappa-\beta)w}}{(i\kappa+\beta)}. \tag{77}$$

It follows that

$$\begin{aligned}\alpha A_{11}(w,\lambda) &= -r(\kappa)s(\kappa)e^{(i\kappa-\beta)w} - r(-\kappa)t(\kappa)e^{(-i\kappa-\beta)w}\\ &= \lambda t(\kappa)e^{(i\kappa-\beta)w} + \lambda s(\kappa)e^{(-i\kappa-\beta)w}\\ &= \lambda e^{-\beta w}\phi(-w,w)/C,\end{aligned} \tag{78}$$

and

$$\alpha A'_{11}(x,\lambda) = \alpha e^{-\beta x}\phi(x,w)/C. \tag{79}$$

Thus the eigenvalue condition (49) in this case reduces to the condition

$$\phi(-w,w) = 0. \tag{80}$$

To satisfy (80), we set

$$s(\kappa) = \rho e^{-i\gamma}, \tag{81}$$

with $\rho = |r(\kappa)|^{1/2}$ and $\gamma = \frac{1}{2}\arg r(\kappa)$:

$$\rho = |\lambda|^{1/2}, \quad \gamma = \tfrac{1}{2}\arctan(\kappa/\beta). \tag{82}$$

Then we have

$$t(\kappa) = \begin{cases} \rho e^{i\gamma} & \text{if } \lambda < 0,\\ -\rho e^{i\gamma} & \text{if } \lambda > 0,\end{cases} \tag{83}$$

and (76) becomes

$$\phi(x,w) = \begin{cases} 2|\lambda|^{1/2}C\cos(\kappa x - \gamma) & \text{if } \lambda < 0,\\ 2i|\lambda|^{1/2}C\sin(\kappa x - \gamma) & \text{if } \lambda > 0.\end{cases} \tag{84}$$

Then (80) requires

$$\begin{aligned}&\cos(\kappa w + \gamma) = 0 \quad \text{if } \lambda < 0,\\ &\sin(\kappa w + \gamma) = 0 \quad \text{if } \lambda > 0,\end{aligned} \tag{85}$$

or

$$\kappa w + \gamma = \begin{cases} (n+\frac{1}{2})\pi & \text{if } \lambda < 0,\\ (n+1)\pi & \text{if } \lambda > 0,\end{cases} \tag{86}$$

where $n = 0, \pm 1, \pm 2, \ldots$, and in either case we are led to the transcendental equation

$$\kappa/\beta + \tan 2\kappa w = 0 \tag{87}$$

for the admissible solutions of κ, and hence of λ, in terms of w. The associated eigenvalues and eigenfunctions are then just the admissible values λ_n of λ, and

$$\phi_n(x,w) = \begin{cases} C_n\cos(\kappa_n x - \gamma_n) & \text{if } \lambda_n < 0,\\ C_n\sin(\kappa_n x - \gamma_n) & \text{if } \lambda_n > 0,\end{cases} \tag{88}$$

where C_n is merely a normalizing constant.

The reader can now verify that if $\lambda^2 > \alpha^2/\beta^2$, then $\pm\kappa$ are replaced throughout by $\pm i\mu$ with μ real, so that (85) is replaced by

$$\begin{cases}\cosh(\mu w + \gamma) = 0 & \text{if } \lambda < 0,\\ \sinh(\mu w + \gamma) = 0 & \text{if } \lambda > 0,\end{cases} \tag{89}$$

admitting no new admissible values for λ. This verifies what already seems reasonable, that

$$0 \leqslant \lambda_n^2 \leqslant \alpha^2/\beta^2, \tag{90}$$

i.e., that the λ_n^2 must lie in the range of $|r(k)|^2$.

The kernel $K(x,y,w)$ is given by (61) with $\lambda = -1$, which, in view of (78), reduces to

$$K(x,y,w) = -(\alpha\cos(\kappa y - \gamma)/\cos(\kappa w + \gamma))e^{\beta(w-x)}. \tag{91}$$

Here we suppose that $\alpha^2 > \beta^2$, in which case $\lambda^2 = 1 < \alpha^2/\beta^2$, $\kappa = +(\alpha^2 - \beta^2)^{1/2}$ is real, and $\gamma = \frac{1}{2}\arctan((\alpha^2/\beta^2) - 1)^{1/2}$. If $1 > \alpha^2/\beta^2$, then $\kappa = i\mu$ is imaginary, with $\mu = +(\beta^2 - \alpha^2)^{1/2}$. Then $s(\kappa) = (\alpha/(\beta - \mu))^{1/2} = e^{\delta}$, and $t(\kappa) = (\alpha/(\beta - \mu))^{-1/2} = e^{-\delta}$, where now $\delta = \frac{1}{2}\log(\alpha/(\beta - \mu)) = \operatorname{arctanh}(\mu/\beta)$. Then $\phi(x,w) = 2C\cosh(\mu x - \delta)$ and $\alpha A_{11}(\omega, -1) = -e^{-\beta w} 2\cosh(\mu w + \delta)/C$. Hence if $\alpha^2 < \beta^2$, then (91) is replaced by

$$K(x,y,w) = -(\alpha\cosh(\mu y - \delta)/\cosh(\mu w + \delta))e^{\beta(w-x)}. \tag{92}$$

The intractability of (87) prohibits an explicit determination of $\lambda_n(w)$, or of $\Delta(w)$, in general. In the limiting case $\alpha = 1$, $\beta = 0$, however, we have $\gamma = \pi/4$, and (87) becomes

$$\kappa w + \frac{\pi}{4} = \begin{cases} (n + \frac{1}{2})\pi & \text{if } \lambda < 0, \\ (n+1)\pi & \text{if } \lambda > 0. \end{cases} \tag{93}$$

The positive admissible values of κ are

$$\kappa_n = \begin{cases} (4n+1)\pi/4w & \text{if } \lambda < 0, \\ (4n+3)\pi/4w & \text{if } \lambda > 0, \end{cases} \tag{94}$$

for $n = 0,1,2,\ldots$, and the admissible values of λ are

$$\lambda_n = \begin{cases} -1/\kappa_n = -4w/(4n+1)\pi & \text{if } \lambda_n < 0, \\ +1/\kappa_n = +4w/(4n+3)\pi & \text{if } \lambda_n > 0, \end{cases} \tag{95}$$

or

$$\lambda_n = (-1)^{n+1} 4w/(2n+1)\pi, \quad n = 0,1,2,\ldots. \tag{96}$$

Hence in this case

$$\begin{aligned} \Delta(w) &= \prod_{n=0}^{\infty}(1 + \lambda_n) \\ &= \prod_{n=0}^{\infty}\left(1 + \frac{(-1)^{n+1}4w}{(2n+1)\pi}\right) \\ &= 2^{1/2}\cos(w + \pi/4). \end{aligned} \tag{97}$$

On the other hand, from (78) we have in this case

$$A_{11}(w,\lambda) = (\lambda/C)\phi(-w,w). \tag{98}$$

In particular, for $\lambda = -1$, $\kappa = +1$,

$$A_{11}(w,-1) = -2(\cos\kappa w + \pi/4) \tag{99}$$

so that $\Delta(w)$ and $\det(A_{11}(w,-1))$ differ by the constant factor $-2^{1/2}$. The eigenfunctions in this case are given by

$$\phi_n(x,w) = \begin{cases} C_n\cos((4n+1)\pi x/4w - \pi/4) & \text{if } \lambda_n < 0, \\ C_n\sin((4n+3)\pi x/4w - \pi/4) & \text{if } \lambda_n < 0, \end{cases}$$

and the kernel $K(x,y,w)$ is given by [cf. (91)]

$$K(x,y,w) = \frac{-\cos(y - \pi/4)}{\cos(w + \pi/4)}. \tag{100}$$

We have assumed throughout this section that $\alpha > 0$ in (64). The reader may verify that if $\alpha < 0$, then everything is exactly the same except that the phase $\gamma = \frac{1}{2}\arg r(\kappa)$ is then augmented by π. We have avoided the case $\alpha^2/\beta^2 = 1$, since then, when $\lambda = -1$, $\kappa = 0$, and so $\pm\kappa$ are not distinct (cf. Ref. 10).

[1] I. M. Gel'fand and B. M. Levitan, Izv. Akad. Nauk SSSR, Math Series **15**, 309 (1951).

[2] I. Kay and H. E. Moses, *Inverse Scattering Papers: 1955–1963* (Math. Sci. Press, Brookline, MA, 1982).

[3] B. Simon, "Notes on Infinite Determinants of Hilbert Space Operators," Advances in Math. **24**, 244 (1977).

[4] H. Cornille, "Connection between the Marchenko formalism and N/D equations I.," J. Math. Phys. **8**, 2268 (1967).

[5] F. J. Dyson, "Fredholm Determinants and Inverse Scattering Problems," Comm. Math. Phys. **47**, 171 (1976); R. G. Newton, *Scattering Theory of Waves and Particles*, 2nd. ed. (McGraw-Hill, New York) (to appear); R. T. Prosser, "A General Solution for the One-Dimensional Inverse Scattering Problem," Dartmouth College, 1981 (unpublished).

[6] I. Kay, "The Inverse Scattering Problem When the Reflection Coefficient is a Rational Function," Comm. Pure Appl. Math. **13**, 371 (1960).

[7] K. R. Pechenick and J. Cohen "Inverse scattering—exact solution of the Gel'fand–Levitan equation," J. Math. Phys. **22**, 1513 (1981).

[8] H. E. Moses, "An Example of the Effect of the Rescaling of the Reflection Coefficient on the Scattering Potential for the One-Dimensional Schrödinger Equation," Stud. Appl. Math. **60**, 177 (1979).

[9] P. B. Abraham and H. E. Moses, "Exact Solutions of the One-Dimensional Acoustic Wave Equation for Several New Velocity Profiles. Transmission and Reflection Coefficients," J. Acoust. Soc. Am. **71**, 1391 (1982).

[10] P. B. Abraham, B. De Facio, and H. E. Moses, "Two Distinct Local Potentials with No Bound States Can Have the Same Scattering Operator: A Non-Uniqueness in Inverse Spectral Transformations," Phys. Rev. Lett. **46**, 1657 (1981).

The causal automorphism of de Sitter and Einstein cylinder spacetimes

J. A. Lester
Department of Pure Mathematics, University of Waterloo, Waterloo, Ontario, Canada

(Received 20 July 1982; accepted for publication 10 December 1982)

A well-known result, due originally to Alexandrov in 1953 and subsequently rediscovered by Zeeman in 1964, states that transformations of Minkowski spacetime which preserve causality are essentially orthochronous Lorentz transformations. In this article, we first exhibit a proof of this result by using a lemma of Zeeman to reduce the proof to another well-known theorem of Alexandrov involving transformations preserving light speed. Then, by generalizing Zeeman's lemma and using recent extensions of Alexandrov's light-speed theorem, we determine the causal automorphisms of de Sitter and Einstein cylinder spacetimes.

PACS numbers: 04.20. − q

1. INTRODUCTION: THE CAUSAL AUTOMORPHISM OF MINKOWSKI SPACETIME

Minkowski spacetime may be thought of as $\mathbb{R}^4$ equipped with the metric $(\ ,\ )$ given by

$$(x, y) := -x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4$$

for all $x := (x_1,x_2,x_3,x_4)$, $y := (y_1, y_2, y_3, y_4) \in \mathbb{R}^4$. The separation between events $x, y \in M_4$ is the quantity $(x-y, x-y)$, and is preserved by all translations and Lorentz transformations (linear, metric-preserving bijections) of M_4.

The separation between events in M_4 is zero iff they are joined by an unreflected light signal. Alexandrov's "light-speed" theorem[1,2] states that bijections of M_4 preserving separation zero in both directions must be Lorentz transformations, up to translations and dilatations (scale changes). The significance of this result is that, unlike Einstein's original derivation of Lorentz transformations,[3] it assumes no regularity conditions (e.g., linearity, or even continuity) for the transformations.

A vector $x \in M_4$ is said to be timelike, null, or spacelike whenever (x,x) is negative, zero, or positive, respectively. The nonzero null and timelike vectors lie, respectively, on and inside one of the two halves of a circular cone in M_4. They are thus segregated into two disconnected components, which we may (arbitrarily) label future-pointing vectors and past-pointing vectors. It is easily checked that, unless they are parallel null vectors, two nonspace-like vectors $x \neq 0$, $y \neq 0$ lie in the same component iff $(x, y) < 0$. Lorentz transformations which preserve future-pointing vectors are said to be orthochronous, and form a subgroup of the full Lorentz group.

Causality on M_4 may be defined in terms of future-pointing vectors as follows. A line in M_4 with timelike direction represents the spacetime history of a material particle experiencing no external force, while a line with null direction describes the history of an unreflected photon. Since an event $x \in M_4$ can cause an event $y \in M_4$ iff a material particle or photon can experience both events in that order, two corresponding causal relations, symbolized by $\prec$ and $\prec\cdot$, may be formulated.

Definition 1.1: For $x,y \in M_4$,

(i) $x \prec y$ iff $y - x$ is timelike and future pointing,

(ii) $x \prec\cdot y$ iff $y - x$ is null and future pointing.

The result which interests us here appeared first as one of several related results in Ref. 4; its rediscovery by Zeeman[5] appears to be better known (at least among physicists), possibly because the former article is in Russian (see Ref. 6 for historical background). The theorem states that bijections of M_4, which preserve the relation $\prec$ in both directions (Zeeman's "causal automorphisms"), must be orthochronous Lorentz transformations, up to translations and dilatations.

The significance of this result is again, as with Alexandrov's light-speed theorem, the absence of regularity assumptions on the transformations involved: preservation of a simple, physical condition is sufficient. For this reason, interest in these and similar characterizations has been growing steadily in recent years, particularly among geometers. Many generalizations now exist; these involve other spacetimes, other separations, more abstract light-cone structures, spaces over more general fields, etc. The bibliographies of Refs. 6 and 7, for example, provide a cross section of such works.

Zeeman's proof of the causality-preservation theorem on M_4 begins by showing that causal automorphisms must also preserve the relation $\prec\cdot$ in both directions. The crux of the matter is the following condition, for which we supply the proof omitted in Ref. 5.

Lemma 1.1: For distinct $x, y \in M_4$,

$$x \prec\cdot y \text{ iff } \begin{cases} x \not\prec y, \\ \text{for all } z \in M_4,\ z \prec x \text{ implies } z \prec y. \end{cases}$$

Proof: (a) Assume that $x \prec\cdot y$; then clearly $x \not\prec y$. For any $z \in M_4$ with $z \prec x$, write $y - z = (y - x) + (x - z)$; then

$$(y - z, y - z) = (y - x, y - x) + 2(y - x, x - z) + (x - z, x - z) < 0$$

since $y - x$ is null, $x - z$ is timelike, and both are future pointing. Thus $y - z$ is timelike, and, from $(y - z, y - x) = (x - z, y - x) < 0$, $y - z$ is future pointing (since $y - x$ is). Hence $z \prec y$.

(b) Assume that $x \not\prec\cdot y$ and $x \not\prec y$. For any timelike future-pointing vector t not parallel to $y - x$, the two-space spanned by t and $y - x$ contains a spacelike vector of the form $(y - x) + \alpha t$. If $y - x$ is spacelike, we may choose $\alpha > 0$ for small enough α; otherwise $y - x$ must be past pointing (else

$x \lessdot y$ or $x \lessdot \cdot y$), which implies $\alpha > 0$. In either case, the vector $z := x - \alpha t$ satisfies $z \lessdot x$, but $z \not\lessdot y$. ■

After restricting attention to the relation $\lessdot \cdot$, Zeeman's proof proceeds through properties of quadric surfaces, compositions of parallel displacements, Cauchy's functional equation, etc., eventually reaching the required result. However, since two events $x, y \in M_4$ have zero separation iff $x \lessdot \cdot y$ or $y \lessdot \cdot x$, the theorem follows immediately via Alexandrov's light-speed theorem. (The above lemma and the consequent shortcut actually work for Minkowski space M_n of any dimension $n \geqslant 3$, where both theorems are valid. For $n = 2$, both theorems fail.) In the next sections, generalizations of Lemma 1.1 and Alexandrov's result will yield the causal automorphisms of de Sitter and Einstein cylinder spacetimes.

2. CAUSAL AUTOMORPHISMS OF DE SITTER SPACETIME

de Sitter spacetime $\mathscr{S}_4$ can be embedded as a hyperboloid in five-dimensional Minkowski space M_5 (see Ref. 8, Sec. 5.2), i.e., if $(\ ,\ )$ denotes the metric of M_5, then $\mathscr{S}_4 := \{x | x \in M_5, (x,x) = 1\}$, and the (differential) metric of $\mathscr{S}_4$ is given by $ds^2 := (dx,dx)$. Events $x, y \in \mathscr{S}_4$ with $(x, y) > -1$ are joined by a geodesic (given by a section of $\mathscr{S}_4$ with a two-space in M_5; see Ref. 7) and their separation is s^2, where s (found by integrating ds along this geodesic) is the real or pure imaginary number given by $4 \sin^2(s/2) = (x - y, x - y)$. A direct generalization of Alexandrov's light-speed theorem[7] states that bijections of $\mathscr{S}_4$ preserving separation $s = 0$ [i.e., preserving the relation $(x, y) = 1$] in both directions must be induced on $\mathscr{S}_4$ by the Lorentz transformations of M_5.

The causal structure of $\mathscr{S}_4$ is induced by that of M_5; upon distinguishing the past-pointing and future-pointing vectors of M_5 as in Sec. 1, we define the causal relations $\lessdot$ and $\lessdot \cdot$ on $\mathscr{S}_4$ essentially as before.

Definition 2.1: For $x, y \in \mathscr{S}_4$,

(i) $x \lessdot y$ iff $y - x$ is timelike in M_5 and future pointing,
(ii) $x \lessdot \cdot y$ iff $y - x$ is null in M_5 and future pointing.

Clearly, the orthochronous Lorentz transformations of M_5 induce causality-preserving transformations of $\mathscr{S}_4$. Zeeman's condition, which generalizes exactly to $\mathscr{S}_4$ (see below), will enable us to establish these induced transformations as the only causal automorphisms of $\mathscr{S}_4$.

Lemma 2.1: For distinct $x, y \in \mathscr{S}_4$,

$$x \lessdot \cdot y \text{ iff } \begin{cases} x \not\lessdot y, \\ \text{for all } z \in \mathscr{S}_4,\ z \lessdot x \text{ implies } z \lessdot y. \end{cases}$$

Proof: If $x \lessdot \cdot y$, repeat part (a) of the proof of lemma 1.1 with $z \in \mathscr{S}_4$ to get the required results.

Assume that $x \not\lessdot \cdot y$ and $x \not\lessdot y$. Choose a future-pointing $t \in M_5$ with $(t,t) = -1$, $(t,x) = 0$, and for $\epsilon > 0$, define $z := (1 + \epsilon^2)^{1/2} x - \epsilon t$. Then $z \in \mathscr{S}_4$, $z \lessdot x$, and, for $\lambda := (x, y)$ and $\mu := (t, y)$, $(y - z, t) = \mu - \epsilon$ and $(y - z, y - z) = 2\{1 + \epsilon\mu - (1 + \epsilon^2)^{1/2}\lambda\}$. We find choices of ϵ for which $z \not\lessdot y$.

If $\mu > 0$, choose $\epsilon < \mu$; then $(y - z, t) > 0$, so $z \not\lessdot y$.

If $\mu = 0$, the space spanned by x and y is orthogonal to t, and is thus positive definite. The Cauchy–Schwarz inequality gives $\lambda^2 \leqslant 1$, so since $\lambda \neq 1$ $(x \neq y)$ we have $\lambda < 1$. Then for some $\epsilon > 0$, $(1 + \epsilon^2)^{1/2}\lambda < 1$, so for this ϵ, $(y - z, y - z) > 0$ and hence $z \not\lessdot y$.

If $\mu < 0$, then $\lambda < 1$ (else $x \lessdot \cdot y$ or $x \lessdot y$). If $\lambda > 0$, choose ϵ with $\epsilon(\lambda - \mu) < 1$; then $(1 + \epsilon^2)^{1/2}\lambda < (1 + \epsilon)\lambda < 1 + \mu\epsilon$, so $(y - z, y - z) > 0$. If $\lambda \leqslant 0$, choose $\epsilon < -\mu^{-1}$; then $1 + \mu\epsilon > 0$, so again $(y - z, y - z) > 0$. In either case, then, $z \not\lessdot y$. ■

Using the generalized light-speed theorem exactly as in Sec. 1, we have that bijections of $\mathscr{S}_4$ which preserve the relation $\lessdot$ in both directions must be induced by the orthochronous Lorentz transformations of M_5. We note that since the generalized Alexandrov result is in fact true for de Sitter spaces $\mathscr{S}_n$ of any dimension $n \geqslant 3$, so is our present result.

3. CAUSAL AUTOMORPHISMS OF EINSTEIN'S CYLINDER UNIVERSE

Einstein's cylinder universe $\mathscr{C}_4$ can be visualized as a circular cylinder in $\mathbb{R}^5$ (see Ref. 8, p. 121), i.e., if "$\cdot$" denotes the usual dot product of $\mathbb{R}^4$, then

$$\mathscr{C}_4 := \{(\rho, r) | \rho \in \mathbb{R}, r \in \mathbb{R}^4, r \cdot r = 1\} \text{ and}$$
$$ds^2 := -d\rho^2 + dr \cdot dr.$$

Its geodesics are either circular sections of $\mathscr{C}_4$ (which are spacelike) or of the form $r = \cos(\alpha\rho)a + \sin(\alpha\rho)b$ for some constant $\alpha \geqslant 0$ and orthonormal $a, b \in \mathbb{R}^4$ (and are timelike, null, or spacelike whenever $\alpha < 1$, $\alpha = 1$, or $\alpha > 1$, respectively). In general, two points of $\mathscr{C}_4$ are joined by many geodesics (e.g., for orthonormal $a, b \in \mathbb{R}^4$, the points (0,a) and $(\pi/2, b)$ are joined by all geodesics of the form $r = \cos[(1 + 4k)\rho]a + \sin[(1 + 4k)\rho]b$ for integral k, thus the separation s^2 between them (obtained by integrating ds along a joining geodesic) will be multivalued. For events (ρ_1, r_1), $(\rho_2, r_2) \in \mathscr{C}_4$ we obtain $s^2 = -(\rho_1 - \rho_2)^2 + \{\cos^{-1}(r_1 \cdot r_2)\}^2$, which is negative, zero, or positive whenever the geodesic is timelike, null, or spacelike, respectively.

The transformation group of $\mathscr{C}_4$ (i.e., the group of transformations of $\mathscr{C}_4$ which preserve ds^2 at each point) consists of mappings of the form $(\rho, r) \to (\pm\rho + \text{const.}, Ar)$, where A is a 4×4 orthogonal matrix. The light-speed theorem does *not* generalize to $\mathscr{C}_4$, since there exist rather pathological transformations of $\mathscr{C}_4$ which preserve separation zero (see Ref. 9 for details; the relevant points follow). For example, for fixed $(\rho, r) \in \mathscr{C}_4$, arbitrary permutations within the subset $\{(\rho + k\pi, (-1)^k r) | k \text{ an integer}\} \subset \mathscr{C}_4$ preserve separation zero. Up to such permutations, bijections of $\mathscr{C}_4$ which preserve separation zero [or equivalently, the relation $\cos(\rho_1 - \rho_2) = r_1 \cdot r_2$] in both directions have the form $(r, \cos\rho, \sin\rho)^t \to \lambda T(r, \cos\rho, \sin\rho)^t$ for a scalar function $\lambda = \lambda(\rho, r)$ (determined up to sign by the requirement that the condition $r \cdot r = 1$ be preserved) and a (constant) 6×6 matrix T satisfying $T^t G T = G$, where $G := \text{diag}\{1,1,1,1,-1,-1\}$ (superscript t denotes transpose).

Causality is defined on $\mathscr{C}_4$ by using the ρ-coordinate as a criterion of temporal order, i.e., an event $(\rho_1, r_1) \in \mathscr{C}_4$ can cause an event $(\rho_2, r_2) \in \mathscr{C}_4$ iff they can be joined by a timelike

or null geodesic and $\rho_1 < \rho_2$. Specifically, we define the causal relations $\prec$ and $\prec\cdot$ as follows.

Definition 3.1: For $(\rho_1,r_1), (\rho_2,r_2) \in \mathscr{C}_4$,

(i) $(\rho_1,r_1) \prec (\rho_2,r_2)$ iff $s^2 < 0$ for some geodesic joining them and $\rho_1 < \rho_2$,

(ii) $(\rho_1,r_1) \prec\cdot (\rho_2,r_2)$ iff $s^2 = 0$ for some geodesic joining them and $\rho_1 < \rho_2$.

A more useful characterization of these relations follows:

Lemma 3.1: For $(\rho_1,r_1), (\rho_2,r_2) \in \mathscr{C}_4$,

(i) $(\rho_1,r_1) \prec (\rho_2,r_2)$ iff $\rho_1 < \rho_2$ and either $\rho_2 - \rho_1 > \pi$ or $\cos(\rho_2 - \rho_1) < r_1 \cdot r_2$,

(ii) $(\rho_1,r_1) \prec\cdot (\rho_2,r_2)$ iff $\rho_1 < \rho_2$ and $\cos(\rho_1 - \rho_2) = r_1 \cdot r_2$.

Proof: for some $0 \leqslant \theta \leqslant \pi$, $r_1 \cdot r_2 = \cos\theta$, thus

$$\begin{aligned} s^2 &= -(\rho_1 - \rho_2)^2 + \{\cos^{-1}(\cos\theta)\}^2 \\ &= -(\rho_1 - \rho_2)^2 + (\pm\theta + 2k\pi)^2 \end{aligned}$$

for integral k. A tedious but elementary analysis of which k's are possible for $s^2 < 0$ and $s^2 = 0$ yields the required results. ■

Zeeman's condition must be slightly modified to hold on $\mathscr{C}_4$.

Lemma 3.2: For distinct $(\alpha,a), (\beta,b) \in \mathscr{C}_4$,

$$\left.\begin{array}{l}(\alpha,a) \prec\cdot (\beta,b) \\ \text{and } \beta - \alpha \leqslant \pi\end{array}\right\}$$

$$\text{iff} \begin{cases} (\alpha,a) \not\prec (\beta,b), \text{ and for all } (\gamma,c) \in \mathscr{C}_4, \\ (\gamma,c) \prec (\alpha,a) \text{ implies } (\gamma,c) \prec (\beta,b). \end{cases}$$

Proof: Without loss of generality we may assume that $\alpha = 0$. For some $0 \leqslant \omega \leqslant \pi$, $a \cdot b = \cos\omega$. Recall that the cosine function is decreasing on $[0,\pi]$.

(a) Assume that $\beta \leqslant \pi$ and that $(0,a) \prec\cdot (\beta,b)$. Then $0 < \beta \leqslant \pi$ and $\cos\beta = b \cdot a = \cos\omega$, so $\omega = \beta > 0$. Suppose there exists a $(\gamma,c) \in \mathscr{C}_4$ with $(\gamma,c) \prec (0,a)$ but $(\gamma,c) \not\prec (\beta,b)$. We have $\gamma < 0 < \beta$ and $0 - \gamma \leqslant \pi$ [else $\beta - \gamma > \pi$, which implies $(\gamma,c) \prec (\beta,b)$]; thus $\cos(-\gamma) < a \cdot c$. For $0 \leqslant \theta, \phi \leqslant \pi$ defined by $\cos\theta := a \cdot c$, $\cos\phi = b \cdot c$, we have $\cos(-\gamma) < \cos\theta$, $\cos(\beta - \gamma) \geqslant \cos\phi$; thus $-\gamma > \theta$ and $\beta - \gamma \leqslant \phi$, from which $\phi > \theta + \omega$ and $\cos\phi < \cos(\theta + \omega)$.

Since the subspace of $\mathbb{R}^4$ spanned by a, b, and c is positive definite,

$$0 \leqslant \begin{vmatrix} a\cdot a & b\cdot a & c\cdot a \\ a\cdot b & b\cdot b & c\cdot b \\ a\cdot c & b\cdot c & c\cdot c \end{vmatrix} = \begin{vmatrix} 1 & \cos\omega & \cos\theta \\ \cos\omega & 1 & \cos\phi \\ \cos\theta & \cos\phi & 1 \end{vmatrix},$$

which may be written

$$\{\cos(\theta + \omega) - \cos\phi\}\{\cos(\theta - \omega) - \cos\phi\} \leqslant 0.$$

The first factor has been proven positive, so $\cos\phi \geqslant \cos(\theta - \omega)$, whence $\theta + \omega \leqslant |\theta - \omega|$. This last implies the contradiction that either θ or ω is negative, thus no $(\gamma,c) \in \mathscr{C}_4$ with $(\gamma,c) \prec (0,a)$ and $(\gamma,c) \not\prec (\beta,b)$ exists.

(b) Assume that $(0,a) \not\prec (\beta,b)$ and that for all $(\gamma,c) \in \mathscr{C}_4$, $(\gamma,c) \prec (0,a)$ implies that $(\gamma,c) \prec (\beta,b)$. If $\beta > \omega$, then $\pi \geqslant \beta > \omega \geqslant 0$, and consequently $\cos\beta < \cos\omega = a \cdot b$. Hence $(0,a) \prec (\beta,b)$, a contradiction. If $\beta < \omega$, define $(\gamma,c) := (-\epsilon,a)$ for $0 < \epsilon < \omega - \beta$; then $\gamma < 0$ and $\cos(-\gamma) = \cos\epsilon < 1 = a \cdot c$, so $(\gamma,c) \prec (0,a)$. But $\cos(\beta - \gamma) = \cos(\beta + \epsilon) > \cos\omega = b \cdot c$ and $\beta - \gamma = \beta + \epsilon < \omega \leqslant \pi$, so $(\gamma,c) \not\prec (\beta,b)$, a contradiction.

If $\beta = \omega = 0$, then $a \cdot b = 1$, so $a = b$ and $(0,a) = (\beta,b)$, a contradiction. There remains the case $\beta = \omega > 0$, which yields $0 < \beta \leqslant \pi$ and $\cos\beta = a \cdot b$, from which $(0,a) \prec\cdot (\beta,b)$ as required. ■

We see that bijections of $\mathscr{C}_4$ which preserve the relation $\prec$ in both directions preserve zero separation for "close enough" points. The following lemma, which rules out the existence of "null triangles" in $\mathscr{C}_4$, will enable us to extend this result to more distant points.

Lemma 3.3: Three distinct points $(\alpha,a), (\beta,b), (\gamma,c) \in \mathscr{C}_4$ with pairwise zero separation lie on a common null geodesic.

Proof: Without loss of generality $\alpha = 0$, so $\cos\beta = a \cdot b$, $\cos\gamma = a \cdot c$, and $\cos(\beta - \gamma) = b \cdot c$. If b and c are parallel to a, then for some integers k,n, $(\beta,b) = (k\pi,(-1)^k a)$ and $(\gamma,c) = (n\pi,(-1)^n a)$. Then for any unit $d \in \mathbb{R}^4$ orthogonal to a, all three points lie on the null geodesic with equation $r = (\cos\rho)a + (\sin\rho)d$.

We may now assume that b is not parallel to a; thus $\sin\beta \neq 0$. Define $d := -\cot\beta a + \csc\beta b$; then d is unit and orthogonal to a, and $(0,a)$ and (β,b) lie on the null geodesic with equation $r = (\cos\rho)a + (\sin\rho)d$. For some scalars Ψ,ϕ and some $e \in \mathbb{R}^4$ orthogonal to a and d, $c = \Psi a + \phi d + e$, so $\Psi = a \cdot c = \cos\gamma$ and $\phi = c \cdot d = \sin\gamma$, from which $1 = c \cdot c = \cos^2\gamma + \sin^2\gamma + e \cdot e$. It follows that $e = 0$, so (γ,c) is also on the null geodesic. ■

Now consider two "distant" points $(\rho_1,r_1), (\rho_2,r_2) \in \mathscr{C}_4$ with separation zero. Cover the null geodesic segment joining them by a collection of open, overlapping subsegments whose points are "close enough," i.e., whose ρ-coordinates differ by at most π. By Lemma 3.3, the images of these subsegments under a causal automorphism are also overlapping segments of null geodesics. But the points of each image overlap lie on at most a single null geodesic, so in fact, all image segments lie on the *same* null geodesic. This geodesic joins the images of (ρ_1,r_1), (ρ_2,r_2), so these image points also have separation zero.

Since they preserve separation zero in both directions, our causal automorphisms have the form

$$(r,\cos\rho,\sin\rho)^t \to \lambda T (r,\cos\rho,\sin\rho)^t$$

for scalar λ and matrix T as described earlier, up to permutations within subsets of the form

$$\{(\rho + k\pi,(-1)^k r) | k \text{ an integer}\} \text{ for fixed } (\rho,r) \in \mathscr{C}_4.$$

But such permutations must now preserve causality, so since all points of the subset lie on a common null geodesic, their order must be preserved. It follows that if $(\bar\rho,\bar r)$ denotes the image of (ρ,r) under a causal automorphism, then the image of $(\rho + k\pi,(-1)^k r)$ is $(\bar\rho + k\pi,(-1)^k \bar r)$ for all integers k.

The scalar $\lambda = \lambda(\rho,r)$ is fixed up to sign by the requirement that $\bar r \cdot \bar r = 1 = \cos^2\bar\rho + \sin^2\bar\rho$. For (β,b), $(\gamma,c) \in \mathscr{C}_4$ we have

$$\begin{aligned} b \cdot c - \cos(\bar\beta - \bar\gamma) &= (\bar b,\cos\bar\beta,\sin\bar\beta) G (\bar c,\cos\bar\gamma,\sin\bar\gamma)^t \\ &= \lambda(\beta,b)\lambda(\gamma,c)(b,\cos\beta,\sin\beta)(T^t G T)(c,\cos\gamma,\sin\gamma)^t \\ &= \lambda(\beta,b)\lambda(\gamma,c)\{b \cdot c - \cos(\beta - \gamma)\} \end{aligned}$$

since $T^t G T = G := \operatorname{diag}\{1,1,1,1,-1,-1\}$. Since causality is preserved, all λ's have the same sign. We may in fact take

$\lambda > 0$: since $T^t GT = G$ iff $(-T)^t G(-T) = G$, minus signs may be absorbed into T.

For $(\alpha,a) \in \mathscr{C}_4$, consider the subset

$$\mathscr{M}(\alpha,a) := \{(\rho,r) \in \mathscr{C}_4 \mid (\alpha,a) \lessdot (\rho,r) \text{ and either } (\rho,r) \lessdot (\alpha + 2\pi,a) \text{ or } (\rho,r) \lessdot \cdot (\alpha + 2\pi,a) \text{ or } (\rho,r) = (\alpha + 2\pi,a)\}.$$

Clearly, if $(\alpha,a) \to (\overline{\alpha},\overline{a})$, then $\mathscr{M}(\alpha,a)$ maps into $\mathscr{M}(\overline{\alpha},\overline{a})$. Furthermore, careful examination of the definitions of $\lessdot$ and $\lessdot\cdot$ shows that

$$\mathscr{M}(\alpha,a) = \{(\rho,r) \in \mathscr{C}_4 \mid 0 < \rho - \alpha \leqslant 2\pi,\ \cos(\rho - \alpha) \leqslant r \cdot a, \text{ and if } \cos(\rho - a) = r \cdot a, \text{ then } \rho - \alpha > \pi\},$$

from which it can be checked that any point $(\rho,r) \in \mathscr{C}_4$ can be uniquely expressed as $(\sigma + k\pi, (-1)^k s)$ for some $(\sigma,s) \in \mathscr{M}(\alpha,a)$ and integer k. It follows that the image of any point of $\mathscr{C}_4$ is determined by the image of $\mathscr{M}(\alpha,a)$ for any given $(\alpha,a) \in \mathscr{C}_4$.

In summary, we may describe any bijection of $\mathscr{C}_4$ which preserves the relation $\lessdot$ in both directions as follows: choose $(\alpha,a) \in \mathscr{C}_4$ with image $(\overline{\alpha},\overline{a})$. Then for some 6×6 matrix T with $T^t GT = G := \operatorname{diag}\{1,1,1,1,-1,-1\}$ and for a uniquely determined scalar function $\lambda = \lambda(\sigma,s) > 0$, the causal automorphism maps $\mathscr{M}(\alpha,a)$ onto $\mathscr{M}(\overline{\alpha},\overline{a})$, and has the form $(\sigma,s) \to (\overline{\sigma},\overline{s})$, where

$$(\overline{s},\cos\overline{\sigma},\sin\overline{\sigma})^t = \lambda T(s,\cos\sigma,\sin\sigma)^t$$

on $\mathscr{M}(\alpha,a)$. Any point $(\rho,r) \in \mathscr{C}_4$ has the form $(\rho,r) = (\sigma + k\pi, (-1)^k s)$ for some unique $(\sigma,s) \in \mathscr{M}(\alpha,a)$ and integer k: its image is then $(\overline{\sigma} + k\pi, (-1)^k \overline{s})$.

It is easily checked that, given any point $(\alpha,a) \in \mathscr{C}_4$, any image point $(\overline{\alpha},\overline{a})$, and a 6×6 matrix T satisfying $T^t GT = G$, then the bijection of $\mathscr{C}_4$ defined as above by $(\alpha,a),(\overline{\alpha},\overline{a})$, and T is a causal automorphism of $\mathscr{C}_4$; we have thus characterized all causal automorphisms. As for Minkowski and de Sitter spacetimes, the characterization is in fact valid for n-dimensional Einstein cylinder spaces $\mathscr{C}_n$ for $n \geqslant 3$.

We note finally that the occurrence of the subsets $\mathscr{M}(\alpha,a)$ above is no accident: the interior of each is conformal to Minkowski spacetime M_4 (see Ref. 8, p. 122). The translations, dilatations, and orthochronous Lorentz transformations of M_4 induced transformations on $\mathscr{M}(\alpha,a)$ which, since they preserve the signs of separations between points, preserve causality on $\mathscr{M}(\alpha,a)$. The causal automorphisms obtained above are in fact compositions of these transformations with those of the transformation group of $\mathscr{C}_4$ described earlier.

[1] A. D. Alexandrov, "On Lorentz transformations," Usp. Mat. Nauk **5**, 187 (1950).

[2] A. D. Alexandrov, "A contribution to chronogeometry," Can. J. Math. **19**, 1110–1128 (1967).

[3] A. Einstein, "Zur Elektrodynamik bewegter Körper," Ann. Phys. **17**, 891–921 (1905).

[4] A. D. Alexandrov and V. V. Ovchinnikova, "Notes on the foundations of relativity theory," Vestn. Leningr. Univ. **11**, 95–100 (1953).

[5] E. C. Zeeman, "Causality implies the Lorentz group," J. Math. Phys. **5**, 490–493 (1964).

[6] A. D. Alexandrov, "Mappings of space with families of cones and space-time transformations," Ann. Mat. Pura Appl. (4) **103**, 229–257 (1975).

[7] J. A. Lester, "Separation-preserving transformations of de Sitter space-time," Abh. Math. Sem. Univ. Hamburg (to appear).

[8] S. W. Hawking and G. F. R. Ellis, in *The Large-Scale Structure of Space-time* (Cambridge University Press, Cambridge, 1973).

[9] J. A. Lester, "Alexandrov-type transformations of Einstein's cylinder universe," C. R. Math. Rep. Acad. Sci. Can. **4**, 175–178 (1982).

Spherically symmetric solution in the nonsymmetric Kaluza–Klein theory

M. W. Kalinowski[a] and G. Kunstatter
Physics Department, University of Toronto, Toronto, Ontario, M5S 1A7, Canada

(Received 6 May 1983; accepted for publication 26 August 1983)

In this paper we find an exact, static, spherically symmetric solution for the nonsymmetric Kaluza–Klein theory. This solution has the remarkable property of describing "mass without mass" and "charge without charge." We examine its properties and a physical interpretation.

PACS numbers: 04.50. + h, 11.10.Ef

INTRODUCTION

The aim of this paper is to find an exact spherically symmetric solution to the nonsymmetric Kaluza–Klein equations (see Refs. 1–7) in the electromagnetic case.[1,3]

The nonsymmetric Kaluza–Klein theory provides a true unification of the electromagnetic and gravitational fields in the following sense. It not only reduces two major principles of invariance (i.e., the local coordinate invariance principle and the local gauge invariance principle) to the local coordinate invariance principle, but it also gives rise to new effects, which are absent in the classical Kaluza–Klein theory. These effects do not appear in either Moffat's theory of gravitation (see Refs. 8–10) or in Maxwell's electromagnetism. They are therefore interference effects between the gravitational and electromagnetic fields. We outline these new features of the nonsymmetric Kaluza–Klein theory below (see Ref. 1):

1. A new term appears in the electromagnetic Lagrangian of the form

$$(1/4\pi)(g^{[\mu\nu]}F_{\mu\nu})^2.$$

2. There exists a vacuum electromagnetic polarization tensor $M_{\alpha\beta}$ which has a geometrical interpretation as torsion in the fifth dimension. Thus, there are two electromagnetic field strength tensors $F_{\alpha\beta}$ and $H_{\alpha\beta}$.

3. There is an additional term for the Lorentz force in the equation of motion for a test particle:

$$(q/m_0)g^{[\gamma\alpha]}H_{\gamma\beta}U^{\beta},$$

where q is the charge of the test particle and m_0 is its rest mass. This term plays the role of a reaction force for nonholonomic constraints.[1]

4. A new traceless energy-momentum tensor $T^{\rm em}_{\alpha\beta}$ appears for the electromagnetic field.

5. There exists a source for the electromagnetic field, i.e., the conserved current j^{α}.

All of the above effects vanish when the metric of spacetime is symmetric, in which case we get the classical Kaluza–Klein theory. Moreover, the new effects do not contradict any experimental or observational data.[1] The nonsymmetric Kaluza–Klein theory has a well-defined linear approximation.[11] In the electromagnetic case it has been shown[11] that there is no coupling between skewon and electromagnetic fields up to the first order in $h_{\mu\nu}\equiv g_{\mu\nu}-\eta_{\mu\nu}$ (where $\eta_{\mu\nu}$ is the Minkowski tensor). The nonsymmetric Kaluza–Klein theory also has a well-defined geometry on the five-dimensional manifold, which one calls Einstein geometry.[1] When the electromagnetic field vanishes, we get Moffat's nonsymmetric gravitation theory (NGT) which is able to fit the perihelion shift of Mercury in the presence of a nonzero quadrupole moment of mass for the sun.[12,13]

It is possible to extend the formalism of the nonsymmetric Kaluza–Klein theory to the nonabelian case[2,6] (including such features as spontaneous symmetry breaking and the Higgs mechanism) as well as to the Jordan–Thiry case[4,5,7], which possesses a scalar field connected to the gravitational constant. Material sources have also been incorporated[3] into this formalism.

It is of course important to find significant physical consequences of the "interference effects" present in the nonsymmetric Kaluza–Klein theory. The best way to achieve this is to find an exact solution of the full field equations, and this is the aim of this paper. We find an exact solution of the field equations in the static, spherically symmetric case in the form suggested in Sec. 6 of Ref. 1. Even in this, the simplest case, we get the following interesting results:

1. The electric field is nonsingular at $r=0$ and has Coulomb like behavior for large r. This is similar to the situation in Born–Infeld electrodynamics.[14] Thus, there is a maximal value of the electric field.

2. Asymptotically (for large r) the full solution behaves like the charged Reissner–Nördström type solution in NGT.[10]

3. The Newtonian mass is constructed from an electric charge Q and from a fermion charge l.

4. The energy distribution is not singular and is negative in a small region around $r=0$. This means that the solution describes a bounded system of electromagnetic and gravitational fields.

5. The total mass (i.e., total energy) of the solution is greater than the Newtonian mass (the mass which is seen at infinity).

6. There is no singularity at $r=0$ in the function $\alpha=g_{11}$; that is, $g_{11}(r=0)=1$.

7. The only singularities at $r=0$ are in $\omega\equiv g_{[14]}=l^2/r^2$ and in a factor $(1+l^4/r^4)$ in the function $\gamma=g_{44}$. There is also the usual singularity in the determinant of the full nonsymmetric tensor $\sqrt{-g}=r^2\sin\theta$ at $r=0$.

8. The charge distribution is nonsingular.

9. For sufficiently large charge Q there exists one or two

[a] On leave of absence from the Institute of Philosophy and Sociology of the Polish Academy of Sciences, 00-330 Warsaw, Nowy Swiat 72, Poland.

event horizons, just as in the Reissner–Nördström solution to the Einstein–Maxwell equations. Sufficiently large charge in the present case means sufficiently large Newtonian mass as well.

This solution is interesting as a classical model of a charged particle constructed from gravitational and electromagnetic fields. If we suppose that the Newtonian mass of our solution is the mass of an electron, we get a relationship between the classical radius of an electron and the parameter l from Moffat's theory of gravitation. The most facinating aspect of our solution is that it describes "mass without mass" and "charge without charge" in the following sense. At the origin $r = 0$ (or anywhere) there are no Coulomb-like or Newton-like first- and second-order poles with charge and mass as residues. This is true for the metric and for the electric field.

The paper is organized as follows. In the first section we describe some elements of the nonsymmetric Kaluza–Klein theory. The second section deals with the spherically symmetric fields in the nonsymmetric Kaluza–Klein theory, and presents the field equations in this case. The third section is devoted to the exact, static, spherically symmetric solution of the nonsymmetric Kaluza–Klein theory. We find this solution and examine its properties. In the fourth section we discuss our conclusions and prospects for further research. Appendices A and B contain some details of calculations; in Appendix A we derive the Ricci tensor in the general (nonstatic) spherically symmetric case, while in Appendix B we deal with some details concerning the static, spherically symmetric case. In Appendix C we write down the coefficients of the connection $\overline{\Gamma}$ and the Christoffel symbols for our solution as well as the equations of motion for uncharged and charged test particles.

1. ELEMENTS OF THE NONSYMMETRIC KALUZA–KLEIN THEORY

Let P be a principal fiber bundle with structural group $G = \mathrm{U}(1)$ over space-time E with projection π and let us define on this bundle a connection α. We call this bundle an electromagnetic bundle and α an electromagnetic connection. We define a curvature 2-form for the connection α:

$$\Omega = d\alpha = \tfrac{1}{2}\pi^*(F_{\mu\nu}\overline{\theta}^{\mu}\wedge\overline{\theta}^{\nu}), \tag{1.1}$$

where

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \quad e^*\alpha = A_\mu\overline{\theta}^{\mu}. \tag{1.2}$$

A_μ is a 4-potential of the electromagnetic field, e is a local section of $\underline{P}$, $F_{\mu\nu}$ is an electromagnetic field strength, and $\overline{\theta}^{\mu}$ is a frame on E. Bianchi's identity is

$$d\Omega = 0, \tag{1.3}$$

so that the 4-potential exists. This is of course simply the first Maxwell equation. On space-time E we define a nonsymmetric metric tensor $g_{\alpha\beta}$ such that

$$\begin{aligned} &g_{\alpha\beta} = g_{(\alpha\beta)} + g_{[\alpha\beta]}, \\ &g_{\alpha\beta}g^{\gamma\beta} = g_{\beta\alpha}g^{\beta\gamma} = \delta^{\gamma}_{\alpha}, \end{aligned} \tag{1.4}$$

where the order of indices is important. We define also on E two connections $\overline{\omega}^{\alpha}{}_{\beta}$ and $\overline{W}^{\alpha}{}_{\beta}$:

$$\overline{\omega}^{\alpha}{}_{\beta} = \overline{\Gamma}^{\alpha}_{\beta\gamma}\overline{\theta}^{\gamma} \tag{1.5}$$

and

$$\overline{W}^{\alpha}{}_{\beta} = \overline{W}^{\alpha}{}_{\beta\gamma}\overline{\theta}^{\alpha},$$

such that

$$\overline{W}^{\alpha}{}_{\beta} = \overline{\omega}^{\alpha}{}_{\beta} - \tfrac{2}{3}\delta^{\alpha}_{\beta}\overline{W}, \tag{1.6}$$

where

$$\overline{W} = \overline{W}_{\gamma}\overline{\theta}^{\gamma} = \tfrac{1}{2}(\overline{W}^{\sigma}_{\cdot\gamma\sigma} - \overline{W}^{\sigma}_{\cdot\sigma\gamma})\overline{\theta}^{\gamma}.$$

For the connection $\overline{\omega}^{\alpha}{}_{\beta}$ we suppose the following conditions:

$$\begin{aligned} &\overline{D}g_{\alpha+\beta-} = \overline{D}g_{\alpha\beta} - g_{\alpha\delta}\overline{Q}^{\delta}_{\beta\gamma}(\overline{\Gamma})\overline{\theta}^{\gamma} = 0, \\ &\overline{Q}^{\alpha}{}_{\beta\alpha}(\overline{\Gamma}) = 0, \end{aligned} \tag{1.7}$$

where $\overline{D}$ is the exterior covariant derivative with respect to $\overline{\omega}^{\alpha}{}_{\beta}$ and $\overline{Q}^{\alpha}{}_{\beta\gamma}(\overline{\Gamma})$ is the torsion of $\overline{\omega}^{\alpha}{}_{\beta}$. Thus we have defined on space-time E all quantities present in Moffat's theory of gravitation (see Refs. 8–10). Let us introduce on $\underline{P}$ a frame

$$\theta^{A} = (\pi^*(\overline{\theta}^{\alpha}), \lambda\alpha = \theta^{5}). \tag{1.8}$$

Now we turn to the natural nonsymmetric metrization of the bundle $\underline{P}$. According to Refs. 1–3 we have

$$\begin{aligned} &\overline{\gamma} = \pi^*\overline{g} - \theta^5\otimes\theta^5 = \pi^*(g_{(\alpha\beta)}\overline{\theta}^{\alpha}\otimes\overline{\theta}^{\beta}) - \theta^5\otimes\theta^5, \\ &\gamma = \pi^*g = \pi^*(g_{[\alpha\beta]}\overline{\theta}^{\alpha}\wedge\overline{\theta}^{\beta}). \end{aligned} \tag{1.9}$$

From the classical Kaluza–Klein theory we know that $\lambda = 2\sqrt{G}/c^2$ (see Ref. 1). We work with a system of units such that $G = c = 1$ and $\lambda = 2$. We have

$$\gamma_{AB} = \left(\begin{array}{c|c} g_{\alpha\beta} & 0 \\ \hline 0 & -1 \end{array}\right), \tag{1.10}$$

where

$$\gamma_{AB} = \gamma_{(AB)} + \gamma_{[AB]} \tag{1.11}$$

and

$$\overline{\gamma} = \gamma_{(AB)}\theta^{A}\otimes\theta^{B}, \tag{1.12}$$

$$\gamma = \gamma_{[AB]}\theta^{A}\wedge\theta^{B} \tag{1.13}$$

(see Refs. 1–3 for more details). Now we define on $\underline{P}$ a connection $\omega^{A}{}_{B}$ such that

$$D\gamma_{A+B-} = D\gamma_{AB} - \gamma_{AD}Q^{D}_{BC}(\Gamma)\theta^{C} = 0, \tag{1.14}$$

which is invariant with respect to the action of the group U(1) on $\underline{P}$. D is the exterior covariant derivative with respect to the connection $\omega^{A}{}_{B}$ and $Q^{D}{}_{BC}(\Gamma)$ is the tensor of torsion for the connection $\omega^{A}{}_{B}$. In Refs. 1 and 2 it is shown that

$$\omega^{A}{}_{B} = \left(\begin{array}{c|c} \pi^*(\overline{\omega}^{\alpha}{}_{\beta}) + g^{\gamma\alpha}H_{\gamma\beta}\theta^{5} & H_{\beta\gamma}\theta_{\gamma} \\ \hline g^{\alpha\beta}(H_{\gamma\beta} + 2F_{\beta\gamma})\theta^{\gamma} & 0 \end{array}\right), \tag{1.15}$$

where $H_{\beta\gamma}$ is a tensor on E such that

$$g_{\delta\beta}g^{\gamma\delta}H_{\gamma\alpha} + g_{\alpha\delta}g^{\delta\gamma}H_{\beta\gamma} = 2g_{\alpha\delta}g^{\delta\gamma}F_{\beta\gamma}. \tag{1.16}$$

In order to get the usual interpretation of geodesics in the classical Kaluza–Klein theory we must assume[1–3]

$$H_{\alpha\beta} = -H_{\beta\alpha}. \tag{1.17}$$

We define on $\underline{P}$ a second connection

$$W^{A}{}_{B} = \left(\begin{array}{c|c} \pi^{*}(\overline{W}^{\alpha}{}_{\beta}) + g^{\gamma\alpha}H_{\gamma\beta}\theta^{5} & H_{\beta\gamma}\theta^{\gamma} \\ \hline g^{\alpha\beta}(H_{\gamma\beta} + 2F_{\beta\gamma})\theta^{\gamma} & 0 \end{array}\right) \tag{1.18}$$

Let us define a Moffat–Ricci curvature scalar for $W^{A}{}_{B}$. One gets[1-3]

$$R(W) = \overline{R}(\overline{W}) + (2(g^{[\mu\nu]}F_{\mu\nu})^{2} - H^{\mu\alpha}F_{\mu\alpha}), \tag{1.19}$$

where

$$\overline{R}(\overline{W}) = g^{\mu\nu}\overline{R}_{\mu\nu}(\overline{\Gamma}) + \tfrac{2}{3}g^{[\beta\mu]}\overline{W}_{[\beta,\mu]} \tag{1.20a}$$

is a Moffat–Ricci curvature scalar for the connection $\overline{W}^{\alpha}{}_{\beta}$ and $\overline{R}_{\alpha\beta}(\overline{\Gamma})$ is a Moffat–Ricci curvature for the connection $\overline{\omega}^{\alpha}{}_{\beta}$. In particular,

$$\overline{R}_{\mu\nu}(\overline{\Gamma}) = \overline{R}^{\alpha}{}_{\mu\nu\alpha}(\overline{\Gamma}) + \tfrac{1}{2}\overline{R}^{\alpha}{}_{\alpha\mu\nu}(\overline{\Gamma}), \tag{1.20b}$$

where $\overline{R}^{\alpha}{}_{\mu\nu\rho}(\overline{\Gamma})$ are the components of the ordinary curvature tensor for $\overline{\Gamma}$. In addition

$$H^{\mu\alpha} = g^{\beta\mu}g^{\gamma\alpha}H_{\beta\gamma}. \tag{1.21}$$

From Eq. (1.19) one gets the field equations[1]

$$\overline{R}_{\alpha\beta}(\overline{W}) - \tfrac{1}{2}g_{\alpha\beta}\overline{R}(\overline{W}) = 8\pi T^{\mathrm{em}}_{\alpha\beta}, \tag{1.22}$$

$$\mathbf{g}^{[\mu\nu]}{}_{,\nu} = 0, \tag{1.23}$$

$$g_{\mu\nu,\sigma} - g_{\zeta\nu}\overline{\Gamma}^{\zeta}_{\mu\sigma} - g_{\mu\zeta}\overline{\Gamma}^{\zeta}_{\sigma\nu} = 0, \tag{1.24}$$

$$\partial_{\mu}(\mathbf{H}^{\alpha\mu}) = 4\mathbf{g}^{[\alpha\beta]}\partial_{\beta}(g^{[\mu\nu]}F_{\mu\nu}), \tag{1.25}$$

where

$$T^{\mathrm{em}}_{\alpha\beta} = (1/4\pi)(g^{\gamma\mu}H_{\gamma\alpha}F_{\mu\beta} - 2g^{[\mu\nu]}F_{\mu\nu}F_{\alpha\beta} - \tfrac{1}{4}g_{\alpha\beta}(H^{\mu\nu}F_{\mu\nu} - 2(g^{[\mu\nu]}F_{\mu\nu})^{2})), \tag{1.26}$$

$$\mathbf{g}^{[\mu\nu]} = \sqrt{-g}\,g^{[\mu\nu]},$$

$$\mathbf{H}^{\mu\alpha} = \sqrt{-g}\,g^{\beta\mu}g^{\gamma\alpha}H_{\beta\gamma}. \tag{1.27}$$

The tensor $H_{\mu\nu}$ has an interpretation as a second electromagnetic field strength tensor.[1-3] We have

$$g^{\alpha\beta}T^{\mathrm{em}}_{\alpha\beta} = 0. \tag{1.28}$$

Equations (1.22)–(1.25) can be written in the form

$$\overline{R}_{(\alpha\beta)}(\overline{\Gamma}) = 8\pi T^{\mathrm{em}}_{(\alpha\beta)}, \tag{1.29}$$

$$\overline{R}_{[[\alpha\beta],\gamma]}(\overline{\Gamma}) - 8\pi T^{\mathrm{em}}_{[[\alpha\beta],\gamma]} = 0, \tag{1.30}$$

$$\overline{\Gamma}_{\mu} = 0, \tag{1.31}$$

$$g_{\mu\nu,\sigma} - g_{\zeta\nu}\overline{\Gamma}^{\zeta}_{\mu\sigma} - g_{\mu\zeta}\overline{\Gamma}^{\zeta}_{\sigma\nu} = 0, \tag{1.32}$$

$$\partial_{\mu}(\mathbf{H}^{\alpha\mu} - 4\mathbf{g}^{[\alpha\mu]}(g^{[\nu\beta]}F_{\nu\beta})) = 0, \tag{1.33}$$

where $\overline{R}_{\alpha\beta}(\overline{\Gamma})$ is a Moffat–Ricci tensor for the connection

$$\overline{\omega}^{\alpha}{}_{\beta} = \overline{\Gamma}^{\alpha}_{\beta\gamma}\overline{\theta}^{\gamma}, \qquad \overline{\Gamma}_{\mu} = \overline{\Gamma}^{\alpha}_{[\mu\alpha]}. \tag{1.34}$$

The condition (1.31) is equivalent to (1.23).

2. SPHERICALLY SYMMETRIC FIELDS IN THE NONSYMMETRIC KALUZA–KLEIN THEORY

Let us suppose that the fundamental fields in the nonsymmetric Kaluza–Klein theory possesses spherical symmetry. According to Refs. 15–23 one gets

$$g_{\mu\nu} = \begin{pmatrix} -\alpha & 0 & 0 & \omega \\ 0 & -\beta & f\sin\theta & 0 \\ 0 & -f\sin\theta & -\beta\sin^{2}\theta & 0 \\ -\omega & 0 & 0 & \gamma \end{pmatrix}, \tag{2.1}$$

where α, β, γ, f, and ω are real functions of r and t with $\alpha, \gamma > 0$. In addition

$$F_{14} = E(r,t), \quad F_{23} = B\sin\theta \tag{2.2}$$

and all other components of $F_{\mu\nu}$ vanish. For $g^{\mu\nu}$, the only nonvanishing components are

$$\begin{aligned} g^{11} &= \gamma/(\omega^{2} - \alpha\gamma), \\ g^{22} &= g^{23}\sin^{2}\theta = -\beta/(\beta^{2} + f^{2}), \\ g^{44} &= -\alpha/(\omega^{2} - \alpha\gamma), \\ g^{[14]} &= \omega/(\omega^{2} - \alpha\gamma), \\ g^{[23]}\sin\theta &= f/(\beta^{2} + f^{2}). \end{aligned} \tag{2.3}$$

We suppose that

$$\omega^{2} - \alpha\gamma \neq 0 \quad \text{and} \quad \beta^{2} + f^{2} \neq 0. \tag{2.4}$$

Let us suppose that $H_{\alpha\beta}$ is also spherically symmetrical, so that

$$H_{14} = D(r,t), \quad H_{23} = H\sin\theta \tag{2.5}$$

and the other components vanish. Using Eqs. (1.16), (2.1), and (2.3) it can be shown that

$$\begin{aligned} H_{14} &= F_{14} = E(r,t), \\ H_{23} &= F_{23} = B\sin\theta. \end{aligned} \tag{2.6}$$

The Bianchi identity equation (1.3) yields

$$B = B_{0} = \text{const}. \tag{2.7}$$

From Eq. (1.23) one gets

$$\frac{\omega^{2}}{\alpha\gamma - \omega^{2}} = \frac{l^{4}}{\beta^{2} + f^{2}}, \tag{2.8}$$

where l is a constant of integration. In Moffat's theory of gravitation this constant has an interpretation as fermion charge. From Eq. (1.33) we have

$$\frac{E}{\omega} = \frac{-(Q/l^{2})(\beta^{2} + f^{2}) + 8fB_{0}}{(\beta^{2} + f^{2} + 8f^{4})}, \tag{2.9}$$

where Q is an integration constant. In the intermediate stages of calculation we used the following expressions for $H^{\mu\alpha}$ and $\sqrt{-g}$

$$H^{14} = -\frac{H_{14}}{(\alpha\gamma - \omega^{2})} = \frac{-E}{(\alpha\gamma - \omega^{2})}, \tag{2.10}$$

$$H^{23} = \frac{B_{0}}{\beta^{2} + f^{2}}, \tag{2.11}$$

$$\sqrt{-g} = \sin\theta\,[(\alpha\gamma - \omega^{2})(\beta^{2} + f^{2})]^{1/2}. \tag{2.12}$$

Thus finally we get Eqs. (1.29)–(1.32) plus the algebraic relations (2.7)–(2.9). From Eq. (1.30) we get immediatelly

$$R_{[23]}(\overline{\Gamma}) - 8\pi T^{\mathrm{em}}_{[23]} = C_{1}\sin\theta, \tag{2.13}$$

where $C_{1} = \text{const}$ is an integration constant and

$$\frac{8\pi}{\sin\theta}T^{\rm em}_{[23]} = -\frac{7fB_0^2}{\beta^2+f^2} + \frac{fl^4}{(\beta^2+f^2)}\left(\frac{E}{\omega}\right)^2 + 4f\left(\frac{fB_0}{(\beta^2+f^2)} - \frac{l^4}{(\beta^2+f^2)}\frac{E}{\omega}\right)^2 + \frac{8B_0l^4}{(\beta^2+f^2)}\left(\frac{E}{\omega}\right). \quad (2.14)$$

Equations (1.31) and (1.32) were solved in Ref. 17 in which Pant wrote down the Ricci tensor for such a connection.

Note that the Moffat–Ricci tensor [Eq. (1.20b)] is a linear combination of the ordinary Ricci tensor and the second contraction of the curvature tensor. However, Eqs. (1.23) and (1.24) imply that[10]

$$\bar{\Gamma}^{\alpha}_{[\mu\alpha]} = 0 \quad (2.15)$$

and

$$\bar{\Gamma}^{\beta}_{\nu\beta} = [\ln((-g)^{1/2})]_{,\nu} \quad (2.16)$$

so that the second contraction

$$\bar{R}^{\alpha}_{\alpha\mu\nu} = \tfrac{1}{2}(\bar{\Gamma}^{\beta}_{(\mu\beta),\nu} - \bar{\Gamma}^{\beta}_{(\nu\beta),\nu}) = 0. \quad (2.17)$$

Consequently the Moffat–Ricci tensor in this case is identically equal to the ordinary Ricci tensor used by Pant,[17] which we shall denote by $A_{\mu\nu}(\bar{\Gamma})$.

Thus we get the equations

$$A_{(\mu\nu)}(\bar{\Gamma}) = 8\pi T^{\rm em}_{(\mu\nu)}, \quad (2.18)$$

$$A_{[23]}(\bar{\Gamma}) - 8\pi T^{\rm em}_{[23]} = C_1\sin\theta,$$

where

$$8\pi T^{\rm em}_{11} = \alpha\left(\frac{l^4}{\beta^2+f^2}\right)\frac{E^2}{\omega^2} + \frac{\alpha B_0^2}{\beta^2+f^2} - 4\alpha\left(\frac{fB_0}{\beta^2+f^2} - \frac{l^4}{(\beta^2+f^2)}\frac{E}{\omega}\right)^2. \quad (2.19)$$

Using Eq. (2.9), the last term in Eq. (2.19) can be written in the form

$$-4\alpha\left(\frac{fB_0+Ql^2}{\beta^2+f^2+8l^4}\right)^2. \quad (2.20)$$

Moreover, it can be shown that

$$8\pi T^{\rm em}_{44} = -(\gamma/\alpha)8\pi T^{\rm em}_{11}, \quad (2.21)$$

$$8\pi T^{\rm em}_{22} = \frac{8\pi}{\sin^2\theta}T^{\rm em}_{33} = \frac{\beta}{\alpha}8\pi T^{\rm em}_{11} = -\frac{\beta B_0^2}{(\beta^2+f^2)} - \frac{\beta l^4}{(\beta^2+f^2)}\left(\frac{E^2}{\omega^2}\right) - 4\beta\left(\frac{fB_0}{(\beta^2+f^2)} - \frac{l^4}{(\beta^2+f^2)}\frac{E}{\omega}\right)^2, \quad (2.22)$$

$$8\pi T^{\rm em}_{14} = -8\pi T^{\rm em}_{41} = \frac{\omega}{(\beta^2+f^2)}\left(7l^4\left(\frac{E^2}{\omega^2}\right) - 8fB_0\left(\frac{E}{\omega}\right) - B_0^2\right) - \omega\left(\frac{fB_0}{(\beta^2+f^2)} - \frac{l^4}{(\beta^2+f^2)}\left(\frac{E}{\omega}\right)\right)^2. \quad (2.23)$$

The rest of the components of $T^{\rm em}_{\mu\nu}$ vanish. The electromagnetic Lagrangian in this case is

$$\mathscr{L}_{\rm em} = \frac{1}{8\pi}(2(g^{[\mu\nu]}F_{\mu\nu})^2 - H^{\mu\nu}F_{\mu\nu}) = \frac{1}{8\pi}\left[\frac{8\omega^4}{(\alpha\gamma-\omega^2)^2}\left(\frac{fB_0}{l^4} - \frac{E}{\omega}\right)^2 - \frac{2\omega^2}{(\alpha\gamma-\omega^2)}\left(\frac{B_0^2}{l^4} - \frac{E^2}{\omega^2}\right)\right]. \quad (2.24)$$

Finally, we have the following equations:

$$A_{11}(\bar{\Gamma}) = 8\pi T^{\rm em}_{11}, \quad (2.25a)$$

$$A_{44}(\bar{\Gamma}) = 8\pi T^{\rm em}_{44}, \quad (2.25b)$$

$$A_{22}(\bar{\Gamma}) = 8\pi T^{\rm em}_{22}, \quad (2.25c)$$

$$A_{33}(\bar{\Gamma}) = 8\pi T^{\rm em}_{33}, \quad (2.25d)$$

$$A_{[23]}(\bar{\Gamma}) - 8\pi T^{\rm em}_{23} = C_1\sin\theta, \quad (2.25e)$$

$$A_{(14)}(\bar{\Gamma}) = 0. \quad (2.25f)$$

Using results from Ref. 17 and Eq. (2.22) one finds the identity (see Appendix A)

$$(A_{22}(\bar{\Gamma}) - 8\pi T^{\rm em}_{22}) \equiv (1/\sin^2\theta)(A_{33}(\bar{\Gamma}) - 8\pi T^{\rm em}_{33}), \quad (2.26)$$

so that Eq. (2.25d) is not independent. In the above

$$8\pi T^{\rm em}_{11} = \alpha\left[\frac{(8l^2fB_0 - Q(\beta^2+f^2))^2 + B_0^2(\beta^2+f^2+8l^4)^2 - (fB_0+Ql^2)(\beta^2+f^2)}{(\beta^2+f^2)(\beta^2+f^2+8l^4)^2}\right], \quad (2.27)$$

$$\frac{8\pi}{\sin\theta}T^{\rm em}_{23} = \frac{8\pi}{\sin\theta}T^{\rm em}_{[23]} = \frac{[-7fB_0(\beta^2+f^2+8l^4)^2 - f(8fB_0 - Q(\beta^2+f^2))^2]}{(\beta^2+f^2)(\beta^2+f^2+8l^4)^2} + \frac{[8B_0l^4(8B_0l^2 - Q(\beta^2+f^2))(\beta^2+f^2+8l^4) + 4f(\beta^2+f^2)(fB_0+Ql^2)^2]}{(\beta^2+f^2)(\beta^2+f^2+8l^4)^2}. \quad (2.28)$$

For $T^{\rm em}_{14}$ one finds

$$8\pi T^{\rm em}_{14} = 8\pi T^{\rm em}_{[14]} = \frac{\omega}{(\beta^2+f^2)}\frac{[7l^2(8l^2fB_0 - Q(\beta^2+f^2))^2 - 8B_0f(8l^2fB_0 - Q(\beta^2+f^2))(-l^2B_0(\beta^2+f^2+8l^4)^2]}{l^2(\beta^2+f^2+8l^4)} \quad (2.29)$$

$A_{11}(\bar{\Gamma})$, $A_{44}(\bar{\Gamma})$, $A_{33}(\bar{\Gamma})$, $A_{(14)}(\bar{\Gamma})$, $A_{[14]}(\bar{\Gamma})$, and $A_{[23]}(\bar{\Gamma})$ are given by the formulas (2.11) (see Appendix A) from Ref. 17. For $\mathscr{L}_{\rm em}$ one easily gets, using (2.24),

$$\mathscr{L}_{\rm em} = \frac{1}{4\pi}\frac{l^4}{(\beta^2+f^2)}\left[\frac{4^4}{(\beta^2+f^2)}\left(\frac{fB_0}{l^2} - \frac{(8l^2fB_0 - Q(\beta^2+f^2))}{(\beta^2+f^2+8l^4)}\right)^2 - \frac{1}{l^4}\left(B_0^2 - \frac{(8fl^2B_0 - Q(\beta^2+f^2))^2}{(\beta^2+f^2+8l^4)^2}\right)\right]. \quad (2.30)$$

3. STATIC, SPHERICALLY SYMMETRIC SOLUTION

Let us consider a spherical field configuration such that

$$B_0 = f = 0. \tag{3.1}$$

Later we suppose that

$$\beta = r^2, \tag{3.2}$$

which is simply a coordinate choice. In addition, our quantities do not depend on time (static case). One finds [see Eq. (2.9)]

$$E = -\omega \frac{Q}{l^2}\left(\frac{r^4}{r^4 + 8l^2}\right) \tag{3.3}$$

(substituting $\beta = r^2$). Equations (2.29) now read

$$\begin{aligned} &A_{11}(\bar{\Gamma}) - \frac{\alpha Q^2}{(\beta^2 + 8l^4)}\left(\frac{\beta^2 - 4l^4}{\beta^2 + 8l^4}\right) = 0, \\ &A_{44}(\bar{\Gamma}) + \frac{\gamma Q^2}{(\beta^2 + 8l^4)}\frac{(\beta^2 - 4l^4)}{(\beta^2 + 8l^4)} = 0, \\ &A_{22}(\bar{\Gamma}) - \frac{\beta Q^2}{(\beta^2 + 8l^4)}\frac{(\beta^2 - 4l^4)}{(\beta^2 + 8l^4)} = 0, \\ &A_{(14)} = 0, \\ &A_{[23]} - 8\pi T^{\text{em}}_{23} = C_1 \sin\theta, \end{aligned} \tag{3.4}$$

and we have

$$8\pi T^{\text{em}}_{[14]} = 8\pi T^{\text{em}}_{14} = \omega Q \frac{(7\beta^2 + 16l^4)}{(\beta^2 + 8l^4)^2}, \tag{3.5}$$

$$\omega = l^2/r^2. \tag{3.6}$$

One gets

$$E = -\frac{Q}{r^2}\left(\frac{r^4}{r^4 + 8l^4}\right). \tag{3.7}$$

It is easy to see that the function (3.7) is bounded

$$|E| < E_{\max} = |E(\sqrt[4]{8}l)| = |Q|/8l^2. \tag{3.7a}$$

From (3.4), using results from Ref. 17, one gets (see Appendix B)

$$\frac{d}{dr}(r\alpha^{-1}) = 1 - Q^2 r^2 \frac{(r^4 + 4l^4)}{(r^4 + 8l^4)^2}. \tag{3.8}$$

Thus we have

$$\frac{1}{\alpha} = 1 + \frac{C}{r} + \frac{Q^2}{r}K(r,l), \tag{3.9}$$

where

$$K(r,l) = -\int r^2 \frac{(r^4 + 4l^4)}{(r^4 + 8l^4)^2}\,dr \tag{3.10}$$

and C is a constant of integration. Moreover,

$$\gamma = \left(1 + \frac{C}{r} + \frac{Q^2}{r}K(r,l)\right)\left(1 + \frac{l^4}{r^4}\right) \tag{3.11}$$

[see Eqs. (B8) and (B11) in Appendix B]. Performing the integration in (3.10) one gets

$$\frac{1}{\alpha} = 1 + \frac{C}{r} + \frac{Q^2 b}{b^2 r}g\left(\frac{b}{r}\right), \tag{3.12}$$

where $b^4 = 8l^4$ and

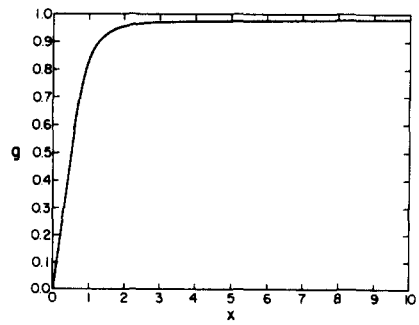

FIG. 1. The function $g = g(x)$ vs x.

$$g(x) = \frac{1}{8}\left(\frac{x}{x^4 + 1}\right) + \frac{7}{32\sqrt{2}}\left(\log\left(\frac{x^2 + \sqrt{2}x + 1}{x^2 - \sqrt{2}x + 1}\right) + 2\arctan(\sqrt{2}x + 1) + 2\arctan(\sqrt{2}x - 1)\right). \tag{3.13}$$

The function $g(x)$ is plotted in Fig. 1. Let us examine the properties of the function

$$g(b/r).$$

It can be shown that

$$\lim_{r=0} g\left(\frac{b}{r}\right) = \frac{7}{16}\frac{\pi}{\sqrt{2}}. \tag{3.14}$$

Thus for small r we get

$$\alpha^{-1} \simeq 1 + \frac{1}{r}\left(C + \frac{7}{16\sqrt{2}}\pi\left(\frac{Q^2}{b}\right)\right). \tag{3.15}$$

We can avoid a singularity in α at $r = 0$ by choosing

$$C = -\frac{7}{16\sqrt{2}}\pi\left(\frac{Q^2}{b}\right) \tag{3.16}$$

so that

$$\lim_{r=0}(\alpha^{-1}) = 1. \tag{3.17}$$

Let us examine the asymptotic properties of α and γ. One gets

$$\alpha^{-1} \underset{r\to\infty}{\to} \left(1 - \frac{[(7/16\sqrt{2}\pi)Q^2/b]}{r} + \frac{Q^2}{r^2}\right). \tag{3.18}$$

For large r, α clearly behaves like the analogous function in the Reissner–Nördström solution, with Q as the electric charge and with

$$m_{\text{N}} = \frac{7}{32\sqrt{2}}\pi\left(\frac{Q^2}{b}\right) \tag{3.19}$$

playing the role of the Newtonian mass. To summarize, we have

$$\alpha^{-1} = \left(1 - \frac{7}{8\sqrt{2}}\left(\frac{\pi}{2}\right)\frac{Q^2/b}{r} + \frac{Q^2}{r^2}\tilde{g}\left(\frac{b}{r}\right)\right), \tag{3.20}$$

where

$$\lim_{r\to\infty}\tilde{g}\left(\frac{b}{r}\right) = 1, \tag{3.21}$$

$$\tilde{g}\left(\frac{b}{r}\right) = \frac{g(b/r)}{(b/r)}, \tag{3.21a}$$

and

$$\lim_{r\to 0} \alpha^{-1} = 1. \tag{3.22}$$

In the neighborhood of $r = 0$ one gets for our metric

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & \frac{l^2}{r^2} \\ 0 & -r^2 & 0 & 0 \\ 0 & 0 & -r^2 \sin\theta & 0 \\ -\frac{l^2}{r^2} & 0 & 0 & \left(1+\frac{l^4}{r^4}\right) \end{pmatrix} \tag{3.23}$$

(for $r\to 0$). The determinant of the symmetric part of the metric is

$$(-\tilde{g})^{1/2} = (r^4 + l^4)^{1/2} \sin\theta. \tag{3.24}$$

The full determinant is

$$\sqrt{-g} = r^2 \sin\theta. \tag{3.25}$$

Thus there is a singularity at $r = 0$. It is worth noting, however, that there is no singularity in α and only one singularity in γ due to the $(1 + l^4/r^4)$ factor. ω, the skew-symmetric part of $g_{\mu\nu}$, is also singular at $r = 0$.

Let us examine properties of the electric field:

$$E = -\frac{Q}{r^2}\left(\frac{r^4}{r^4 + 8l^4}\right). \tag{3.26}$$

One easily sees that

$$E(0) = 0 \tag{3.27}$$

and

$$E \underset{r\to\infty}{\to} -\frac{Q}{r^2}. \tag{3.28}$$

Thus there is no singularity at $r = 0$. This is similar to the situation in Born–Infeld electrodynamics.[14] Let us calculate the charge distribution and total charge for the electric field. It is known that

$$4\pi\sqrt{-g}\rho = \mathbf{H}^{4i}{}_{,i} \sim \operatorname{div}\vec{D}, \tag{3.29}$$

where ρ is the charge density distribution and $\vec{D}$ is an electric induction vector. One gets

$$\mathbf{H}^{41} = \sqrt{-g}E/(\alpha\gamma - \omega^2) = \sqrt{-g}E \tag{3.30}$$

and

$$\sqrt{-g}\rho = -\frac{1}{\pi}\frac{Q}{r}\frac{(8l^4r^4)}{(r^4 + 8l^4)^2}\sin\theta. \tag{3.31}$$

The total charge is

$$Q_{\text{tot}} = \int \sqrt{-g}\rho\, d^3x = -32Ql^4 \int_0^\infty \frac{1}{r}\frac{r^4}{(r^4 + 8l^4)^2}\, dr = -Q. \tag{3.32}$$

Thus we find the following interesting feature: the total electric charge defined above is the same as the charge obtained from the asymptotic properties of the electric field E and the metric (functions α and γ). Let us pass on the calculation of the energy of the electromagnetic field. One has

$$T^4{}_4 = \frac{1}{8\pi}Q^2\left(\frac{(r^4 - 10l^4)}{(r^4 + 8l^4)^2}\right). \tag{3.33}$$

The total energy

$$E_{\text{tot}} = 4\pi\int_0^\infty r^2 \overset{\text{em}}{T}{}^4{}_4\, dr = \left(\frac{Q^2}{b}\right)\frac{\pi}{\sqrt{2}}\left(\frac{59}{64}\right), \tag{3.34}$$

where $b^4 = 8l^4$. Thus we get that the total mass is

$$m_{\text{tot}} = \left(\frac{59}{64}\right)\left(\frac{\pi}{\sqrt{2}}\right)\left(\frac{Q^2}{b}\right). \tag{3.35}$$

and the Newtonian mass is

$$m_{\text{N}} = \left(\frac{7}{32}\right)\left(\frac{\pi}{\sqrt{2}}\right)\left(\frac{Q^2}{b}\right). \tag{3.36}$$

Thus,

$$m_{\text{N}}/m_{\text{tot}} = \tfrac{14}{59}. \tag{3.37}$$

Equation (3.37) implies that asymptotically we see only $[\frac{14}{59}]$ of the total energy as a Newtonian gravitational mass. Let us divide the total energy into two parts: Newtonian and electromagnetic. That is

$$m_{\text{tot}} = m_{\text{N}} + m_{\text{em}}. \tag{3.38}$$

One gets

$$m_{\text{em}} = \frac{\pi}{\sqrt{2}}\left(\frac{Q^2}{b}\right)\left(\frac{45}{64}\right). \tag{3.39}$$

This energy could be treated as the energy of the electric field of the charge Q distributed over a sphere of radius r_0. That is,

$$c^2 m_{\text{em}} = Q^2/r_0, \tag{3.40}$$

so that

$$r_0 = b\frac{r^2}{\pi}\left(\frac{64}{45}\right) = l^4\sqrt{2}\left(\frac{128}{45\pi}\right). \tag{3.41}$$

Let us suppose that the Newtonian mass is the mass of an electron.

$$m_{\text{N}} = m_{\text{e}}. \tag{3.42}$$

One gets

$$m_{\text{e}}c^2 = \left(\frac{Q^2}{b}\right)\left(\frac{\pi}{\sqrt{2}}\frac{7}{32}\right). \tag{3.43}$$

Thus we get

$$l = \left(\frac{7}{64}\right)\left(\frac{\pi}{\sqrt{2}}\right)\left(\frac{e^2}{m_{\text{e}}c^2}\right), \tag{3.44}$$

where e is an elementary charge. For r_0 we get similarly

$$r_0 = (\tfrac{14}{45})(e^2/m_{\text{e}}c^2). \tag{3.45}$$

The classical radius of an electron is defined as

$$r_{\text{cl}} = e^2/m_{\text{e}}c^2 \simeq 2.81\times 10^{-13}\ \text{cm}. \tag{3.46}$$

Thus we get

$$r_0 = (\tfrac{14}{45})r_{\text{cl}} \simeq 10^{-13}\ \text{cm} \tag{3.47}$$

and

$$l = \left(\frac{7}{64}\right)\left(\frac{\pi}{\sqrt[4]{2}}\right)r_{\text{cl}} \simeq 10^{-13}\ \text{cm}. \tag{3.48}$$

Let us introduce the dimensionless variables

$$q \equiv Q/b = Q/\sqrt[4]{8}l, \tag{3.49}$$

$$R \equiv r/b = r/\sqrt[4]{8}l. \tag{3.50}$$

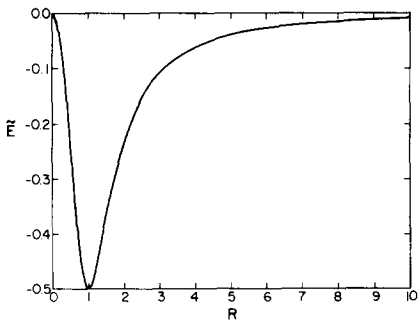

FIG. 2. The function $\tilde{E} = \tilde{E}(R)$ vs R (normalized electric field).

Using Eqs. (3.49) and (3.50) we have

$$\alpha^{-1} = \left(1 - \frac{7}{8\sqrt{2}}\left(\frac{\pi}{2}\right)\frac{q}{R} + \frac{q^2}{R^2}\tilde{g}\left(\frac{1}{R}\right)\right) = (1 - q^2 P(R)), \tag{3.51}$$

$$E = -\frac{q^2}{R^2}\left(\frac{R^4}{R^4+1}\right) = q^2\tilde{E}, \tag{3.52}$$

$$e = 4\pi T^{em}_{44} r^2 = \frac{q^2 R^2}{2}\frac{(R^4 - \frac{5}{4})}{(R^4+1)^2} = q^2\tilde{e}, \tag{3.53}$$

$$\rho_R = \frac{4\pi\rho r^2}{8l^4} = \frac{4\pi\rho r^2}{8l^4} = \frac{4\pi\rho r^2}{b^4} = -\frac{2q}{R}\left(\frac{R^4}{R^4+1}\right) = q\tilde{\rho}_R, \tag{3.54}$$

where q is a normalized charge, R is a normalized radial coordinate, and $\tilde{E}, \tilde{e}, \tilde{\rho}_R$ are normalized, electric field, radial energy distribution, and radial charge distribution, respectively. These functions are plotted in Figs. 2–4. The function

$$P(R) = \frac{1}{R}\left(-q\left(\frac{1}{R}\right) + \frac{7\pi}{16\sqrt{2}}\right) \tag{3.55}$$

is plotted in Fig. 5. It expresses the properties of the generalized Newtonian potential for our solution. Notice that the function $e \leqslant 0$ for $0 \leqslant R \leqslant \sqrt[4]{5}/\sqrt{2}$. This means that our solution corresponds to a kind of bounded system of gravitational and electromagnetic fields.

An interesting question which we can pose here concerns the existence of event horizons. This problem reduces to finding real roots for the function $\alpha^{-1} = f(R,q)$. This depends of course on the value of the parameter q. Let us consider the function

$$f(R,q) = 1 + q^2(1/R)(-7\pi/16\sqrt{2} + g(1/R)). \tag{3.56}$$

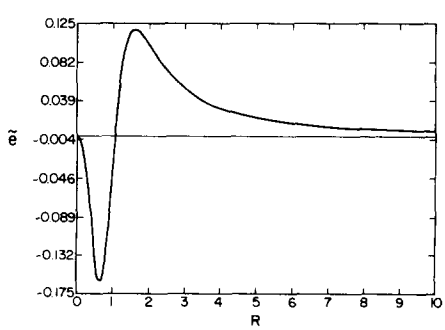

FIG. 3. The function $\tilde{e} = \tilde{e}(R)$ vs R (normalized radial energy distribution).

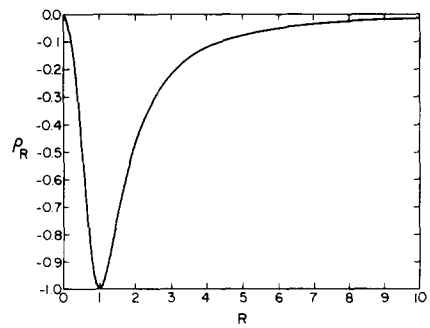

FIG. 4. The function $\tilde{\rho}_R = \tilde{\rho}_R(R)$ vs R (normalized radial charge distribution).

We have

$$f(0,q) = 1 \tag{3.57a}$$

and

$$\lim_{R\to\infty} f(R,q) = 1. \tag{3.57b}$$

Consider now the function

$$h(x) = -7\pi/16\sqrt{2} + g(x) \tag{3.58}$$

and look for a value of $x = x_1$ such that

$$h(x_1) < 0. \tag{3.59}$$

The function $g(x)$ is monotonic in the interval $(0, +\infty)$ and positive. Moreover,

$$\lim_{x=\infty} g(x) = \frac{7\pi}{16\sqrt{2}} \tag{3.60}$$

so that

$$g(1/R) < 7\pi/16\sqrt{2}. \tag{3.61}$$

Consequently,

$$h(1/R_1) < 0 \tag{3.62}$$

for every $R_1 > 0$. Let us suppose that

$$q > \frac{\sqrt{R_1}}{\sqrt{-g(1/R_1) + 7\pi/16\sqrt{2}}}. \tag{3.63}$$

It is easy to check that if (3.63) is satisfied then

$$f(q,R_1) < 0. \tag{3.64}$$

The function $f(q,R)$ changes sign in the interval $\langle 0,R_1\rangle$. This means that there exists a value $R_H \in \langle 0,R_1\rangle$ such that

$$f(q,R_H) = 0. \tag{3.65}$$

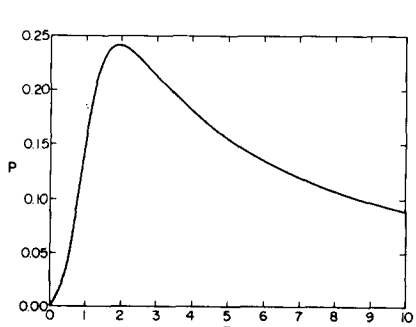

FIG. 5. The function $p = p(R)$ vs R (generalized Nördström function).

The function $f(q,R)$ changes sign in the interval $\langle R_1, +\infty)$ too. Thus there exists a value $R_{\bar{H}} \in \langle R_1, +\infty)$ such that

$$f(q,R_{\bar{H}}) = 0 \tag{3.66}$$

[if condition (3.67) is satisfied]. Hence there are two event horizons for sufficiently large q in general.

Let us examine the situation with only one event horizon. The conditions necessary for the existence of a single horizon are

$$f(q,R) = 0, \tag{3.67a}$$

$$\frac{df}{dR}(q,R) = 0. \tag{3.67b}$$

From (3.67b) one easily gets

$$\frac{1d}{R\,dr} g\left(\frac{1}{R}\right) = g\left(\frac{1}{R}\right) - \frac{7\pi}{16\sqrt{2}}. \tag{3.68}$$

Equation (3.67) is equivalent to

$$\frac{7\pi}{16\sqrt{2}} - \frac{R(R^4 + \frac{1}{2})}{(R^4+1)^2} = g\left(\frac{1}{R}\right). \tag{3.69}$$

In terms of the variable $x \equiv 1/R$ we have

$$\frac{7\pi}{16\sqrt{2}} - \frac{(x^4+2)x}{2(x^4+1)^2} = g(x). \tag{3.70}$$

The soluiton x_0 of Eq. (3.70) is

$$x_0 = 0.516\,288\,994\,64\cdots. \tag{3.71}$$

Let us solve Eq. (3.67a) with respect to q. One gets

$$q_0 = \frac{1}{\sqrt{x_0(7\pi/16\sqrt{2} - g(x_0))}} \tag{3.72}$$

or

$$q_0 = \frac{x_0(x_0^4+1)\sqrt{2x_0}}{\sqrt{x_0^4+2}}. \tag{3.73}$$

Thus there is exactly one event horizon when

$$R_H = 1/x_0 \simeq 1.9369\cdots \tag{3.74}$$

and

$$\left(\frac{r_H}{l}\right) = \frac{\sqrt[4]{8}}{x_0} \simeq 3.2575\cdots, \tag{3.73a}$$

$$q_0 = \frac{x_0(x_0^4+1)\sqrt{2x_0}}{\sqrt{x_0^4+2}} \simeq 2.038\,6231\cdots. \tag{3.75}$$

In this case we have for the Newtonian and total mass,

$$m_{\mathrm{N}}^0 = \pi^4\sqrt{2}\left(\frac{7}{16}\right)\frac{x_0^3(x_0^4+1)^2}{(x_0^4+2)}\left(\frac{c^2 l}{G}\right), \tag{3.76}$$

$$m_{\mathrm{tot}}^0 = \pi^4\sqrt{2}\left(\frac{59}{32}\right)\frac{x_0^3(x_0^4+1)^2}{(x_0^4+2)}\left(\frac{c^2 l}{G}\right), \tag{3.77}$$

or

$$m_{\mathrm{N}}^0 = 3.39\left(\frac{c^2 l}{G}\right) \cong 10^7\ \mathrm{g}, \tag{3.76a}$$

$$m_{\mathrm{tot}}^0 = 14.31\left(\frac{c^2 l}{G}\right) \cong 10^7\ \mathrm{g} \tag{3.77a}$$

for $l = 10^{-20}$ cm. The total charge is

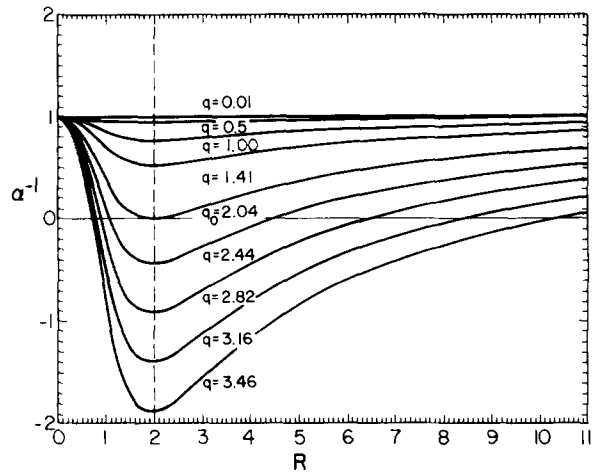

FIG. 6. The function $\alpha^{-1} = f(q,R)$ vs R for various values of parameters q. q_0 means the critical value for which we have only one event horizon for the value $R = R_H$. For the value $R = R_H$ the function $f(q,R)$ has a minimum regardless the value of q. If $q > q_0$ we have two event horizons [two real roots of $f(q,R)$, R_{H_1}, R_{H_2}]. If $q < q_0$ there are not any event horizons [no real roots for $f(q,R)$].

$$Q_0 = q_0{}^4\sqrt{8}lc^2/\sqrt{G} \simeq 2.82(lc^2/\sqrt{G})$$
$$\simeq 10^5\ \mathrm{esu} \simeq 10^{14}\ \text{elementary charges}$$
$$(\text{for } l \simeq 10^{-20}\ \mathrm{cm}). \tag{3.78}$$

It is easy to see that if $q > q_0$ we hve two horizons. This also implies that

$$m_{\mathrm{N}} > m_{\mathrm{N}}^0. \tag{3.79}$$

In other words the Newtonian mass is large enough to form event horizons. If $q = q_0$ we have only one horizon and if $q < q_0$ we have no horizons. This situation is described in Fig. 6 where we plot the function $\alpha^{-1} = f(q,R)$ for various values of the parameter q. For example for an electron one has

$$q_{\mathrm{electron}} = e\sqrt{G}/{}^4\sqrt{8}lc^2 \simeq 10^{-37} \ll q_0. \tag{3.80}$$

Thus there are no event horizons. It is worth noting that if there exists only one event horizon the solution is unstable due to pair creation and Hawking radiation. Such "black holes" are "very hot" (see Ref. 23) and decay very quickly. In the case of two event horizons the solution is unstable because of pair creation. If the Newtonian mass is sufficiently big this solution could be more stable because the Hawking effect is not important for very massive black holes (see Ref. 23). The situation without any event horizons is very interesting from a physical point of view, because it corresponds to the parameter q for electron (in general for any elementary particle). Thus we have in this case a singularity without a horizon. The structure of this singularity is different from the Nördström-like or Schwarzchild-like singularity in the nonsymmetric theory of gravitation [see Refs. 15, 16, and 23 and Eq. (3.23)].

To summarize, we have found the following exact solution (in the form suggested in Sec. 6 of Ref. 1):

$$g_{\mu\nu} = \begin{pmatrix} -\alpha & 0 & 0 & l^2/r^2 \\ 0 & -r^2 & 0 & 0 \\ 0 & 0 & -r^2\sin\theta & 0 \\ -l^2/r^2 & 0 & 0 & \gamma \end{pmatrix}, \tag{3.81}$$

$$\alpha = \left(1 - \frac{7\pi}{16\sqrt{2}}\left(\frac{Q^2}{b}\right)\frac{1}{r} + \frac{Q^2}{rb}g\left(\frac{b}{r}\right)\right)^{-1}, \quad (3.82)$$

$$\gamma = \left(1 + \frac{l^4}{r^4}\right)\left(1 - \frac{7\pi}{16\sqrt{2}}\left(\frac{Q^2}{b}\right)\frac{1}{r} + \frac{Q^2}{rb}g\left(\frac{b}{r}\right)\right), \quad (3.82a)$$

$$b^4 = 8l^4, \quad (3.83)$$

$$E = -\frac{Q}{r^2}\left(\frac{r^4}{r^4 + 8l^4}\right). \quad (3.84)$$

The function g is plotted on Fig. 1 [see Eq. (3.13)]. The solution has one horizon if

$$Q = Q_0 = 2.82(lc^2/\sqrt{G}). \quad (3.85)$$

If $Q < Q_0$ there no horizons. If $Q > Q_0$ we have two horizons (as for the Nördström solution to the Einstein–Maxwell equations).In other words, the horizons exist if the mass is sufficiently big [see Eq. (3.79)]. Finally let us calculate the ratio Q/m_N for our solution. One gets using (3.36) and (3.49)

$$Q/m_N = 32\sqrt{2G}/7\pi q. \quad (3.86)$$

However, for an electron,

$$\frac{e}{m_e} = \frac{32\sqrt{2G}}{7\pi q_{\text{electron}}} \quad (3.87)$$

so that

$$\frac{Q}{m_N} = \left(\frac{q_{\text{electron}}}{q}\right)\left(\frac{e}{m_e}\right). \quad (3.88)$$

4. CONCLUSIONS AND PROSPECTS

We have found an exact static, spherically symmetric solution for the nonsymmetric Kaluza–Klein theory.[1,3] Our solution has the following properties: The metric (symmetric part of $g_{\alpha\beta}$) behaves asymptotically like the Reissner–Nördström solution of general relativity [apart from a factor of $(1 + l^4/r^4)$ which is typical in the nonsymmetric gravitational theory[15,16]]. The most remarkable feature of this metric is that the function α is not singular at $r = 0$ and goes to 1 as $r \to 0$. We have calculated the total energy of the solution and its Newtonian mass. Both quantities are constructed from Q and l, the charge and fermion number parameters respectively.[10] The electric field in our solution asymptotically behaves like the Coulomb field generated by a charge Q. However, this field vanishes at $r = 0$ and is nonsingular for all r. We get a maximal value of this field similar to the one in Born–Infeld electrodynamics.[14] We calculated the charge distribution for such a field and showed that it is nonsingular and equal to zero at $r = 0$. Asymptotically our solution behaves similarly to the Reissner–Nördström-like solution in NGT.[16] Although asymptotically we see a Newtonian mass and an electric charge, at the origin ($r = 0$) there is no mass or electric charge (only fermion charge l). Thus it seems that we get "mass" without mass and "charge" without charge. The total charge for our solution is the same as the Coulomb charge (charge seen at infinity). The total mass, on the other hand, is not the same as the Newtonian mass (mass seen at infinity). In this sense we get a kind of finite mass renormalization. If we consider this solution to be a model for a charged particle constructed from gravitational and electromagnetic fields, this mass renormalization is understandable. The Newtonian mass is the mass of the particle and the remainder is the mass of the external electric field. For example, if we consider this solution as a model of an electron we get a connection between the classical radius of an electron and its fermion number parameter l. Note that in general relativity the total energy associated with the electric field of a pointlike electron is infinite.

Our solution possesses a singularity at $r = 0$ in the determinant of the full nonsymmetric metric. However, the (symmetric) metric seems to be less singular. There is no singularity for the function α. The function γ has a singularity only in the factor $(1 + l^4/r^4)$ and the function $\omega = l^2/r^2$ has the usual singularity at $r = 0$. The electric field is not singular. Our solution posseses one or two event horizons if the charge Q (and consequently the Newtonian mass) is sufficiently large. The solution seems to represent a bounded system of gravitational and electromagnetic fields [c.f. the behavior of the function $\tilde{e}$ (see Fig. 3)]. The radial energy density is zero at the origin, and finite everywhere. In a small region around $r = 0$ it is negative. The metric is spatially flat at the origin. For a very small value of the parameter q (see Fig. 6) the function $\alpha \simeq 1$, and $\gamma = (1 + l^4/r^4)$. If the parameter q is equal to q_{electron}, one gets

$$1 \geqslant \alpha^{-1} = (1 - q^2_{\text{electron}} P(R)) \geqslant (1 - q^2_{\text{electron}} P_{\max}) \geqslant 1 - 10^{-74} \simeq 1. \quad (4.1)$$

Thus α is almost exactly one and γ is almost exactly $(1 + l^4/r^4)$. The metric is then as follows:

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & l^2/r^2 \\ 0 & -r^2 & 0 & 0 \\ 0 & 0 & -r^2\sin^2\theta & 0 \\ -l^2/r^2 & 0 & 0 & (1 + l^4/r^4) \end{pmatrix}. \quad (4.2)$$

The symmetric part of this metric is spatially flat. It is easy to see that such behavior is valid for every elementary particle. The remarkable property of (4.2) is that it is described completely by the parameter l (fermion number) which plays the role of the second gravitational charge in the nonsymmetric theory of gravitation. It seems that the fermion number parameter should play a significant role in the unification of elementary particle theory and gravity. In Eq. (4.2) the fermion number parameter is much more important than mass. Thus the geometry of space-time on the level of elementary particles is determined by the second gravitational charge. The function α^{-1} in general relativity has the form

$$\alpha^{-1} = 1 - 2m/r. \quad (4.3)$$

This function describes the difference between the Schwarzschild solution and a Minkowski metric; in particular the curvature of a space. In the solar system at the earth's orbit one finds

$$\alpha^{-1}(1\text{ au}) \cong 1 - 3 \times 10^{-8}, \quad (4.4)$$

where $1\text{ au} = 1.45 \times 10^8$ km, is one astronomical unit (the radius of earth's orbit) and we have put into Eq. (4.3)

$$2m \cong 5\text{ km} \quad (4.5)$$

which is the Schwarzschild radius of the sun. If we compare

Eq. (4.4) with Eq. (4.1) we easily see that our solution with $q = q_{\text{electron}}$ is spatially much more flat *everywhere* than 3-space at the orbit of the earth.

Note that in Eq. (4.2) we get in a natural constant l which has the dimension of length. Some authors claim that it is impossible to get a true unification of the gravitational field and elementary particles without a new universal constant dimensions of length. In the nonsymmetric theory of gravitation there exists such a constant connected to fermion number. The nonsymmetric Kaluza–Klein theory which unifies the nonsymmetric theory of gravitation with a gauge field theory (i.e., the a electromagnetic field), possesses this constant as well.[1-7] This fact might enable these investigations to lead ultimately to a true unification of gravity and elementary particles.

Here are some prospects for further investigation:

1. Find more general spherical solutions with nonzero f and B_0, including nonstatic solutions.

2. Find axially symmetric solutions of the field equations. This is more difficult, because there is no known axially symmetric solution in the Einstein unified field theory and in NGT.

3. Extend our formalism to the nonabelian-nonsymmetric Kaluza–Klein theory (see Refs. 2 and 6), i.e., to find such a solution for the case $G = \mathrm{SU}(2)$ and $G = \mathrm{SU}(2)\times\mathrm{U}(1)$. This will offer a model of an electron or a lepton constructed from gravitational, electromagnetic, and weak interactions.

4. Extend our solution for the nonsymmetric Jordan–Thiry theory (see Ref. 4).

ACKNOWLEDGMENTS

One of us (M.W.K.) would like to thank Professor M. W. Moffat and Dr. R. B. Mann for their kind hospitality and numerous extremely valuable discussions during my stay at the Physics Department of the University of Toronto.

APPENDIX A

Using Eqs. (2.9) and (2.11) from Ref. 17 and the equation

$$\frac{\omega^2}{\alpha\gamma-\omega^2} = \frac{l^4}{\beta^2+f^2}, \tag{A1}$$

one gets

$$\begin{aligned} A_{11}(\bar{\Gamma}) = &-\frac{1}{2}\phi'' - \frac{1}{8}\{(\phi')^2+4C^2\} + \frac{\alpha'}{4\alpha}\phi' \\ &+ \frac{\omega^2}{8\gamma^2}(3(\dot\phi)^2+4D^2) + \left(\frac{\omega^2}{2\alpha\gamma}\phi' + \frac{\gamma'}{2\gamma}\right)\left(\frac{\alpha'}{2\alpha} - \frac{\omega^2}{2\alpha\gamma}\phi' - \frac{\gamma'}{2\gamma}\right) \\ &- \frac{\partial}{\partial r}\left(\frac{\omega^2}{2\alpha\gamma}+\frac{\gamma'}{2\gamma}\right) + \frac{\partial}{\partial t}\left(\frac{\omega^2}{\gamma^2}\dot\phi + \frac{\dot\alpha}{2\gamma}\right) \\ &+ \left(\frac{\omega^2}{\gamma^2}\dot\phi + \frac{\dot\alpha}{2\gamma}\right)\left(\frac{\dot\gamma}{2\gamma} - \frac{\omega^2}{2\alpha\gamma}\dot\phi - \frac{\dot\alpha}{2\alpha} + \frac{1}{2}\dot\phi\right), \end{aligned} \tag{A2}$$

$$\begin{aligned} A_{44}(\bar{\Gamma}) = &-\frac{1}{2}\ddot\phi - \frac{1}{8}((\dot\phi)^2+4D^2) + \frac{\dot\gamma}{4\gamma}\dot\phi \\ &+ \frac{\omega^2}{8\alpha^2}(3(\phi')^2+4C^2) + \left(\frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha}\right)\left(\frac{\gamma'}{2\alpha\gamma}\dot\phi - \frac{\omega^2}{2\alpha\gamma}\dot\phi\frac{\dot\alpha}{2\alpha}\right) \\ &- \frac{\partial}{\partial t}\left(\frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha}\right) + \left(\frac{\alpha'}{2\alpha} - \frac{\omega^2}{2\alpha\gamma}\phi' - \frac{\gamma'}{2\gamma}\right) + \frac{\partial}{\partial r}\left(\frac{\omega^2}{\alpha^2}\phi' + \frac{\gamma'}{2\alpha}\right), \end{aligned} \tag{A3}$$

$$\begin{aligned} A_{22}(\bar{\Gamma}) = &\left[\left(\frac{2fC-\beta\phi'}{4\alpha}\right) + \frac{(2fC-\beta\phi')}{8\alpha}\frac{\partial}{\partial r}\log(\omega^2(\beta^2+f^2)) + \frac{B(f\phi'+2\beta C)}{4\alpha} + 1 - \frac{\partial}{\partial t}\left(\frac{2fD-\beta\dot\phi}{4\gamma}\right)\right. \\ &\left. - \frac{(2fD-\beta\dot\phi)}{8\gamma}\frac{\partial}{\partial t}\log(\omega^2(\beta^2+f^2)) - \frac{D}{4\gamma}(f\dot\phi+2\beta D)\right] \\ = &\frac{1}{\sin^2\theta}A_{33}(\bar{\Gamma}), \end{aligned} \tag{A4}$$

$$\begin{aligned} A_{(14)}(\bar{\Gamma}) = &\frac{\partial}{\partial r}\left(\frac{\omega^2}{4\alpha\gamma}\dot\phi + \frac{\dot\alpha}{4\alpha} - \frac{\dot\gamma}{4\gamma} - \frac{1}{4}\dot\phi\right) + \frac{\partial}{\partial t}\left(\frac{\omega^2}{4\alpha\gamma}\phi' + \frac{\gamma'}{4\gamma} - \frac{\alpha'}{4\alpha} - \frac{1}{4}\phi'\right) \\ &+ \frac{1}{2}\phi'\left(\frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha} - \frac{1}{4}\dot\phi\right) - \left(\frac{\omega^2}{\gamma^2}\dot\phi + \frac{\dot\alpha}{2\gamma}\right)\left(\frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha}\right) \\ &+ \left[\frac{\omega^2}{2\alpha\gamma}\phi' + \frac{\gamma'}{2\gamma}\right]\left(\frac{1}{2}\dot\phi + \frac{\omega^2}{2\alpha\gamma}\dot\phi + \frac{\dot\alpha}{2\alpha}\right) + \frac{\omega^2}{2\alpha\gamma}\phi'\cdot\dot\phi - \frac{DC}{2l^4}\frac{(\beta^2+f^2)}{\alpha\gamma}, \end{aligned} \tag{A5}$$

$$A_{[23]}(\bar{\Gamma}) = \sin\theta\left(\left(\frac{f\phi' - 2\beta C}{4\alpha}\right)' - \frac{C}{4\alpha}(2fC - \beta\phi') + \frac{1}{8\alpha}(f\phi' + 2\beta C)\left(\frac{\alpha'}{\alpha} + \frac{\omega^2}{\alpha\gamma}\phi' + \frac{\gamma'}{\gamma}\right)\right.$$
$$\left. + \frac{1}{8\gamma}(f\dot{\phi} + 2\beta D)\left(\frac{\dot{\gamma}}{\gamma} + \frac{\omega^2}{2\alpha\gamma}\dot{\phi} + \frac{\dot{\alpha}}{2\alpha}\right) - \frac{\partial}{\partial t}\left(\frac{f\dot{\phi} + 2\beta D}{4\gamma}\right) + \frac{D}{4\gamma}(2fD - \beta\dot{\phi})\right), \tag{A6}$$

where

$$\phi = \log(\beta^2 + f^2), \tag{A7}$$

$$C = \frac{f\beta' - \beta f'}{\beta^2 + f^2}, \quad D = \frac{\dot{\beta}f - \beta\dot{f}}{\beta^2 + f^2}, \tag{A8}$$

· means derivative with respect to time t, and ′ means derivative with respect to radius r. (A9)

$$A_{[14]}(\bar{\Gamma}) = \frac{\omega}{8\alpha}((\phi')^2 + 4C^2) - \frac{\omega}{8\gamma}((\dot{\phi})^2 + 4D^2) + \frac{\omega^2}{4\alpha}\phi'(\phi' + \dot{\phi}) - \frac{1}{2}\frac{\partial}{\partial t}\left(\dot{\phi}\frac{\omega}{\gamma}\right) \tag{A10}$$

APPENDIX B

Using condition (3.1) in the static case and the following ideas from Ref. 17 we get from (A2)–(A4) and from Eqs. (3.4) in the static case,

$$-\frac{1}{\alpha}(A_{11}(\bar{\Gamma}) - 8\pi T^{\text{em}}_{11}) + \frac{2}{\beta}(A_{22}(\bar{\Gamma}) - 8\pi T^{\text{em}}_{22}) + \frac{1}{\gamma}(A_{44} - 8\pi T^{\text{em}}_{44})$$
$$= -\frac{1}{\alpha}A_{11}(\bar{\Gamma}) + \frac{2}{\beta}A_{22}(\bar{\Gamma}) + \frac{1}{\gamma}A_{44}(\bar{\Gamma}) - \frac{4}{\alpha}\frac{\beta^2}{(\beta^2 + 4l^2)}8\pi T^{\text{em}}_{11} = P. \tag{B1}$$

One gets

$$0 = \frac{1}{\alpha}(A_{11}(\bar{\Gamma}) - 8\pi T^{\text{em}}_{11}) + \frac{1}{2}P$$
$$= \frac{1}{2\alpha}A_{11}(\bar{\Gamma}) + \frac{1}{\beta}A_{22}(\bar{\Gamma}) + \frac{1}{2\gamma}A_{44}(\bar{\Gamma}) - \frac{8\pi}{\alpha}\left(\frac{\beta^2 + 4l^4}{\beta^2 + 4l^4}\right)T^{\text{em}}_{11}, \tag{B2}$$

$$0 = \frac{1}{\beta}(A_{22}(\bar{\Gamma}) - 8\pi T^{\text{em}}_{22}) + \frac{1}{2}P$$
$$= -\frac{1}{2\alpha}A_{11}(\bar{\Gamma}) + \frac{1}{\beta}A_{22}(\bar{\Gamma}) + \frac{1}{2\gamma}A_{44}(\bar{\Gamma}) - \frac{8\pi}{\alpha}\left(\frac{3\beta^2 - 4l^4}{\beta^2 + 4l^4}\right)T^{\text{em}}_{11}, \tag{B3}$$

$$0 = -\frac{1}{\gamma}(A_{44}(\bar{\Gamma} - 8\pi T^{\text{em}}_{44}) + \frac{1}{2}P$$
$$= -\frac{1}{2\alpha}A_{11}(\bar{\Gamma}) + \frac{1}{\beta}A_{22}(\bar{\Gamma}) - \frac{1}{2\gamma}A_{44}(\bar{\Gamma}) - \frac{8\pi}{\alpha}\left(\frac{\beta^2 - 4l^4}{\beta^2 + 4l^4}\right)T^{\text{em}}_{11}, \tag{B4}$$

where

$$8\pi T^{\text{em}}_{11} = \frac{\alpha Q^2}{\beta^2}\left(\frac{\beta^2 - 4l^4}{(\beta^2 + 8l^4)^2}\right). \tag{B5}$$

From Eqs. (B2)–(B4) one gets

$$(1/\alpha)A_{11}(\bar{\Gamma}) + (1/\gamma)A_{44}(\bar{\Gamma}) = 0. \tag{B6}$$

Let us substitute

$$\alpha = \exp(M), \quad \gamma = \exp(N), \tag{B7}$$

where $M = M(r)$ and $N = N(r)$ are real functions of r. From (B5) one gets

$$\frac{M' + N'}{r} + \frac{4}{r^2}H = 0, \tag{B8}$$

where

$$H = (l^4/(l^4 + \beta^2). \tag{B9}$$

Let us take

$$\beta = r^2 \tag{B10}$$

and substitute Eqs. (B8)–(B10) to Eq. (B4). One gets, using Eqs. (B5) and (B6),

$$\frac{d}{dr}(r\exp(-M)) = 1 - \frac{Q^2}{r^2}\frac{(r^2 + 4l^4)}{(r^4 + 8l^4)^2}. \tag{B11}$$

APPENDIX C

Let us calculate the connection $\bar{\Gamma}^{\alpha}_{\beta\gamma}$ and the Christoffel symbols for our solution. One gets (using results from Ref. 17)

$$\bar{\Gamma}^1_{[14]} = 2l^2/\alpha r^3, \ \bar{\Gamma}^2_{33} = -\tfrac{1}{2}\sin 2\theta, \ \bar{\Gamma}^2_{23} = \bar{\Gamma}^3_{23} = \cot\theta,$$
$$\bar{\Gamma}^1_{22} = (1/\sin^2\theta)\bar{\Gamma}^1_{33} = -r/\alpha,$$
$$\bar{\Gamma}^2_{(12)} = \bar{\Gamma}^3_{(13)} = 1/r,$$
$$\bar{\Gamma}^2_{[24]} = \bar{\Gamma}^3_{[34]} = -l^2/\alpha r^3, \tag{C1}$$
$$\bar{\Gamma}^1_{11} = \alpha'/2\alpha,$$
$$\Gamma^1_{44} = \frac{4l^4}{r^5\alpha^2} + \frac{\gamma'}{2\alpha} = \frac{7l^4}{8\alpha^2 r^5} - \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3},$$
$$\bar{\Gamma}^4_{(14)} = \frac{2l^4}{r^5\alpha\gamma} + \frac{\gamma'}{2\gamma} = \frac{3l^4}{2r^5}\left(1 + \frac{l^4}{r^4}\right)^{-1} - \frac{\alpha'}{2\alpha}.$$

The remaining $\bar{\Gamma}$'s are zero. Let us consider the symmetric part of our soluiton, i.e.,

$$g_{(\mu\nu)} = \begin{pmatrix} -\alpha & 0 & 0 & 0 \\ 0 & -r^2 & 0 & 0 \\ 0 & 0 & -r^2\sin^2\theta & 0 \\ 0 & 0 & 0 & \gamma \end{pmatrix}, \tag{C2}$$

where α and γ are given by the formulas (3.81) and (3.81a). One easily finds the determinant

$$\tilde{g} = \det[g_{(\mu\nu)}] = -(1 + l^4/r^4)r^4 \sin^2\theta. \tag{C3}$$

The determinant is not singular at $r = 0$. The inverse tensor for $g_{(\mu\nu)}$,

$$\tilde{g}^{(\mu\alpha)}g_{(\alpha\nu)} = \delta^{\mu}_{\nu}, \tag{C4}$$

is

$$\tilde{g}^{(\mu\nu)} = \begin{pmatrix} -1/\alpha & 0 & 0 & 0 \\ 0 & -1/r^2 & 0 & 0 \\ 0 & 0 & -1/r^2\sin^2\theta & 0 \\ 0 & 0 & 0 & 1/\gamma \end{pmatrix}. \tag{C5}$$

Let us calculate the Christoffel symbols for $g_{(\mu\nu)}$.

$$\begin{pmatrix} \alpha \\ \beta\gamma \end{pmatrix} = \frac{1}{2}\tilde{g}^{(\alpha\mu)}(g_{(\beta\mu),\gamma} + g_{(\gamma\mu),\beta} - g_{(\beta\gamma),\mu}). \tag{C6}$$

One easily finds

$$\begin{aligned} &\begin{pmatrix} 1 \\ 11 \end{pmatrix} = \frac{\alpha'}{2\alpha}, \\ &\begin{pmatrix} 1 \\ 22 \end{pmatrix} = \frac{r}{\alpha}, \\ &\begin{pmatrix} 1 \\ 33 \end{pmatrix} = \frac{r}{\alpha}\sin^2\theta, \\ &\begin{pmatrix} 2 \\ 33 \end{pmatrix} = -\frac{1}{2}\sin 2\theta, \quad \begin{pmatrix} 3 \\ 32 \end{pmatrix} = \cot\theta, \\ &\begin{pmatrix} 1 \\ 44 \end{pmatrix} = \frac{\gamma'}{2\alpha} = \frac{-l^4}{2\alpha^2 r^5} - \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}, \\ &\begin{pmatrix} 2 \\ 21 \end{pmatrix} = \frac{1}{r} = \begin{pmatrix} 3 \\ 31 \end{pmatrix}, \\ &\begin{pmatrix} 4 \\ 41 \end{pmatrix} = -\frac{\gamma'}{2\gamma} = \frac{\alpha'}{2\alpha} + \frac{l^4}{r^5}\left(1 + \frac{l^4}{r^4}\right)^{-1}. \end{aligned} \tag{C7}$$

The remaining Christoffel symbols are zero. Let us write equations of motion for an uncharged test particle for our solution, i.e., equation for geodesics.

$$\frac{d^2x^{\alpha}}{d\tau^2} + \overline{\Gamma}^{\alpha}_{(\beta\gamma)}\frac{dx^{\beta}}{d\tau}\frac{dx^{\gamma}}{d\tau} = 0. \tag{C8}$$

One easily finds, from (C1),

$$\begin{aligned} &\frac{d^2r}{d\tau^2} + \frac{\alpha'}{2\alpha}\left(\frac{dr}{d\tau}\right)^2 + \left(\frac{7l^4}{8\alpha^2r^5} - \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}\right) \\ &\quad\times\left(\frac{dt}{d\tau}\right)^2 - \frac{r}{\alpha}\left[\left(\frac{d\theta}{d\tau}\right)^2 + \sin^2\theta\left(\frac{d\gamma}{d\tau}\right)^2\right] = 0, \\ &\frac{d^2\theta}{d\tau^2} + \frac{2}{r}\left(\frac{dr}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) - \frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right)^2 = 0, \\ &\frac{d^2\phi}{dt^2} + \frac{2}{r}\left(\frac{dr}{d\tau}\right)\left(\frac{d\phi}{d\tau}\right) + 2\cot\theta\left(\frac{d\phi}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) = 0, \\ &\frac{d^2t}{d\tau^2} + \left(\frac{3l^4}{2r(l^4 + r^4)} - \frac{\alpha'}{2\alpha}\right)\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right) = 0. \end{aligned} \tag{C9}$$

In the nonsymmetric theory of gravitation uncharged particles move along geodesics in Riemanian geometry formed from $g_{(\mu\nu)}$ (see Ref. 13), i.e., in Christoffels' symbols

$$\frac{d^2x^{\alpha}}{d\tau^2} + \begin{pmatrix} \alpha \\ \beta\gamma \end{pmatrix}\frac{dx^{\gamma}}{d\tau}\frac{dx^{\beta}}{d\tau} = 0. \tag{C10}$$

One easily finds, from (C7),

$$\begin{aligned} &\frac{d^2r}{d\tau^2} + \frac{\alpha'}{2\alpha}\left(\frac{dr}{d\tau}\right)^2 - \left(\frac{l^4}{2\alpha^2r^5} + \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}\right)\left(\frac{dt}{d\tau}\right)^2 \\ &\quad + \frac{r}{\alpha}\left(\left(\frac{d\theta}{d\tau}\right)^2 + \sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right) = 0, \\ &\frac{d^2\theta}{d\tau^2} - \frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right)^2 + \frac{2}{r}\left(\frac{d\theta}{d\tau}\right)\left(\frac{dr}{d\tau}\right) = 0, \\ &\frac{d^2\phi}{d\tau^2} + \frac{2}{r}\left(\frac{d\phi}{d\tau}\right)\left(\frac{dr}{d\tau}\right) + 2\cot\theta\left(\frac{d\phi}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) = 0, \\ &\frac{d^2t}{d\tau^2} + \left(\frac{\alpha'}{2\alpha} + \frac{l^4}{r(l^4 + r^4)}\right)\left(\frac{dt}{d\tau}\right)\left(\frac{dr}{d\tau}\right) = 0. \end{aligned} \tag{C11}$$

Let us find equations of motion for a charged test particle. In the nonsymmetric Kaluza–Klein theory one derived such equations, (see Ref. 1)

$$\begin{aligned} &\frac{d^2x^{\alpha}}{d\tau^2} + \overline{\Gamma}^{\alpha}_{(\beta\gamma)}\frac{dx^{\gamma}}{d\tau}\frac{dx^{\beta}}{d\tau} + \left(\frac{q}{m_0}\right) \\ &\quad\times[g^{\alpha\gamma}F_{\gamma\beta} - g^{[\alpha\gamma]}H_{\gamma\beta}]\frac{dx^{\beta}}{d\tau} = 0, \end{aligned} \tag{C12}$$

where q is a charge and m_0 a rest mass of a test particle. Using (C9) and (3.7) one gets

$$\begin{aligned} &\frac{d^2r}{d\tau^2} + \frac{\alpha'}{2\alpha}\left(\frac{dr}{d\tau}\right)^2 + \left(\frac{7l^4}{8\alpha^2r^5} - \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}\right) \\ &\quad\times\left(\frac{dt}{d\tau}\right)^2 - \frac{r}{\alpha}\left(\left(\frac{d\theta}{d\tau}\right)^2 + \sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right) \\ &\quad - \left(\frac{q}{m_0}\right)\frac{Q}{\alpha r^2}\frac{(r^4 + l^4)}{(r^4 + 8l^4)}\left(\frac{dt}{d\tau}\right) = 0, \\ &\frac{d^2\theta}{d\tau^2} + \frac{2}{r}\left(\frac{dr}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) - \frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right) = 0, \\ &\frac{d^2\phi}{d\tau^2} + \frac{2}{r}\left(\frac{dr}{d\tau}\right)\left(\frac{d\phi}{d\tau}\right) + 2\cot\theta\left(\frac{d\phi}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right) = 0, \\ &\frac{d^2t}{d\tau^2} + \left(\frac{3l^4}{2r(l^4 + r^4)} - \frac{\alpha'}{2\alpha}\right)\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right) \\ &\quad + \left(\frac{q}{m_0}\right)\left(\frac{r^2\alpha Q}{r^4 + 8l^4}\right)\left(\frac{dr}{d\tau}\right) = 0. \end{aligned} \tag{C13}$$

In Ref. 3 a different possibility is considered for the equations of motion for a charged test particle.

$$\begin{aligned} &\frac{d^2x^{\alpha}}{d\tau^2} + \begin{pmatrix} \alpha \\ \beta\gamma \end{pmatrix}\left(\frac{dx^{\beta}}{d\tau}\right)\left(\frac{dx^{\gamma}}{d\tau}\right) + \left(\frac{q}{m_0}\right) \\ &\quad\times[g^{\alpha\gamma}F_{\gamma\beta} - g^{[\alpha\gamma]}H_{\gamma\beta}]\frac{dx^{\beta}}{d\tau} = 0. \end{aligned} \tag{C14}$$

Using (C9) and (C11) one finds the equations

$$\begin{aligned} &\frac{d^2r}{d\tau^2} + \frac{\alpha'}{2\alpha}\left(\frac{dr}{d\tau}\right)^2 - \left(\frac{l^4}{2\alpha^2r^5} + \left(1 + \frac{l^4}{r^4}\right)\frac{\alpha'}{2\alpha^3}\right) \\ &\quad\times\left(\frac{dt}{d\tau}\right)^2 + \frac{r}{\alpha}\left[\left(\frac{d\theta}{d\tau}\right)^2 + \sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right] \\ &\quad - \left(\frac{q}{m_0}\right)\frac{Q}{\alpha r^2}\left(\frac{r^4 + l^4}{r^4 8l^4}\right)\frac{dt}{d\tau} = 0, \\ &\frac{d^2\theta}{d\tau^2} - \frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right)^2 + \frac{2}{r}\left(\frac{d\theta}{d\tau}\right)\left(\frac{dr}{d\tau}\right) = 0, \end{aligned}$$

$$\frac{d^2\phi}{d\tau^2}+\frac{2}{r}\left(\frac{d\phi}{d\tau}\right)\left(\frac{dr}{d\tau}\right)+2\cot\theta\left(\frac{d\phi}{d\tau}\right)\left(\frac{d\phi}{d\tau}\right)=0,$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{\alpha'}{2\alpha}+\frac{l^4}{r(l^4+r^4)}\right)\left(\frac{dt}{d\tau}\right)\left(\frac{dr}{d\tau}\right)+\left(\frac{q}{m_0}\right)\left(\frac{r^2\alpha Q}{r^4+8l^4}\right)\left(\frac{dr}{d\tau}\right)=0. \tag{C15}$$

Notice that equations for θ and ϕ are the same in (C9), (C11), (C13), and (C15) regardless of connections and whether the particle is charged or not. For α' we have

$$\alpha'=\frac{\alpha}{r}+\alpha^2\left(\frac{Q^2(r^4+4l^4)}{(r^4+8l^4)^2}-\frac{1}{r}\right), \tag{C16}$$

where α is given by formula (3.81). According to the general properties of the geodetic equations in Einstein's unified theory, nonsymmetric theory of gravitation, and in the nonsymmetric Kaluza–Klein theory, the Eqs. (C9), (C11), (C13), and (C15) have the following first integral (see Refs. 1 and 3):

$$\gamma\left(\frac{dt}{d\tau}\right)^2-\alpha\left(\frac{dr}{d\tau}\right)^2-r^2\times\left[\left(\frac{d\theta}{d\tau}\right)^2+\sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right]=\text{const}. \tag{C17}$$

We can choose const = 1 and

$$\gamma\left(\frac{dt}{d\tau}\right)^2-\alpha\left(\frac{dr}{d\tau}\right)^2-r^2\times\left(\left(\frac{d\theta}{d\tau}\right)^2+\sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2\right)=1. \tag{C18}$$

Let us consider equations for θ and ϕ,

$$\frac{d^2\theta}{d\tau^2}-\frac{\sin 2\theta}{2}\left(\frac{d\phi}{d\tau}\right)^2+\frac{2}{r}\left(\frac{d\theta}{d\tau}\right)\left(\frac{dr}{d\tau}\right)=0, \tag{C19}$$

$$\frac{d^2\phi}{d\tau^2}+2\cot\theta\left(\frac{d\theta}{d\tau}\right)\left(\frac{d\theta}{d\tau}\right)+\frac{2}{r}\left(\frac{d\phi}{d\tau}\right)\left(\frac{dr}{d\tau}\right)=0.$$

One easily finds the first integral of motion of (C19),

$$r^2\left(\left(\frac{d\theta}{d\tau}\right)^2+\sin^2\theta\left(\frac{d\theta}{d\tau}\right)^2\right)=\frac{2E_0}{r^2}, \tag{C20}$$

where

$$E_0=\text{const}. \tag{C21}$$

comparing (C18) and (C20) one gets

$$\gamma\left(\frac{dt}{d\tau}\right)^2-\alpha\left(\frac{dr}{d\tau}\right)^2=1-\frac{2E_0}{r^2}. \tag{C22}$$

Let us consider the second equation of (C19). One easily finds the first integral of motion

$$\frac{d\phi}{d\tau}=\frac{L}{r^2\sin^2\theta}, \tag{C23}$$

where L = const. Comparing (C20) and (C23) one gets

$$\left(\frac{d\theta}{d\tau}\right)^2=\frac{1}{r^4}\left(2E_0-\frac{L^2}{\sin^2\theta}\right). \tag{C24}$$

The first integrals (C20) and (C22) lead to the following simplifications of our equations (C9), (C11), (C13), and (C15):

$$\frac{d^2r}{d\tau^2}+\frac{7l^4}{8r(l^4+r^4)}\left(\frac{dr}{d\tau}\right)^2+\left(\frac{7l^4}{8\alpha r(l^4+r^4)}-\frac{\alpha'}{2\alpha^2}\right)\times\left(1-\frac{2E_0}{r^2}\right)-\frac{2E_0}{\alpha r^3}=0, \tag{C9a}$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{3l^4}{2r(l^4+r^4)}-\frac{\alpha'}{2\alpha}\right)\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right)=0,$$

$$\frac{d^2r}{d\tau^2}-\frac{l^4}{2r(l^4+r^4)}\left(\frac{dr}{d\tau}\right)^2-\left(\frac{l^4}{2r\alpha(l^4+\gamma^4)}+\frac{\alpha'}{2\alpha^2}\right)\left(1-\frac{2E_0}{r^2}\right)+\frac{2E_0}{\alpha r^3}=0,$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{\alpha}{2\alpha}+\frac{l^4}{r(l^4+r^4)}\right)\left(\frac{dt}{d\tau}\right)\left(\frac{dr}{d\tau}\right)=0, \tag{C11a}$$

$$\frac{d^2r}{d\tau^2}+\frac{7l^4}{8r(l^4+r^4)}\left(\frac{dr}{d\tau}\right)^2+\left(\frac{7l^4}{8\alpha r(l^4+r^4)}-\frac{\alpha'}{2\alpha}\right)\left(1-\frac{2E_0}{r^2}\right)-\frac{2E_0}{\alpha r^3}-\left(\frac{q}{m_0}\right)\frac{Q}{\alpha r^2}\left(\frac{r^4+l^4}{r^4+8l^4}\right)\left(\frac{dt}{d\tau}\right)=0,$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{3l^4}{2r(l^4+r^4)}-\frac{\alpha'}{2\alpha}\right)\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right)+\left(\frac{q}{m_0}\right)\left(\frac{r^2\alpha Q}{r^4+8l^4}\right)\left(\frac{dr}{d\tau}\right)=0, \tag{C13a}$$

$$\frac{d^2r}{d\tau^2}-\frac{l^4}{2r(l^4+r^4)}\left(\frac{dr}{d\tau}\right)^2-\left(\frac{l^4}{2r\alpha(l^4+r^4)}+\frac{\alpha'}{2\alpha^2}\right)\left(1-\frac{2E_0}{r^2}\right)+\frac{2E_0}{\alpha r^3}-\left(\frac{q}{m_0}\right)\frac{Q}{\alpha r^2}\left(\frac{r^4+l^4}{r^4+8l^4}\right)\left(\frac{dt}{d\tau}\right)=0,$$

$$\frac{d^2t}{d\tau^2}+\left(\frac{\alpha'}{2\alpha}+\frac{l^4}{r(l^4+r^4)}\right)\left(\frac{dt}{d\tau}\right)\left(\frac{dr}{d\tau}\right)+\left(\frac{q}{m_0}\right)\left(\frac{r^2\alpha Q}{r^4+8l^4}\right)\left(\frac{dr}{d\tau}\right)=0. \tag{C15a}$$

For angular coordinates we have for Eqs. (C9a), (C11a), (C13a), and (C15a) the same equations (C19) and the same first integral of motion (C20), (C22), and (C23).

[1] M. W. Kalinowski, "The nonsymmetric Kaluza–Klein theory," J. Math. Phys. **24**, 1835 (1983).

[2] M. W. Kalinowski, "The nonsymmetric-nonabelian Kaluza–Klein theory," J. Phys. A **16**, 1669 (1983).

[3] M. W. Kalinowski, "Material sources in the nonsymmetric Kaluza–Klein theory," University of Toronto report, September 1982 (to appear in J. Math. Phys., 1984).

[4] M. W. Kalinowski, "The nonsymmetric Jordan–Thiry theory," Can. J. Phys. **61**, 884 (1983).

[5] M. W. Kalinowski, "The nonsymmetric–nonabelian Jordan–Thiry theory," University of Toronto report, September 1982.

[6] M. W. Kalinowski, "Spontaneous symmetry breaking and Higgs' mechanism in the nonsymmetric Kaluza–Klein theory," Ann. Phys. **148**, 214 (1983).

[7] M. W. Kalinowski, "Spontaneous symmetry breaking and Higgs' mechanism in the nonsymmetric Jordan–Thiry theory," University of Toronto report, December 1982.

[8] J. W. Moffat, "New theory of Gravitation," Phys. Rev. D **19**, 3557 (1979).

[9] J. W. Moffat, "Gauge invariance and string interactions in a generalized theory of gravitation," Phys. Rev. D **23**, 2870 (1981).

[10] J. W. Moffat, "Generalized theory of gravitation and its physical consequences," in *Proceedings of the VII International School of Gravitation and Cosmology*, Erice Sicily, edited by V. de Sabatta (World Scientific Publishing, Singapore, 1982), p. 127.

[11] M. W. Kalinowski and R. B. Mann, "Linear approximation in the nonsymmetric Kaluza–Klein theory," University of Toronto report, March 1983.

[12]H. A. Hill, R. J. Bos, and P. R. Goode, "Preliminary determination of the quadrupole moment of the sun from rotational splitting of global oscillations and its relevance to tests of general relativity," Phys. Rev. Lett. **33**, 1497 (1983).
[13]J. W. Moffat, "Consequences of a new experimental determination of the quadrupole moment of the sun for gravitational theory," Phys. Rev. Lett. **50**, 709 (1983).
[14]M. Born and L. Infeld, "Foundations of the new field theory," Proc. Roy. Soc. London, Ser. A **144**, 425 (1934).
[15]J. W. Moffat and D. H. Boal, "Solutions in the nonsymmetric unified field theory," Phys. Rev. D **11**, 1375 (1975).
[16]J. W. Moffat, "Static spherically symmetric solution for the field of a charged particle in a theory of gravity," Phys. Rev. D **19**, 3562 (1978).
[17]D. N. Pant, "Spherically Symmetric Rigorous Solutions in Bonnor's Unified Field Theory," Nuovo Cimento B **25**, 175 (1975).
[18]A. Papapetrou, "Static spherically symmetric solutions in the unitary field theory," Proc. Roy. Irish Acad. **52**, 69 (1948).
[19]M. Wyman, "Unified field theory," Can. J. Math. **2**, 427 (1950).
[20]W. B. Bonnor, "The general static spherically symmetric solution in Einstein's unified field theory," Proc. Roy. Soc. **210**, 427 (1952).
[21]W. B. Bonnor, "Static spherically symmetric solutions in Einstein's unified field theory," Proc. Roy. Soc. **209**, 353 (1951).
[22]J. R. Vanstone, "The general static spherically symmetric solution of the 'weak' unified field theory," Can. J. Math. **14**, 568 (1962).
[23]L. Campbell and J. W. Moffat, "Black Holes in the Nonsymmetric Theory of Gravitation," University of Toronto Report, August 1982.

Solution of multidimensional inverse transport problems[a)]

Edward W. Larsen
University of California, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 19 May 1983; accepted for publication 29 July 1983)

Formulas are derived for energy-dependent, steady-state, and time-dependent neutron transport problems, relating the surface neutron fluxes for a convex, homogeneous, three-dimensional region to the neutron scattering laws that apply within the region. In principle, these formulas can be used to deduce information about the scattering laws.

PACS numbers: 05.60. + w, 42.68.Db

I. INTRODUCTION

In recent years, a substantial effort has been directed toward the problem of obtaining exact formulas relating incoming and exiting neutron fluxes for a homogeneous slab to the scattering laws that apply within the slab.[1-13] Such formulas have generally been obtained by directly manipulating the forward and adjoint one-dimensional slab geometry transport equations, although there are exceptions; some early work of Siewert[1,2] makes use of the Chandrasekhar X and Y functions; recent work by Sanchez and McCormick[11] uses the diffusion equation as an approximation to the transport equation; and a recent article by Siewert and Dunn[9] allows for spatial variations in the angular flux in directions parallel to the edges of the slab. Also, most of this prior work considers only monoenergetic transport problems, although Larsen[6] has considered multigroup problems.

In an effort to obtain a more general, and therefore possibly more useful theory, we shall in this article extend the domain of the previous results to the general case of time- and energy-dependent neutron transport in a three-dimensional, convex, homogeneous region. Specifically, for such transport problems we derive exact formulas relating both steady-state and time-dependent surface neutron fluxes to the neutron scattering laws that apply within the region. In principle, these formulas can be used to determine properties of the material scattering laws. However, there are limitations: a large number of neutron flux measurements generally must be made, and the theory described here is only applicable for homogeneous regions.

Our theory thus cannot be used to determine the structure of a heterogeneous solid by irradiating it with external neutrons and measuring (and processing) the incident and exiting fluxes. However, it can be used to solve the following two general problems for a homogeneous region D: (1) If D consists of a uniform mixture of known materials (with known cross sections) in unknown proportions, then determine the proportions; and (2) if the cross sections in D can be regarded as multigroup with a finite number of groups and a finite Legendre expansion in angle, then determine these cross sections.

The remainder of this article is organized as follows. In Sec. II we establish notation and derive physical interpretations for solutions of certain adjoint neutron transport problems. In Sec. III we use these results to derive the inverse theory for steady-state problems; in Sec. IV we repeat this analysis for time-dependent problems. We conclude, in Sec. V, by describing a way to simplify some of the results obtained in Secs. III and IV.

II. PRELIMINARIES

The main purpose of this section is to show that solutions of adjoint transport problems for a convex solid exist having simple interpretations at points on the surface.

To begin, let us assume that steady-state neutron transport occurs within a homogeneous convex region D according to the standard equations

$$\mathbf{\Omega}\cdot\nabla\psi(\mathbf{r},\mathbf{\Omega},E)+\sigma_T(E)\psi(\mathbf{r},\mathbf{\Omega},E)=\int\int\sigma_s(E'\to E,\mathbf{\Omega}'\cdot\mathbf{\Omega})\psi(\mathbf{r},\mathbf{\Omega}',E')d^2\Omega'\,dE', \tag{2.1}$$

$$\psi(\mathbf{r},\mathbf{\Omega},E)=f(\mathbf{r},\mathbf{\Omega},E),\quad \mathbf{r}\in\partial D,\quad \mathbf{\Omega}\cdot\mathbf{n}<0. \tag{2.2}$$

Here $\mathbf{n}$ is the unit outer normal. The solution ψ of problem (2.1), (2.2) is, physically, the neutron angular flux arising from the incident flux f on the surface of D.

To proceed, let R be the set of all phase-space points $(\mathbf{r},\mathbf{\Omega},E)$, with $\mathbf{r}\in\partial D$ and $\mathbf{\Omega}\cdot\mathbf{n}>0$. Let R_0 be any subset of R, and χ_0 the characteristic function for R_0:

$$\chi_0(\mathbf{r},\mathbf{\Omega},E)=\begin{cases}1, & (\mathbf{r},\mathbf{\Omega},E)\in R_0,\\ 0, & (\mathbf{r},\mathbf{\Omega},E)\in R-R_0.\end{cases} \tag{2.3}$$

For any neutron flux $\psi(\mathbf{r},\mathbf{\Omega},E)$ existing in D, we define

$$\int\int\int_R\mathbf{\Omega}\cdot\mathbf{n}\chi_0(\mathbf{r},\mathbf{\Omega},E)\psi(\mathbf{r},\mathbf{\Omega},E)d^2\Omega\,dE\,d^2r=\int\int\int_{R_0}\mathbf{\Omega}\cdot\mathbf{n}\psi(\mathbf{r},\mathbf{\Omega},E)d^2\Omega\,dE\,d^2r$$
$$=\text{the net current out of } R_0. \tag{2.4}$$

Now, let us consider the steady-state adjoint problem

$$-\mathbf{\Omega}\cdot\nabla\psi^*(\mathbf{r},\mathbf{\Omega},E)+\sigma_T(E)\psi^*(\mathbf{r},\mathbf{\Omega},E)=\int\int\sigma_s(E\to E',\mathbf{\Omega}\cdot\mathbf{\Omega}')\psi^*(\mathbf{r},\mathbf{\Omega}',E')d^2\Omega'\,dE', \tag{2.5}$$

$$\psi^*(\mathbf{r},\mathbf{\Omega},E)=\chi_0(\mathbf{r},\mathbf{\Omega},E),\quad r\in\partial D,\quad \mathbf{\Omega}\cdot\mathbf{n}>0. \tag{2.6}$$

We shall prove the following result:

Lemma 1: For any $\mathbf{r}\in\partial D$, $\mathbf{\Omega}\cdot\mathbf{n}<0$, and any E,

[a)] This research was performed under the auspices of the U. S. Department of Energy.

 0022-2488/84/010131-05$02.50

$\psi^*(\mathbf{r}, \mathbf{\Omega}, E)$ = the net current out of R_0 due to a unit delta incident beam at $(\mathbf{r}, \mathbf{\Omega}, E)$.

Proof: Let $\mathbf{r}_0$ be any point on ∂D, $\mathbf{\Omega}_0$ any unit vector such that $\mathbf{\Omega}_0 \cdot \mathbf{n}_0 < 0$, and E_0 any admissible value of E. Also, let ψ be the solution of the forward problem consisting of Eqs. (2.1) and (2.2), with

$$f(\mathbf{r}, \mathbf{\Omega}, E) = \delta(\mathbf{r} - \mathbf{r}_0)\delta(\mathbf{\Omega} - \mathbf{\Omega}_0)\delta(E - E_0)/|\mathbf{\Omega}_0 \cdot \mathbf{n}_0| . \tag{2.7}$$

Then $\psi(\mathbf{r}, \mathbf{\Omega}, E)$ is the angular flux at any point $(\mathbf{r}, \mathbf{\Omega}, E)$ due to the unit delta incident beam f at $(\mathbf{r}_0, \mathbf{\Omega}_0, E_0)$.

We multiply Eq. (2.1) by ψ^* and Eq. (2.5) by ψ, integrate both equations over $\mathbf{\Omega}$ and E, subtract, and then integrate the resulting single equation over all $\mathbf{r} \in D$ to obtain

$$0 = \int_{\partial D} \iint \mathbf{\Omega} \cdot \mathbf{n} \psi^* \psi d^2\Omega \, dE \, d^2r . \tag{2.8}$$

(This is just the reciprocity relation for the special case of no interior sources for the forward and adjoint transport fluxes.[14]) Next, we use Eqs. (2.2), (2.6), and (2.7) to get

$$0 = \iiint_{R_0} \mathbf{\Omega} \cdot \mathbf{n} \psi(\mathbf{r}, \mathbf{\Omega}, E) d^2\Omega \, dE \, d^2r + \int_{\partial D} \iint_{\mathbf{\Omega}\cdot\mathbf{n}<0} \mathbf{\Omega} \cdot \mathbf{n} \psi^* \frac{\delta(\mathbf{r} - \mathbf{r}_0)\delta(\mathbf{\Omega} - \mathbf{\Omega}_0)\delta(E - E_0)}{|\mathbf{\Omega}_0 \cdot \mathbf{n}_0|} \times d^2\Omega \, dE \, d^2r , \tag{2.9}$$

or

$$\psi^*(\mathbf{r}_0, \mathbf{\Omega}_0, E_0) = \iiint_{R_0} \mathbf{\Omega} \cdot \mathbf{n} \psi(\mathbf{r}, \mathbf{\Omega}, E) d^2\Omega \, dE \, d^2r . \tag{2.10}$$

This proves the result. Q.E.D.

Now let us assume that time-dependent neutron transport occurs within the homogeneous convex region D according to the standard equations

$$\frac{1}{v}\frac{\partial}{\partial t}\psi(\mathbf{r}, \mathbf{\Omega}, E, t) + \mathbf{\Omega} \cdot \nabla\psi(\mathbf{r}, \mathbf{\Omega}, E, t) + \sigma_t(E)\psi(\mathbf{r}, \mathbf{\Omega}, E, t) = \iint \sigma_s(E' \to E, \mathbf{\Omega}' \cdot \mathbf{\Omega})\psi(\mathbf{r}, \mathbf{\Omega}', E', t) d^2\Omega' \, dE' , \tag{2.11}$$

$$\psi(\mathbf{r}, \mathbf{\Omega}, E, t) = f(\mathbf{r}, \mathbf{\Omega}, E, t), \quad \mathbf{r} \in \partial D, \ \mathbf{\Omega} \cdot \mathbf{n} < 0, \ 0 < t , \tag{2.12}$$

$$\psi(\mathbf{r}, \mathbf{\Omega}, E, 0) = 0, \quad \mathbf{r} \in D . \tag{2.13}$$

The solution ψ of Eqs. (2.11)–(2.13) is, physically, the time-dependent neutron angular flux arising from the incident flux f on the surface of D. [Throughout this article, we only treat problems with initial data of the form (2.13), i.e., we assume that initially no free neutrons are present in D.]

We let R, R_0, and χ_0 be defined above, and for any neutron flux $\psi(\mathbf{r}, \mathbf{\Omega}, E, t)$ existing in D and $T > 0$, we define

$$\int_0^T \iiint_R \mathbf{\Omega} \cdot \mathbf{n} \chi_0(\mathbf{r}, \mathbf{\Omega}, E)\psi(\mathbf{r}, \mathbf{\Omega}, E, t) d^2\Omega \, dE \, d^2r \, dt = \int_0^T \iiint_{R_0} \mathbf{\Omega} \cdot \mathbf{n} \psi(\mathbf{r}, \mathbf{\Omega}, E, t) d^2\Omega \, dE \, d^2r \, dt = \text{the net current out of } R_0 \text{ up to time } T. \tag{2.14}$$

We now consider the time-dependent adjoint problem

$$-\frac{1}{v}\frac{\partial}{\partial t}\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t) - \mathbf{\Omega} \cdot \nabla\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t) + \sigma_T(E)\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t) = \iint \sigma_s(E \to E', \mathbf{\Omega} \cdot \mathbf{\Omega}')\psi^*(\mathbf{r}, \mathbf{\Omega}', E', t) d^2\Omega' \, dE' , \tag{2.15}$$

$$\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t) = \chi_0(\mathbf{r}, \mathbf{\Omega}, E), \quad \mathbf{r} \in \partial D, \ \mathbf{\Omega} \cdot \mathbf{n} > 0, \ 0 < t < T , \tag{2.16}$$

$$\psi^*(\mathbf{r}, \mathbf{\Omega}, E, T) = 0, \quad \mathbf{r} \in D . \tag{2.17}$$

We shall prove the following result:

Lemma 2: Let $0 < t < T$. Then for any $\mathbf{r} \in \partial D$, $\mathbf{\Omega} \cdot \mathbf{n} < 0$, and any E, $\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t)$ = the net current out of R_0 up to time T due to a unit delta incident beam at $(\mathbf{r}, \mathbf{\Omega}, E, t)$.

Proof: Let $\mathbf{r}_0$ be any point on ∂D, $\mathbf{\Omega}_0$ any unit vector such that $\mathbf{\Omega}_0 \cdot \mathbf{n}_0 < 0$, E_0 any admissible value of E, and $0 < t_0 < T$. Also, let ψ be the solution of the forward problem consisting of Eqs. (2.11)–(2.13), with

$$f(\mathbf{r}, \mathbf{\Omega}, E, t) = \frac{\delta(\mathbf{r} - \mathbf{r}_0)\delta(\mathbf{\Omega} - \mathbf{\Omega}_0)\delta(E - E_0)\delta(t - t_0)}{|\mathbf{\Omega}_0 \cdot \mathbf{n}_0|} . \tag{2.18}$$

Then $\psi(\mathbf{r}, \mathbf{\Omega}, E, t)$ is the time-dependent angular flux at any point $(\mathbf{r}, \mathbf{\Omega}, E, t)$ due to the unit delta incident beam at $(\mathbf{r}_0, \mathbf{\Omega}_0, E_0, t_0)$.

We multiply Eq. (2.11) by ψ^*, Eq. (2.15) by ψ, integrate both equations over $\mathbf{\Omega}$ and E, and subtract to obtain the single equation

$$0 = \frac{\partial}{\partial t}\iint \frac{1}{v}\psi\psi^* d^2\Omega \, dE + \nabla \cdot \iint \mathbf{\Omega}\psi\psi^* \, d^2\Omega \, dE . \tag{2.19}$$

Next, we operate on Eq. (2.19) by

$$\int_0^T \int_D (\cdot) d^3r \, dt , \tag{2.20}$$

and use the initial conditions, Eqs. (2.13), (2.17), and the boundary conditions, Eqs. (2.12), (2.16), and (2.18) to easily obtain

$$\psi^*(\mathbf{r}_0, \mathbf{\Omega}_0, E_0, t_0) = \int_0^T \iiint_{R_0} \mathbf{\Omega} \cdot \mathbf{n} \psi \, d^2\Omega \, dE \, d^2r \, dt . \tag{2.21}$$

This proves the result. Q.E.D.

The main purpose of Lemmas 1 and 2 is to establish the following: (1) there exist solutions ψ^* of the steady-state adjoint transport Eq. (2.5) for which $\psi^*(\mathbf{r}, \mathbf{\Omega}, E)$ is physically measurable for all $\mathbf{r} \in \partial D$, all $\mathbf{\Omega}$, and all E; and (2) there exist solutions ψ^* of the time-dependent adjoint transport Eq. (2.15) and initial condition Eq. (2.17) for which $\psi^*(\mathbf{r}, \mathbf{\Omega}, E, t)$ is physically measurable for all $\mathbf{r} \in \partial D$, all $\mathbf{\Omega}$, all E, and all $t < T$. Such solutions will play a key role in the remainder of this article.

III. STEADY-STATE THEORY

Let ψ be any solution of Eq. (2.1) and ψ^* any solution of Eq. (2.5). We multiply Eq. (2.1) by $\nabla\psi^*$, Eq. (2.5) by $\nabla\psi$, integrate over $\mathbf{\Omega}$ and E, and then add the two resulting equa-

tions, obtaining

$$\iint [(\nabla\psi^*)(\mathbf{\Omega}\cdot\nabla\psi) - (\nabla\psi)(\mathbf{\Omega}\cdot\nabla\psi^*)]d^2\Omega\, dE + \nabla\iint \sigma_T(E)\psi(\mathbf{r},\mathbf{\Omega},E)\psi^*(\mathbf{r},\mathbf{\Omega},E)d^2\Omega\, dE = \nabla\iiiint \sigma_s(E'\to E,\mathbf{\Omega}'\cdot\mathbf{\Omega})\psi(\mathbf{r},\mathbf{\Omega}',E') \times \psi^*(\mathbf{r},\mathbf{\Omega},E)d^2\Omega'\, dE'\, d^2\Omega\, dE\,. \tag{3.1}$$

However, elementary operations give

$$\begin{aligned}(\nabla\psi^*)(\mathbf{\Omega}\cdot\nabla\psi) - (\nabla\psi)(\mathbf{\Omega}\cdot\nabla\psi^*) &= \nabla(\psi^*\mathbf{\Omega}\cdot\nabla\psi) - \mathbf{\Omega}\cdot\nabla(\psi^*\nabla\psi)\\ &= \mathbf{\Omega}\cdot\nabla(\psi\nabla\psi^*) - \nabla(\psi\mathbf{\Omega}\cdot\nabla\psi^*)\,.\end{aligned} \tag{3.2}$$

Introducing Eq. (3.2) into Eq. (3.1) and integrating over $\mathbf{r}$, we obtain

$$\mathbf{S} + \int_{\partial D}\mathbf{n}\iint \sigma_T(E)\psi(\mathbf{r},\mathbf{\Omega},E)\psi^*(\mathbf{r},\mathbf{\Omega},E)d^2\Omega\, dE\, d^2r = \int_{\partial D}\mathbf{n}\iiiint \sigma_s(E'\to E,\mathbf{\Omega}'\cdot\mathbf{\Omega})\psi(\mathbf{r},\mathbf{\Omega}',E') \times \psi^*(\mathbf{r},\mathbf{\Omega},E)d^2\Omega'\, dE'\, d^2\Omega\, dE\, d^2r\,, \tag{3.3}$$

where, using a standard vector identity,[15] we have

$$\mathbf{S} = \int_{\partial D}\iint \psi^*[\mathbf{n}(\mathbf{\Omega}\cdot\nabla\psi) - (\mathbf{\Omega}\cdot\mathbf{n})(\nabla\psi)]d^2\Omega\, dE\, d^2r = \int_{\partial D}\iint \psi^*[\mathbf{\Omega}\times(\mathbf{n}\times\nabla\psi)]d^2\Omega\, dE\, d^2r\,, \tag{3.4a}$$

or

$$\mathbf{S} = \int_{\partial D}\iint \psi[(\mathbf{\Omega}\cdot\mathbf{n})(\nabla\psi^*) - \mathbf{n}(\mathbf{\Omega}\cdot\nabla\psi^*)]d^2\Omega\, dE\, d^2r = \int_{\partial D}\iint \psi[\mathbf{\Omega}\times(\nabla\psi^*\times\mathbf{n})]d^2\Omega\, dE\, d^2r\,, \tag{3.4b}$$

However, if ∇_T denotes the gradient operator in the plane tangent to ∂D, then for any point on ∂D we may use

$$\nabla\psi = \mathbf{n}(\mathbf{n}\cdot\nabla\psi) + \nabla_T\psi \tag{3.5a}$$

in Eq. (3.4a), and

$$\nabla\psi^* = \mathbf{n}(\mathbf{n}\cdot\nabla\psi^*) + \nabla_T\psi^* \tag{3.5b}$$

in (3.4b). Making these substitutions (and noting that $\mathbf{n}\times\mathbf{n} = \mathbf{0}$) we obtain

$$\mathbf{S} = \int_{\partial D}\iint \psi^*[\mathbf{\Omega}\times(\mathbf{n}\times\nabla_T\psi)]d^2\Omega\, dE\, d^2r\,, \tag{3.6a}$$

or

$$\mathbf{S} = \int_{\partial D}\iint \psi[\mathbf{\Omega}\times(\nabla_T\psi^*\times\mathbf{n})]d^2\Omega\, dE\, d^2r\,. \tag{3.6b}$$

Our result is Eq. (3.3) and Eq. (3.6). Each of the terms in these equations consists only of a surface integral involving ψ, ψ^*, $\nabla_T\psi$, or $\nabla_T\psi^*$. Since boundary conditions for ψ and ψ^* have not yet been imposed, we can choose these boundary conditions so that both ψ and ψ^* are physically measurable on ∂D. Doing this, then $\nabla_T\psi$ and $\nabla_T\psi^*$ can also be obtained, and the vector equation (3.3) reduces (for general three-dimensional geometry) to three linear scalar constraints involving σ_T and σ_s. For different combinations of ψ and ψ^*, different constraints are derived, and one can use these constraints to determine properties of σ_s and σ_T, such as described above in Sec. I.

To obtain new constraints on σ_T and σ_s, one does not have to determine new values of both ψ and ψ^*. For instance, one could experimentally determine a specific, unique ψ^*, and then three new constraints are determined by each different value of ψ. Alternatively, one could determine a unique ψ and then derive three different constraints using each different value of ψ^*. [This can easily be done if in evaluating the "first" ψ^* using the theory in Sec. II, one determines the exiting angular fluxes for all points $(\mathbf{r},\mathbf{\Omega},E)\in R_0$. Then, the "first" ψ^* arises from R_0, and arbitrarily many other solutions ψ^* arise from arbitrary subsets of R_0.]

Whichever way one chooses to determine different constraints, it is clear that the experimental determination of the necessary data will require a large number of measurements. In addition, because the problem under consideration is truly inverse in nature, it is likely that our set of constraints will be sensitive to errors in neutron flux measurements. However, only experiment can determine just how accurately the fluxes need to be determined so that errors in measurements of ψ do not lead to unacceptable errors in σ_T or σ_s.

IV. TIME-DEPENDENT THEORY

Let ψ be any solution of Eqs. (2.11) and (2.13), and ψ^* any solution of Eqs. (2.15) and (2.17). We multiply Eq. (2.11) by $\nabla\psi^*$, Eq. (2.15) by $\nabla\psi$, integrate over $\mathbf{\Omega}$ and E, and then add the two resulting equations, obtaining

$$\iint \frac{1}{v}\left[(\nabla\psi^*)\frac{\partial\psi}{\partial t} - (\nabla\psi)\frac{\partial\psi^*}{\partial t}\right]d^2\Omega\, dE + \iint [(\nabla\psi^*)(\mathbf{\Omega}\cdot\nabla\psi) - (\nabla\psi)(\mathbf{\Omega}\cdot\nabla\psi^*)]d^2\Omega\, dE + \nabla\iint \sigma_T\psi\psi^*\, d^2\Omega\, dE = \nabla\iiiint \sigma_s\psi\psi^*\, d^2\Omega'\, dE'\, d^2\Omega\, dE\,. \tag{4.1}$$

Equation (3.2) can be used to rewrite the second term on the left side of Eq. (4.1), while the first term can be rewritten using

$$\begin{aligned}(\nabla\psi^*)\frac{\partial\psi}{\partial t} - (\nabla\psi)\frac{\partial\psi^*}{\partial t} &= \nabla\left(\psi^*\frac{\partial\psi}{\partial t}\right) - \frac{\partial}{\partial t}(\psi^*\nabla\psi)\\ &= \frac{\partial}{\partial t}(\psi\nabla\psi^*) - \nabla\left(\psi\frac{\partial\psi^*}{\partial t}\right).\end{aligned} \tag{4.2}$$

Introducing Eqs. (3.2) and (4.2) into Eq. (4.1), operating by

$$\int_0^T\int_D (\cdot)d^3r\, dt\,,$$

and using the initial conditions (2.13) and (2.17) and the formulas (3.5), we obtain

$$\mathbf{U} + \mathbf{V} + \int_0^T\int_{\partial D}\mathbf{n}\iint \sigma_T(E)\psi(\mathbf{r},\mathbf{\Omega},E,t) \times \psi^*(\mathbf{r},\mathbf{\Omega},E,t)d^2\Omega\, dE\, d^2r\, dt = \int_0^T\int_{\partial D}\mathbf{n}\iiiint \sigma_s(E'\to E,\mathbf{\Omega}'\cdot\mathbf{\Omega})\psi(\mathbf{r},\mathbf{\Omega}',E',t) \times \psi^*(\mathbf{r},\mathbf{\Omega},E,t)d^2\Omega'\, dE'\, d^2\Omega\, dE\, d^2r\, dt\,, \tag{4.3}$$

where

$$\mathbf{U} = \int_0^T \int_{\partial D} \mathbf{n} \iint \frac{1}{v} \psi^* \frac{\partial \psi}{\partial t} \, d^2\Omega \, dE \, d^2r \, dt \tag{4.4a}$$

or

$$\mathbf{U} = -\int_0^T \int_{\partial D} \mathbf{n} \iint \frac{1}{v} \psi \frac{\partial \psi^*}{\partial t} \, d^2\Omega \, dE \, d^2r \, dt \tag{4.4b}$$

and

$$\mathbf{V} = \int_0^T \int_{\partial D} \iint \psi^* [\boldsymbol{\Omega} \times (\mathbf{n} \times \nabla_T \psi)] \, d^2\Omega \, dE \, d^2r \, dt \tag{4.5a}$$

or

$$\mathbf{V} = \int_0^T \int_{\partial D} \iint \psi [\boldsymbol{\Omega} \times (\nabla_T \psi^* \times \mathbf{n})] \, d^2\Omega \, dE \, d^2r \, dt \, . \tag{4.5b}$$

Our result consists of Eqs. (4.3)–(4.5). As with the steady-state analysis, each of the expressions in these equations involving ψ or ψ^* can, in principle, be determined by a suitable interpretation of ψ^* (see Sec. II) together with suitable measurements of surface neutron fluxes. The comments at the end of Sec. III regarding (1) the likely sensitivity of our equations to experimental errors, and (2) the effort that appears necessary to determine acceptable measurements, apply here to an even greater degree than in Sec. III. This is because one must now make accurate measurements for each value of t; therefore, the dimensionality of the space in which measurements must be made, recorded, and processed, is increased by one.

To conclude this section, we note that there is a simple instance in which time-dependent results can be analyzed directly by the steady-state results of Sec. III. This occurs for the case of a subcritical medium and $T = \infty$. Then, assuming that a source of neutrons is beamed onto D for only a finite amount of time, the angular flux ψ will tend to zero as $t \to \infty$. Thus, one can integrate Eq. (2.11) from $t = 0$ to $t = \infty$ and define

$$\psi(\mathbf{r}, \boldsymbol{\Omega}, E) = \int_0^\infty \psi(\mathbf{r}, \boldsymbol{\Omega}, E, t) dt$$

to obtain exactly Eq. (2.1) for the steady-state ψ. The boundary condition is just the time-integrated boundary condition for the time-dependent ψ. Sanchez and McCormick have discussed this (and more general) procedure for slab geometry problems.[10]

V. ADDITIONAL RESULTS

In the previous sections of this article we have considered the problem of forward (and adjoint) transport with boundary conditions that are as general as possible, constrained only by the requirement that ψ and ψ^* are both measurable for all $\mathbf{r} \in \partial D$, all $\boldsymbol{\Omega}$, all E, and all suitable t if the problem is time dependent. In this section, we show that by placing additional constraints on these boundary conditions, a simplification of our results can occur. For brevity and simplicity, we only consider the case of steady-state transport as described in Sec. III.

To be specific, we prove that for certain types of boundary conditions on ψ and ψ^*, the expressions (3.6) for $\mathbf{S}$ simplify to line integrals involving only ψ and ψ^* (not their tangential derivatives) over simple closed curves on ∂D. This makes the resulting constraint (3.3) on σ_T and σ_s substantially simpler and almost certainly less prone to experimental error, because errors in measurements of $\nabla_T \psi$ or $\nabla_T \psi^*$ are likely to be much greater than errors in ψ or ψ^*. We shall not attempt to discuss the most general boundary conditions for which this simplification occurs; we just show that it can occur in special cases.

To describe a special case, let Σ_1 and Σ_2 be simply connected subsets of the boundary ∂D of D with the following properties: (1) the boundaries of Σ_1 and Σ_2 are simple closed curves, Γ_1 and Γ_2, having piecewise continuous tangent vectors; and (2) Σ_1 is sufficiently small in diameter that there exists a unit vector $\hat{\boldsymbol{\Omega}}$ with the property that $\hat{\boldsymbol{\Omega}} \cdot \mathbf{n} < 0$ for all unit outer normal vectors $\mathbf{n}$ corresponding to points in Σ_1. (Thus, $\hat{\boldsymbol{\Omega}}$ points into D at all points in Σ_1. If Σ_1 happens to consist of a planar part of ∂D; then $\hat{\boldsymbol{\Omega}}$ exists and can be any unit vector pointing into D through this plane. In general, $\hat{\boldsymbol{\Omega}}$ exists if Σ_1 is "small" enough that $\mathbf{n}_1 \cdot \mathbf{n}_2 > 0$ for all unit outer normals $\mathbf{n}_1$ and $\mathbf{n}_2$ corresponding to points on Σ_1.) Finally, let $\chi_n(\mathbf{r})$, $n = 1,2$, be the characteristic functions for Σ_1 and Σ_2:

$$\chi_n(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} \in \Sigma_n \, , \\ 0, & \mathbf{r} \in \partial D - \Sigma_n \, . \end{cases} \tag{5.1}$$

We now consider the forward transport problem consisting of Eq. (2.1) and the boundary condition

$$\psi(\mathbf{r}, \boldsymbol{\Omega}, E) = \chi_1(\mathbf{r}) \delta(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}), \quad \mathbf{r} \in \partial D, \ \boldsymbol{\Omega} \cdot \mathbf{n} < 0 \, . \tag{5.2}$$

(This equation describes a uniform, monodirectional beam incident on Σ_1.) Also, we consider the adjoint problem consisting of Eq. (2.5) and

$$\psi^*(\mathbf{r}, \boldsymbol{\Omega}, E) = \chi_2(\mathbf{r}), \quad \mathbf{r} \in \partial D, \ \boldsymbol{\Omega} \cdot \mathbf{n} > 0 \, . \tag{5.3}$$

(The physical interpretation of ψ^* with this boundary condition is given in Sec. II.)

To proceed, we use Eqs. (5.2) and (5.3) in Eq. (3.6b) [use of Eq. (3.6a) leads to the same result] and write

$$\mathbf{S} = \mathbf{S}^+ + \mathbf{S}^- \, , \tag{5.4}$$

where

$$\mathbf{S}^+ = \int_{\partial D} \iint_{\boldsymbol{\Omega} \cdot \mathbf{n} > 0} \psi [\boldsymbol{\Omega} \times (\nabla_T \psi^* \times \mathbf{n})] \, d^2\Omega \, dE \, d^2r \tag{5.5}$$

and

$$\begin{aligned} \mathbf{S}^- &= \int_{\partial D} \iint_{\boldsymbol{\Omega} \cdot \mathbf{n} < 0} \psi [\boldsymbol{\Omega} \times (\nabla_T \psi^* \times \mathbf{n})] \, d^2\Omega \, dE \, d^2r \\ &= \iint_{\Sigma_1} \hat{\boldsymbol{\Omega}} \times [\nabla_T \psi^*(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) \times \mathbf{n}] \, dE \, d^2r \, . \end{aligned} \tag{5.6}$$

If we define

$$d(\mathbf{r}, \Gamma_2) = \text{the distance from } \mathbf{r} \text{ to } \Gamma_2 \, , \tag{5.7}$$

then by Eq. (5.3), for $\boldsymbol{\Omega} \cdot \mathbf{n} > 0$,

$$\nabla_T \psi^* = -\delta[d(\mathbf{r}, \Gamma_2)] \mathbf{m} \, , \tag{5.8}$$

where δ is the usual delta function and $\mathbf{m}$ is the unit outer normal to Γ_2 in the plane of ∂D. Introducing Eq. (5.8) into

Eq. (5.5), we obtain

$$\mathbf{S}^{+} = -\int_{\Gamma_2}\iint_{\mathbf{\Omega}\cdot\mathbf{n}>0} \psi[\mathbf{\Omega}\times(\mathbf{m}\times\mathbf{n})]d^{2}\Omega\, dE\, d^{1}r\,. \quad (5.9)$$

Finally, we note that

$$-\mathbf{m}\times\mathbf{n} = \mathbf{t}\,, \quad (5.10)$$

where $\mathbf{t}$ is the unit tangent vector pointing in the direction of the transverse of Γ_2. (This direction is right handed with respect to the outer normals of Σ_2.) Equation (5.9) thus reduces to

$$\mathbf{S}^{+} = \int_{\Gamma_2}\iint_{\mathbf{\Omega}\cdot\mathbf{n}>0} (\mathbf{\Omega}\times\mathbf{t})\psi\, d^{2}\Omega\, dE\, d^{1}r\,, \quad (5.11)$$

which is the desired simplification of Eq. (5.5).

To simplify Eq. (5.6), it is necessary to use vector indicial notation and Stokes' theorem.[15] Then, with

$$\hat{\psi}^{*}(\mathbf{r}, E) \equiv \psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)\,, \quad (5.12)$$

we have

$$\hat{\mathbf{\Omega}}\times[\nabla_T\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)\times\mathbf{n}] = \hat{\mathbf{\Omega}}\times(\nabla\hat{\psi}^{*}\times\mathbf{n}) = \epsilon_{ijk}\hat{\Omega}_j\epsilon_{klm}\hat{\psi}^{*}_{,l}n_m$$
$$= -\epsilon_{mlk}[\epsilon_{ijk}\hat{\Omega}_j\hat{\psi}^{*}]_{,l}n_m\,. \quad (5.13)$$

Thus, by Stokes' theorem,

$$\int_{\Sigma_1}\hat{\Omega}\times[\nabla_T\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)\times\mathbf{n}]d^{2}r$$
$$= -\int_{\Sigma_1}\epsilon_{mlk}[\epsilon_{ijk}\hat{\Omega}_j\hat{\psi}^{*}]_{,l}n_m\, d^{2}r$$
$$= -\int_{\Gamma_1}[\epsilon_{ijk}\hat{\Omega}_j\hat{\psi}^{*}]t_k\, d^{1}r$$
$$= -\int_{\Gamma_1}(\hat{\mathbf{\Omega}}\times\mathbf{t})\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)d^{1}r\,. \quad (5.14)$$

Using this result in Eq. (5.6), we obtain

$$\mathbf{S}^{-} = -\int_{\Gamma_1}\int(\hat{\mathbf{\Omega}}\times\mathbf{t})\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)dE\, d^{1}r\,, \quad (5.15)$$

which is the desired simplification. Combining Eqs. (5.4), (5.11), and (5.15), we obtain the final result

$$\mathbf{S} = \int_{\Gamma_2}\iint_{\mathbf{\Omega}\cdot\mathbf{n}>0}(\mathbf{\Omega}\times\mathbf{t})\psi(\mathbf{r}, \mathbf{\Omega}, E)d^{2}\Omega\, dE\, d^{1}r$$
$$-\int_{\Gamma_1}\int(\hat{\mathbf{\Omega}}\times\mathbf{t})\psi^{*}(\mathbf{r}, \hat{\mathbf{\Omega}}, E)dE\, d^{1}r\,, \quad (5.16)$$

which consists of line integrals of just ψ and ψ^{*}.

Other boundary conditions for ψ and ψ^{*} also lead to expressions of the form (5.16) for $\mathbf{S}$. For example, one could replace the delta function in $\mathbf{\Omega}$ in Eq. (5.2) by a characteristic function in $\mathbf{\Omega}$ over a subset of the cone of directions pointing into D through all of Σ_1. ($\hat{\mathbf{\Omega}}$ belongs to this cone.) However, we shall not consider this topic further here.

ACKNOWLEDGMENTS

I would like to thank Norman McCormick and Richard Sanchez for their interest, encouragement, and helpful suggestions.

[1] C. E. Siewert, "On a possible experiment to establish the validity of the one-speed or constant cross-section model of the neutron transport equation," J. Math. Phys. **19**, 1587 (1978).
[2] C. E. Siewert, "On Establishing a Two-Term Scattering Law in the Theory of Radiative Transfer," Z. Angew. Math. Phys. **30**, 522 (1979).
[3] N. J. McCormick, "Transport scattering coefficients from reflection and transmission measurements," J. Math. Phys. **20**, 1504 (1979).
[4] C. E. Siewert, "On the Inverse Problem for a Three-Term Phase Function," J. Quant. Spectrosc. Radiat. Transfer **22**, 441 (1979).
[5] C. E. Siewert and J. R. Maiorino, "The Inverse Problem for a Finite Rayleigh-Scattering Atmosphere," Z. Angew. Math. Phys. **31**, 767 (1980).
[6] E. W. Larsen, "Solution of the inverse problem in multigroup transport theory," J. Math. Phys. **22**, 158 (1981).
[7] N. J. McCormick and R. Sanchez, "Inverse problem transport calculations for anisotropic scattering coefficients," J. Math. Phys. **22**, 199 (1981).
[8] R. Sanchez and N. J. McCormick, "General solutions to inverse transport problems," J. Math. Phys. **22**, 847 (1981).
[9] C. E. Siewert and W. L. Dunn, "On inverse problems for plane-parallel media with nonuniform source illumination," J. Math. Phys. **23**, 1376 (1982).
[10] R. Sanchez and N. J. McCormick, "Numerical Evaluation of Optical Single-Scattering Properties Using Multiple-Scattering Inverse Transport Methods," J. Quant. Spectrosc. Radiat. Transfer **28**, 169 (1982).
[11] R. Sanchez and N. J. McCormick, "Inverse Problem Calculations for Multigroup Diffusion Theory," Nucl. Sci. Eng. **83**, 63 (1983).
[12] C. E. Siewert, "Solutions to an Inverse Problem in Radiative Transfer with Polarization-I," J. Quant. Spectrosc. Radiat. Transfer (in press).
[13] N. J. McCormick and R. Sanchez, "Solutions to an Inverse Problem in Radiative Transfer with Polarization-II," J. Quant. Spectrosc. Radiat. Transfer (in press).
[14] G. I. Bell and S. Glasstone, *Nuclear Reactor Theory* (Van Nostrand Reinhold, New York, 1970), p. 258.
[15] R. Aris, *Vector, Tensors, and the Basic Equations of Fluid Mechanics* (Prentice-Hall, Englewood Cliffs, NJ, 1965).

Symmetric Hadamard series

M. R. Brown
Department of Astrophysics, South Parks Road, Oxford, OX1 3RQ, United Kingdom

(Received 20 July 1982; accepted for publication 23 December 1982)

In a general curved space–time, the requirements that the Feynman Green's function be symmetric and have the Hadamard form are shown to result in specific constraints on the local behavior of the function. These constraints are solved yielding a general form for the function.

PACS numbers: 11.10.Cd, 02.30.Bi

I. INTRODUCTION

The Feynman Green's function, or time-ordered, two-point function, is a quantity of central importance in the study of quantum field theory in curved, or flat, space–time. In Minkowski space–time there is, for a given field, exactly one such function. When space–time is curved, there are often many candidates for the title. In this paper I wish to discuss the structure of these functions that is required by the two constraints: that they have the Hadamard[1] form and that they be symmetric functions of the two space–time points involved in their definition. I shall not discuss whether they ought to have the Hadamard form, although there is fast growing support for this idea,[2] nor shall I discuss boundary conditions or Cauchy problems. They must be symmetric functions, and it is how this condition affects the Hadamard form that I shall investigate. I shall use the example of the massless, conformally invariant, scalar field in an arbitrary curved space–time. The analysis will be seen to be applicable to more general fields.

Although in writing this paper I have in mind the application to quantum field theory, it is exclusively concerned with properties of the classical wave equation; Planck's constant enters only in spirit. This is an important point: Much of the subsequent analysis is about finding a missing length. In quantum field theory this length might find expression as an arbitrary renormalization length or the Planck length. Here, with a massless, classical field theory, it is a length that can only be constructed from the curvature of space–time itself.

II. THE SYMMETRIC HADAMARD SERIES

In this section I shall derive a necessary condition for the Green's function $G(x,x')$ to be a symmetric solution to the inhomogeneous wave equation,

$$(\Box - \tfrac{1}{6}R)G(x,x') = -g^{-1/2}(x)\delta^4(x-x') \tag{2.1}$$

having the Hadamard form,

$$G(x,x') = i(8\pi^2)^{-1}[\Delta^{1/2}(\sigma + i\epsilon)^{-1} + v\ln(\sigma + i\epsilon) + w]. \tag{2.2}$$

First, note some well-known features of Eq. (2.2): The factors $i\epsilon$ are included to give G the singularity structure that is appropriate for a Feynman Green's function. $2\sigma(x,x')$ is the square of the length along the geodesic joining x and x'. (One can require that x and x' belong to a "simple region"[3]; this ensures that it is meaningful to speak of their being joined by a unique geodesic.) $\Delta(x,x')$ is the symmetric biscalar constructed from the Van Vleck–Morette determinant, viz.,

$$\Delta(x,x') \equiv -g^{-1/2}(x)\,g^{-1/2}(x')\det(-\sigma_{;ab'}). \tag{2.3}$$

Δ satisfies the equation

$$\sigma^{;a}(\ln\Delta)_{;a} = 4 - \Box\sigma. \tag{2.4}$$

The functions $v(x,x')$ and $w(x,x')$ can be represented as the uniformly convergent power series,[1]

$$v(x,x') = \sum_{n=0}^{\infty} v_n(x,x')\sigma^n(x,x'), \tag{2.5}$$

$$w(x,x') = \sum_{n=0}^{\infty} w_n(x,x')\sigma^n(x,x'), \tag{2.6}$$

where the coefficients v_n and w_n satisfy the differential recursion relations

$$(n+1)(n+2)v_{n+1} + (n+1)v_{n+1;c}\sigma^{;c} - (n+1)v_{n+1}\Delta^{-1/2}\Delta^{1/2}_{;c}\sigma^{;c} + \tfrac{1}{2}(\Box - \tfrac{1}{6}R)v_n = 0, \tag{2.7}$$

and

$$\begin{aligned}(n+1)(n+2)w_{n+1} &+ (n+1)w_{n+1;c}\sigma^{;c} \\ &- (n+1)w_{n+1}\Delta^{-1/2}\Delta^{1/2}_{;c}\sigma^{;c} \\ &+ \tfrac{1}{2}(\Box - \tfrac{1}{6}R)w_n + (2n+3)v_{n+1} + v_{n+1;c}\sigma^{;c} \\ &- v_{n+1}\Delta^{-1/2}\Delta^{1/2}_{;c}\sigma^{;c} = 0;\end{aligned} \tag{2.8}$$

the biscalar $v(x,x')$ is completely determined by Eq. (2.7), and the boundary condition

$$v_0 + v_{0;c}\sigma^{;c} - v_0\Delta^{-1/2}\Delta^{1/2}_{;c}\sigma^{;c} + \tfrac{1}{2}(\Box - \tfrac{1}{6}R)\Delta^{1/2} = 0. \tag{2.9}$$

$v(x,x')$ is a solution to the homogeneous wave equation. The functions $v(x,x')$ and $v_n(x,x')$ are known to be symmetric.[2] v and v_1 have the covariant Taylor series expansions

$$v(x,x') = \tfrac{1}{2}v_{ab}(x)\sigma^{;a}\sigma^{;b} - \tfrac{1}{4}v_{ab;c}(x)\sigma^{;a}\sigma^{;b}\sigma^{;c} + O(\sigma^2), \tag{2.10}$$

and

$$v_1(x,x') = v_1(x) - \tfrac{1}{2}v_{1;a}(x)\sigma^{;a} + O(\sigma), \tag{2.11}$$

where

$$\begin{aligned}v^{ab} &= \tfrac{1}{120}(C^{c(ab)d}R_{cd} + 2C^{c(ab)d}{}_{;cd}), \\ &= \tfrac{1}{240}g^{-1/2}\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2}C_{abcd}C^{abcd},\end{aligned} \tag{2.12}$$

and

$$v_1(x) = \tfrac{1}{720}(R_{abcd}R^{abcd} - R_{ab}R^{ab} + \Box R). \tag{2.13}$$

Equations (2.12) and (2.13) are immediate consequences of the formulae given in the Appendix.

Less is known about the biscalar $w(x,x')$. Clearly it must be symmetric if G is to be symmetric. w (and hence G) is completely determined by the recursion relations once the biscalar $w_0(x,x')$ is specified. Thus the requirement that $w(x,x')$ be symmetric can be seen as a condition on $w_0(x,x')$. w, unlike v, is not a solution to the homogeneous wave equation; it is a simple matter to show that it has to satisfy the equation

$$(\Box - \tfrac{1}{6}R)w(x,x') = -6v_1(x) + 2v_{1;a}(x)\sigma^{;a} + O(\sigma). \tag{2.14}$$

w is unlike v in another important respect: The biscalar $v(x,x')$ has a covariant Taylor series expansion, the first few terms of which are given by Eq. (2.10). The complete expansion has the property that the coefficients $v_{ab}(x)$, etc., are polynomial functions of the curvature tensor and its covariant derivatives. One might ask if one should expect the same property to hold for the covariant Taylor seris expansion of w, when, as in often the case, one seeks to find a purely geometrical solution to equation (2.2). The answer is that, in general, one should not: In Eq. (2.2) the argument of the logarithm is a dimensional quantity. Thus w must supply a term $-v(x,x')\ln L(x,x')$, where $L(x,x')$ is a function having the dimensions of area. The requirement that G be geometrical implies that L must be some function of the curvature tensor. I shall return to this point in the next section where I shall be able to specify further $L(x,x')$.

Let me now determine a condition on $w_0(x,x')$ that must be satisfied if $G(x,x')$ is to be symmetric. I begin with some observations on covariant Taylor series: Let A be a biscalar possessing a covariant Taylor series expansion in a neighborhood of the point x, namely,

$$A(x,x') = A(x) + A_a(x)\sigma^{;a} + \tfrac{1}{2}A_{ab}(x)\sigma^{;a}\sigma^{;b} + \tfrac{1}{6}A_{abc}(x)\sigma^{;a}\sigma^{;b}\sigma^{;c} + O(\sigma^2), \tag{2.15}$$

where $A_{ab} = A_{(ab)}$ and $A_{abc} = A_{(abc)}$, etc. The expansion coefficients, A_{ab} etc., can be expressed as coincidence limits of covariant derivatives of the biscalar $A(x,x')$ by means of the equations[4]

$$\begin{aligned} A(x) &= [A], \\ A_a(x) &= [A_{;a}] - [A]_{;a}, \\ A_{ab}(x) &= [A_{;(ab)}] - 2[A_{;(a}]_{;b)} + [A]_{;(ab)}, \\ A_{abc}(x) &= [A_{;(abc)}] - 3[A_{;(ab}]_{;c)} + 3[A_{;(a}]_{;bc)} - [A]_{;(abc)}, \end{aligned} \tag{2.16}$$

where I use the standard notation

$$[A] \equiv \lim_{x' \to x} A(x,x').$$

Using these equations, it is easy to compute the Taylor series for the function $A(x',x)$. The requirement that $A(x,x')$ equal $A(x',x)$ results in the conditions

$$2A_a(x) = -A_{;a}(x), \tag{2.17}$$

$$4A_{abc}(x) = -6A_{(ab;c)}(x) + A_{;(abc)}(x), \tag{2.18}$$

and so on. More generally, the requirement of symmetry determines the odd coefficients, A_a, A_{abc}, A_{abcde}, etc. However, I shall need only Eqs. (2.17) and (2.18) in what follows and shall not record the higher order constraints.

$w(x,x')$ is a symmetric biscalar that, it is supposed, possesses a Taylor series expansion. Therefore, by the above argument, it can be written

$$w(x,x') = w(x) - \tfrac{1}{2}w_{;a}(x)\sigma^{;a} + \tfrac{1}{2}w_{ab}(x)\sigma^{;a}\sigma^{;b} - \tfrac{1}{4}\{w_{ab;c}(x) - \tfrac{1}{6}w_{;abc}(x)\}\sigma^{;a}\sigma^{;b}\sigma^{;c} + O(\sigma^2), \tag{2.19}$$

where $w(x) = [w]$ and $w_{ab} = [w_{;ab}]$.

At this point there are several ways to proceed. Perhaps the most direct is to require that $w(x,x')$, as given by Eq. (2.19), satisfy Eq. (2.14). So doing, one obtains the equations

$$w^a{}_a(x) = \tfrac{1}{6}Rw(x) - 6v_1(x), \tag{2.20}$$

and

$$\{w^a{}_b(x) - \tfrac{1}{2}\delta^a{}_b w^c{}_c(x)\}_{;a} = 2v_{1;b}(x) + \tfrac{1}{4}(\Box w(x))_{;b} + \tfrac{1}{2}R^a{}_b w_{;a}(x) - \tfrac{1}{12}Rw_{;b}(x). \tag{2.21}$$

Next one has to relate these equations to $w_0(x,x')$. This is done as follows: $w_0(x,x')$ has a Taylor series expansion

$$w_0(x,x') = w_0(x) - \tfrac{1}{2}w_{0;a}(x)\sigma^{;a} + \tfrac{1}{2}w_{0ab}(x)\sigma^{;a}\sigma^{;b} + O(\sigma^{3/2}), \tag{2.22}$$

where $w_0(x) = w(x)$. [The form of the second term in Eq. (2.22) is required by the symmetry of $w(x,x')$; it must not be supposed that $w_0(x,x')$ has any particular symmetry property.] $w_1(x,x')$ has a Taylor series

$$w_1(x,x')\sigma = \tfrac{1}{2}w_{1ab}(x)\sigma^{;a}\sigma^{;b} + O(\sigma^{3/2}), \tag{2.23}$$

where, by Eq. (2.8) and (2.6),

$$w_{1ab}(x) = g_{ab}[w_1(x,x')] \tag{2.24}$$

and

$$[w_1(x,x')] = \tfrac{1}{24}Rw_0(x) - \tfrac{1}{4}w_0{}^a{}_a(x) - \tfrac{3}{2}v_1(x). \tag{2.25}$$

Combining Eqs. (2.22) and (2.23) with (2.6), one sees that

$$w(x,x') = w_0(x) - \tfrac{1}{2}w_{0;a}(x)\sigma^{;a} + \tfrac{1}{2}\{w_{0ab}(x) + w_{1ab}(x)\}\sigma^{;a}\sigma^{;b} + O(\sigma^{3/2}). \tag{2.26}$$

Comparing this equation with Eq. (2.19) yields the result

$$w_{ab}(x) = w_{0ab}(x) + w_{1ab}(x). \tag{2.27}$$

Now Eqs. (2.20) and (2.21) can be written in terms of $w_0(x,x')$. The first of these equations is identically satisfied; in other words, it is not a constraint on $w_0(x,x')$. The second is more interesting and becomes

$$\{w_0{}^a{}_b(x) - \tfrac{1}{4}\delta^a{}_b w_0{}^c{}_c(x)\}_{;a} = \tfrac{1}{2}v_{1;b}(x) + \tfrac{1}{4}(\Box w_0(x))_{;b} + \tfrac{1}{2}R^a{}_b w_{0;a}(x) + \tfrac{1}{24}\{R_{;b}w_0(x) - Rw_{0;b}(x)\}. \tag{2.28}$$

Equation (2.28) must be satisfied by the coefficients in the Taylor series expansion of $w_0(x,x')$ if G is to be a symmetric Hadamard solution to Eq. (2.1). Of course, there will be additional constraints on the higher order Taylor series coefficients. These would require some dedication to compute; fortunately, one needs only those terms up to $w_{0ab}(x)$ to understand quantum field theoretic energy densities.[5] In this context, notice that $w_0(x,x') = 0^6$ is not a solution to Eq. (2.28) unless $v_1(x)$ is constant. $v_1(x)$ [Eq. (2.13)] is a function that is commonly known[7] as the "trace anomaly."

In the next section I shall describe the geometrical solutions to Eq. (2.28).

III. THE FORM OF $w(x,x')$

I shall regard Eq. (2.28) as a constraint on $w_{0ab}(x)$ for some given w_0 in a general curved space–time. It can be solved as follows:

Let me write

$$w_{0ab} = s_{ab} + t_{ab}, \tag{3.1}$$

where s_{ab} satisfies

$$(s^a{}_b - \tfrac{1}{4}\delta^a{}_b s^c{}_c)_{;a} = \tfrac{1}{4}(\Box w_0)_{;b} + \tfrac{1}{2}R^a{}_b w_{0;a} + \tfrac{1}{24}(R_{;b} w_0 - R w_{0;b}), \tag{3.2}$$

and t_{ab} satisfies

$$(t^a{}_b - \tfrac{1}{4}\delta^a{}_b t^c{}_c)_{;a} = \tfrac{1}{2}v_{1;b}. \tag{3.3}$$

A solution to Eq. (3.2) is provided by

$$s_{ab} = \tfrac{1}{6}(w_0 R_{ab} - \tfrac{1}{4}g_{ab} w_0 R) + \tfrac{1}{3}(w_{0;ab} - \tfrac{1}{4}g_{ab}\Box w_0). \tag{3.4}$$

This is geometrical, provided, of course, that w_0 is a function of the curvature. That it satisfies Eq. (3.2) is easily checked: One uses the Bianchi identity

$$R^a{}_{b;a} = \tfrac{1}{2}R_{;b}, \tag{3.5}$$

and the differential identity

$$\Box(w_{0;b}) = (\Box w_0)_{;b} + R^a{}_b w_{0;a}. \tag{3.6}$$

Finding a solution to Eq. (3.3) is not so easy. I first gave a solution to this equation some years ago.[8] However, the method I then used is inappropriate in the present context. I think that the following is a more interesting way to proceed.

I define the tensor T:

$$T_{ab} \equiv t_{ab} - \tfrac{1}{4}g_{ab} t^c{}_c - \tfrac{1}{2}v_1 g_{ab}. \tag{3.7}$$

Then Eq. (3.3) implies that

$$T^{ab}{}_{;a} = 0 \tag{3.8}$$

and

$$T^a{}_a = -2v_1. \tag{3.9}$$

Thus one has a geometrical solution to Eq. (3.3) if one can find a geometrical tensor T^{ab} that is conserved [Eq. (3.8)] and whose trace is proportional to the trace anomaly [Eq. (3.9)]. The clue to finding such a tensor is provided by the conservation equation: suppose that T^{ab} is the variation with respect to the metric of an invariant action. In other words, let

$$T^{ab} = 2g^{-1/2}\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2} A(g_{cd}). \tag{3.10}$$

Equation (3.8) is the statement that A be a scalar under general coordinate transformations. Equation (3.9) can be treated as a statement about the scaling behavior of A; more precisely,

$$\frac{\delta}{\delta\omega}\int d^4x\, \hat{g}^{1/2} A(\hat{g}_{cd})\Big|_{\omega=0} = -g^{1/2} T^a{}_a, \tag{3.11}$$

where $\hat{g}$ is related g by the equation

$$\hat{g}_{ab} \equiv e^{-2\omega} g_{ab}.$$

Equation (3.11) suggests that a suitable action can be found by integrating the functional differential equation

$$\frac{\delta}{\delta\omega}\int d^4x\, \hat{g}^{1/2} A(\hat{g}_{cd}) = -\hat{g}^{1/2} T^a{}_a(\hat{g}_{cd}). \tag{3.12}$$

Equation (3.12) clearly reduces to (3.11) in the limit $\omega = 0$. The variation with respect to ω is taken holding the metric g_{ab} fixed. In this sense, Eq. (3.12) is a partial, functional differential equation. Bearing this in mind, it is remarkably simple to integrate it.

Using the formulae in the Appendix, $\hat{g}^{1/2} T^a{}_a(\hat{g}_{cd})$ can be written

$$\begin{aligned}\hat{g}^{1/2} T^a{}_a(\hat{g}_{cd}) &= -2\hat{g}^{1/2} v_1(\hat{g}_{cd}) \\ &= -\tfrac{1}{360}\hat{g}^{1/2}(\hat{R}_{abcd}\hat{R}^{abcd} - \hat{R}_{ab}\hat{R}^{ab} + \hat{\Box}\hat{R}) \\ &= -\tfrac{1}{360}g^{1/2}\{R_{abcd}R^{abcd} - R_{ab}R^{ab} + \Box R \\ &\quad + 2R\,\Box\omega + 2R_{;a}\omega^{;a} + 6\Box(\Box\omega) + 8[(\Box\omega)^2 \\ &\quad - \omega_{;ab}\omega_{;ab} - R_{ab}\omega^{;a}\omega^{;b} \\ &\quad - \omega^{;c}\omega_{;c}\Box\omega - 2\omega_{;ab}\omega^{;a}\omega^{;b}]\}.\end{aligned} \tag{3.13}$$

It is straightforward to see that Eq. (3.12) can be functionally integrated to give

$$\hat{g}^{1/2} A(\hat{g}_{cd}) = g^{1/2} C(\omega; g_{cd}) + g^{1/2} F(g_{cd}), \tag{3.14}$$

where F is a function of the metric (but not ω) and

$$\begin{aligned}C(\omega; g_{cd}) \equiv \tfrac{1}{360}[&(R_{abcd}R^{abcd} - R_{ab}R^{ab} + \Box R)\omega + 3(\Box\omega)^2 \\ &- 2R_{ab}\omega^{;a}\omega^{;b} - 4\omega^{;c}\omega_{;c}\Box\omega + 2(\omega^{;c}\omega_{;c})^2].\end{aligned} \tag{3.15}$$

C is determined uniquely up to total divergences. Equation (3.14) must hold for $\omega = 0$. This implies that

$$F(g_{cd}) = A(g_{cd}). \tag{3.16}$$

Equation (3.14) may now be seen to determine the scaling behavior of the function A:

$$\hat{g}^{1/2} A(e^{-2\omega} g_{cd}) - g^{1/2} A(g_{cd}) = g^{1/2} C(\omega; g_{cd}). \tag{3.17}$$

Thus the problem of finding a tensor satisfying equations (3.8) and (3.9) has been reduced to finding a scalar A that satisfies the scaling equation (3.17).

The solutions to Eq. (3.17) can be found by choosing ω to be a function of the curvature that has the scaling law

$$\omega(e^{-2\chi} g_{ab}) = \omega(g_{ab}) - \chi. \tag{3.18}$$

Equation (3.17) then has the solution $A^*(g_{cd})$, where

$$A^*(g_{ab}) = -C(\omega(g_{ab}); g_{ab}). \tag{3.19}$$

This is clearly a solution since

$$A^*(e^{-2\chi} g_{ab}) = -C(\omega - \chi; e^{-2\chi} g_{ab}). \tag{3.20}$$

Setting $\chi = \omega$ in Eq. (3.20) yields

$$A^*(e^{-2\omega} g_{ab}) = 0. \tag{3.21}$$

More general solutions to Eq. (3.17) are obtained by adding to a solution $g^{1/2}A^*$ any conformal invariant. It is worth noting that, when ω satisfies Eq. (3.18), C has the scaling property

$$C(\omega(e^{-2\chi} g_{ab}); e^{-2\chi} g_{ab}) = C(\omega(g_{ab}); g_{ab}) - C(\chi; g_{ab}). \tag{3.22}$$

Thus, if ω_1 and ω_2 both satisfy Eq. (3.18), the difference $\{C(\omega_1; g_{ab}) - C(\omega_2; g_{ab})\}$ is a conformal invariant.

To summarize these results: I have shown that a solution to Eqs. (3.8) and (3.9) is provided by

$$T^{ab} = -2g^{-1/2}\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2}C(\omega(g_{cd});g_{cd}), \quad (3.23)$$

where ω is a scalar function of the curvature satisfying Eq. (3.18) and $C(\omega;g)$ is given by Eq. (3.15). There exists the freedom to add to T^{ab} any conserved, trace-free tensor. In terms of the function $w(x,x')$, this freedom corresponds to the freedom to add a symmetric solution to the homogeneous wave equation that has zero coincidence limit. {The function $v(x,x')$ provides a particular example. Recall that $v(x,x)$ is zero and $v_{ab}(x)$ is the variation of a conformally invariant action [Eq. (2.12)].}

It now remains to show that there exist scalar functions of the curvature, ω, that satisfy Eq. (3.18). These functions do indeed exist; they are more or less difficult to construct, depending upon whether or not the Weyl curvature of the space–time is zero.

When the space–time is not conformally flat ($C_{abcd} \neq 0$),

$$\omega = -\tfrac{1}{4}\ln C_{abcd}C^{abcd} \quad (3.24)$$

is the simplest to construct. Of course, it may be that C_{abcd} is not zero, but the particular invariant $C_{abcd}C^{abcd}$ is. In this case one can select any other, nonvanishing, invariant. One could take ω to be proportional to the logarithm of the sum of the squares of the independent invariants of the Weyl tensor; this would have some advantages. However, it still may not be the most natural choice. To see what might be more natural, it is necessary to see how T_{ab} contributes to the Green's function $G(x,x')$. It does this through the function $w(x,x')$. Combining Eqs. (2.26), (3.1), (3.4), and (3.7), $w(x,x')$ is now seen to have the form,

$$\begin{aligned} w(x,x') = {} & w_0(x) - \tfrac{1}{2}w_{0;a}(x)\sigma^{;a} \\ & + \tfrac{1}{2}\big[T_{ab} - v_1 g_{ab} + \tfrac{1}{6}R_{ab}w_0 \\ & + \tfrac{1}{3}(w_{0;ab} - \tfrac{1}{4}g_{ab}\Box w_0)\big]\sigma^{;a}\sigma^{;b} \\ & + O(\sigma^{3/2}). \end{aligned} \quad (3.25)$$

In the previous section I made the point that it was artificial to write G in the form of Eq. (2.2); in particular, $w(x,x')$ had to provide a term $-v(x,x')\ln L$. The Taylor series expansion of this term about the point x is provided by Eq. (2.10), and the necessary assumption that $L(x,x)$ is not zero. A term having exactly this structure is indeed provided by $w(x,x')$. Whatever the actual choice for ω, its scaling behavior is characteristic of a function that is the logarithm of a length; Eq. (3.24) is an example.

Consider taking the variation in Eq. (3.23) to obtain an explicit form for the tensor T^{ab}. It is easy to see that the only place where the logarithmic nature of ω survives is in the term

$$-\tfrac{1}{180}g^{-1/2}\omega\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2}(R_{abcd}R^{abcd} - R_{ab}R^{ab} + \Box R). \quad (3.26)$$

Elsewhere ω appears differentiated, either functionally or covariantly. Using the formulae in the Appendix, the term (3.26) can be shown to be equal to

$$-\tfrac{1}{120}g^{-1/2}\omega\frac{\delta}{\delta g_{ab}}\int d^4x\, g^{1/2}(C_{abcd}C^{abcd}). \quad (3.27)$$

(The variation of the other terms vanishes identically.)

Recalling Eqs. (2.12) and (3.25), one sees that the term (3.27) contributes to $w(x,x')$ an amount:

$$-\tfrac{1}{2}v_{ab}(x)\sigma^{;a}\sigma^{;b}\ln L^2(x), \quad (3.28)$$

where $\omega \equiv \ln L$. In short, ω provides the length that is missing in Eq. (2.2).

It would seem natural to choose ω to be a function of the biscalar $v(x,x')$, insofar as it is the existence of v that requires the existence of ω. The tensor v_{ab} is well suited to this purpose: its scaling behavior can be inferred from Eq. (2.12) and is given by

$$v^a{}_b(e^{-2\chi}g_{cd}) = e^{4\chi}v^a{}_b(g_{cd}). \quad (3.29)$$

The eigenvalues v_i of $v^a{}_b$ scale in the same way. Thus it is possible to choose for ω,

$$\omega = -(4d)^{-1}\ln h_d(v_i), \quad (3.30)$$

where $h_d(v_i)$ is any homogeneous function of degree d.

When the space–time is conformally flat v_{ab} vanishes. Indeed it can be shown that $v(x,x')$ vanishes.[9] If this is the case, then there is no pressing need to construct an ω satisfying Eq. (3.18). However, solutions do still exist and can be defined implicitly. For example,

$$\omega = -\ln\psi, \quad (3.31)$$

where ψ is a geometrical solution to the wave equation

$$(\Box - \tfrac{1}{6}R)\psi = 0,$$

which has the scaling behavior

$$\psi(e^{-2\chi}g_{ab}) = e^{\chi}\psi(g_{ab}).$$

Of course, functions of the type (3.31) will continue to provide solutions to Eq. (3.18) when $C_{abcd} = 0$. It can be shown[8] that by proceeding in this way one obtains for the tensor T^{ab} the polynomial expression[10]

$$\begin{aligned} T^{ab} = {} & \tfrac{1}{720}\big[6R^{ac}R^{b}{}_{c} + 2R^{;ab} \\ & - 6RR^{ab} - g^{ab}(2\Box R - 2R^2 + 3R_{cd}R^{cd})\big]. \end{aligned} \quad (3.32)$$

(One chooses for ω the solution that is conformal to a constant, the solution in flat space–time.)

The simple form for T^{ab} in Eq. (3.32) is essentially a feature of the conformal flatness. The variation of Eq. (3.15) is easy to compute because

$$\delta C = \int d^4x\,\{\hat{g}^{1/2}T^a{}_a(\hat{g}_{cd})\delta\omega - 2g^{1/2}T^{ab}\delta g_{ab}\}, \quad (3.33)$$

and, for the above choice of ω, the coefficient of $\delta\omega$ vanishes; one does not have to compute further the variation of ω with respect to the metric.

In general, when the Weyl tensor is nonzero, one can arrange for a similar simplification to take place: Require that $\omega(g_{cd})$ is determined by the condition

$$T^a{}_a(e^{-2\omega}g_{cd}) = 0. \quad (3.34)$$

This equation implies that ω satisfies (3.18) and has some interesting solutions.[11] A nongeometrical solution worth mentioning is provided by

$$\omega = -\tfrac{1}{2}\ln(K^a g_{ab}K^b), \quad (3.35)$$

where K^a is any curl-free, Killing vector field of the Ricci flat metric g_{ab}.[12]

IV. CONCLUSION

To some extent it is artificial to look for more or less natural functions ω that satisfy Eq. (3.18): for a given problem with prescribed boundary conditions an ω will be automatically provided. But, as I said in the Introduction, I was interested in how far the requirements of symmetry and having the Hadamard form determine the local structure of Feynman Green's functions. In this spirit, the hard conclusions of this paper are those contained in Eq. (3.25), (3.23), and (3.15). The rest is more speculative but, I hope, not without interest.

APPENDIX

The conventions used in this paper are consistent with Ref. 13. The following formulae were used in the derivation of the equations appearing in the text:

$$\hat{R}^{ab}{}_{cd} = e^{2\omega}(R^{ab}{}_{cd} + \delta^{[a}{}_{[c}\omega^{b]}{}_{d]}), \tag{A1}$$

$$\hat{R}^{b}{}_{d} = e^{2\omega}[R^{b}{}_{d} + \tfrac{1}{4}(2\omega^{b}{}_{d} + \delta^{b}{}_{d}\omega^{a}{}_{a})], \tag{A2}$$

$$\hat{R} = e^{2\omega}(R + \tfrac{3}{2}\omega^{a}{}_{a}), \tag{A3}$$

$$\hat{\Box}\phi = e^{2\omega}(\Box\phi - 2\omega^{;a}\phi_{;a}), \tag{A4}$$

where

$$\hat{R}_{abcd} \equiv R_{abcd}(e^{-2\omega}g_{ef}),$$

$$\omega_{ab} \equiv 4(\omega_{;ab} + \omega_{;a}\omega_{;b}) - 2g_{ab}\omega^{;c}\omega_{;c},$$

and a semicolon denotes covariant differentiation with respect to the metric g_{ab};

$$\begin{aligned}\sigma_{;ab}(x,x') = {} & g_{ab}(x) - \tfrac{1}{3}R_{acbd}(x)\sigma^{;c}\sigma^{;d} \\ & + \tfrac{1}{12}R_{acbd;e}\sigma^{;c}\sigma^{;d}\sigma^{;e} \\ & - (\tfrac{1}{60}R_{acbd;ef} + \tfrac{1}{45}R_{acgd}R_{be}{}^{g}{}_{f}) \\ & \times\sigma^{;c}\sigma^{;d}\sigma^{;e}\sigma^{;f} + O(\sigma^{5/2}),\end{aligned} \tag{A5}$$

$$\begin{aligned}\Delta^{1/2}(x,x') = {} & 1 + \tfrac{1}{12}R_{ab}\sigma^{;a}\sigma^{;b} - \tfrac{1}{24}R_{ab;c}\sigma^{;a}\sigma^{;b}\sigma^{;c} \\ & + (\tfrac{1}{288}R_{ab}R_{cd} + \tfrac{1}{360}R^{e}{}_{a}{}^{f}{}_{b}R_{ecfd} \\ & + \tfrac{1}{180}R_{ab;cd})\sigma^{;a}\sigma^{;b}\sigma^{;c}\sigma^{;d} + O(\sigma^{5/2}),\end{aligned} \tag{A6}$$

$$\begin{aligned}\Delta^{1/2}{}_{;ab}(x,x') = {} & \tfrac{1}{6}R_{ab} + \tfrac{1}{12}(2R_{c(a;b)} - R_{ab;c})\sigma^{;c} \\ & + (\tfrac{1}{40}R_{ab;cd} + \tfrac{1}{40}R_{cd;(ab)} - \tfrac{1}{15}R_{c(a;b)d} \\ & + \tfrac{1}{72}R_{ab}R_{cd} + \tfrac{1}{36}R_{ac}R_{bd} \\ & + \tfrac{1}{180}R_{e(a}R_{b)c}{}^{e}{}_{d} + \tfrac{1}{90}R_{aebf}R^{e}{}_{c}{}^{f}{}_{d} \\ & - \tfrac{1}{90}R^{e}{}_{cf(a}R_{b)}{}^{f}{}_{ed} \\ & - \tfrac{1}{90}R^{e}{}_{cf(a}R_{b)e}{}^{f}{}_{d} + \tfrac{11}{180}R_{ce}R^{e}{}_{(ab)d}) \\ & \times\sigma^{;c}\sigma^{;d} + O(\sigma^{3/2}),\end{aligned} \tag{A7}$$

[formulae (A5), (A6), and (A7) are taken from Ref. 14]

$$V_{a;[bc]} = \tfrac{1}{2}V_{d}R^{d}{}_{abc}, \tag{A8}$$

$$C_{abcd} = R_{abcd} + g_{a[d}R_{b]} - g_{b[d}R_{c]a} + \tfrac{1}{3}Rg_{a[c}g_{d]b}, \tag{A9}$$

$$C^{a}{}_{bcd;a} = R_{b[d;c]} - \tfrac{1}{6}g_{b[d}R_{;c]}, \tag{A10}$$

$$C_{abcd}C^{abcd} = R_{abcd}R^{abcd} - 2R_{ab}R^{ab} + \tfrac{1}{3}R^{2}, \tag{A11}$$

$$C_{acde}C_{b}{}^{cde} = \tfrac{1}{4}g_{ab}C_{efgh}C^{efgh}, \tag{A12}$$

$$\delta g^{1/2} = \tfrac{1}{2}g^{1/2}g^{ab}\delta g_{ab}, \tag{A13}$$

$$\delta\Gamma_{ab}{}^{c} = \tfrac{1}{2}g^{cd}(\delta g_{ad;b} + \delta g_{bd;a} - \delta g_{ab;d}), \tag{A14}$$

$$\delta R_{abc}{}^{d} = (\delta\Gamma_{ca}{}^{d})_{;b} - (\delta\Gamma_{cb}{}^{d})_{;a}, \tag{A15}$$

$$\delta R_{ab} = g^{cd}(\delta g_{c(a;b)d} - \tfrac{1}{2}\delta g_{ab;cd} - \tfrac{1}{2}\delta g_{cd;ab}), \tag{A16}$$

$$\delta R = g^{ab}g^{cd}(\delta g_{ac;db} - \delta g_{ab;cd}) - R^{ab}\delta g_{ab}. \tag{A17}$$

[1] J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations* (Yale U. P., New Haven, 1923); B. S. DeWitt and R. W. Brehme, "Radiation damping in a gravitational field," Ann. Phys. (N.Y.) **9**, 220 (1960).

[2] S. A. Fulling, M. Sweeny, and R. M. Wald, "Singularity structure of the two point function in quantum field theory in curved space–time," Commun. Math. Phys. **63**, 259 (1978); S. A. Fulling, F. J. Narcovwich, and R. M. Wald, "Singularity structure of the two point function in quantum field theory in curved space–time II," Ann. Phys. (N.Y.) **136**, 243 (1981).

[3] R. Penrose, *Techniques of Differential Topology in Relativity* (SIAM, Philadelphia, 1972).

[4] S. L. Adler, J. Lieberman, and Y. J. Ng, "Regularization of the stress–energy tensor for vector and scalar particles propagating in a general background metric," Ann. Phys. (N.Y.) **106**, 279 (1977); R. M. Wald, Phys. Rev. D **17**, 1477 (1978).

[5] M. R. Brown, A. C. Ottewill, and S. T. C. Siklos, "Comments on conformal Killing vector fields and quantum field theory," Phys. Rev. D **26**, 1881 (1982).

[6] S. L. Adler *et al.*, Ref. 4.

[7] M. J. Duff, "Observations on conformal anomalies," Nucl. Phys. B **125**, 334 (1977).

[8] M. R. Brown, "Actions and anomalies," preprint, University of Texas at Austin, 1978.

[9] F. G. Friedlander, *The Wave Equation on a Curved Space–Time* (Cambridge U.P., Cambridge, 1975).

[10] L. S. Brown and J. Cassidy, Phys. Rev. D **16**, 1712 (1977).

[11] M. R. Brown, "Quantum field theory and conformal transformations," preprint, Oxford University, 1981.

[12] D. N. Page, Phys. Rev. D **25**, 1499 (1982).

[13] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge U.P., Cambridge, 1973).

[14] S. M. Christensen, "Vacuum expectation value of the stress tensor in an arbitrary curved background: The covariant point separation method," Phys. Rev. D **14**, 2490 (1976).

Properties of the Schwinger model

Anton Z. Capri and Ruggero Ferrari[a)]
Theoretical Physics Institute, University of Alberta, Edmonton, Alberta, T6G 2J1, Canada

(Received 31 March 1983; accepted for publication 5 August 1983)

We present all the Wightman functions for an explicit operator solution of the Schwinger model. To understand these better, we study the algebra of fields of this model, representations of this algebra as well as the Hamiltonian. The latter turns out to elucidate the "confinement" of the fermion field. In addition we comment on the renormalization of the theory as well as on the analyticity of the amplitudes in terms of the coupling constant.

PACS numbers: 11.10.Mn

I. INTRODUCTION

The Schwinger model has proved to be a rich source of theoretical results for further conjectures as well as for testing conjectures. This makes it worthwhile to examine this model, in as much detail, and from as many perspectives, as possible. In a previous paper,[1] we presented an explicit operator solution of the Schwinger model for an arbitrary covariant gauge. The solution was local, Lorentz-covariant, chirally invariant, and the gauge transformations of the first kind were implementable.

In this paper we further examine properties of these solutions. In particular, we list all the Wightman functions, construct the Hamiltonian, and examine its spectrum. Finally we also comment briefly on the renormalization of the theory and the analyticity of the Wightman functions with respect to the coupling constant.

Throughout we use the same notation as in Ref. 1. To introduce this notation, we briefly review the results obtained in Ref. 1. When necessary, we use the following explicit conventions:

$$g^{00}=1, \quad \epsilon^{01}=1,$$

$$\gamma^0=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad \gamma^1=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}, \quad \gamma^5=\gamma^0\gamma^1.$$

We also define $F=F^{(+)}+F^{(-)}$ for any free quantized field F to be, respectively, the annihilation and creation parts of F.

The Schwinger model, as considered by us in Ref. 1, is defined by the formal Lagrangian

$$\mathscr{L}=-\tfrac{1}{4}(F_{\mu\nu})^2-\tfrac{1}{2}\alpha(\partial\cdot A)^2+\bar{\phi}\,(i\gamma\cdot\partial-e\gamma\cdot A)\phi. \tag{1}$$

The solutions for the Heisenberg fields ϕ, A_μ are given in terms of certain free "building block" fields as follows:

$$\phi(x)=Z^{-1/2}\exp[-ie\Omega^{(-)}(x)]\psi(x)\exp[-ie\Omega^{(+)}(x)], \tag{2}$$

$$A_\mu(x)=\partial_\mu c(x)+\epsilon_{\mu\nu}\,\partial^\nu d(x), \tag{3}$$

where

$$\Omega(x)=c(x)+\gamma^5 d(x), \tag{4}$$

$$c(x)=a(x)+\beta\rho(x),$$

$$d(x)=(\sqrt{\pi}/e)[\Sigma(x)+\sigma(x)]-(\alpha\pi/e^2)\tilde{b}(x). \tag{5}$$

Here β is a real parameter and the other quantities are free fields defined as follows:

$$\gamma\cdot\partial\psi(x)=0, \tag{6}$$

$$\Box a=b, \quad \Box b=0, \tag{7}$$

$$(\Box+e^2/\pi)\Sigma=0, \tag{8}$$

$$:\bar{\psi}\gamma_\mu\psi:(x)=(1/\sqrt{\pi})\partial_\mu\rho=(1/\sqrt{\pi})\epsilon_{\mu\nu}\,\partial^\nu\sigma, \tag{9}$$

and

$$\partial_\mu b=\epsilon_{\mu\nu}\,\partial^\nu\tilde{b}. \tag{10}$$

The relevant two-point functions for these fields are

$$\langle 0|\psi_a(x)\bar{\psi}_\beta(0)|0\rangle=-i(i\gamma\cdot\partial)_{\alpha\beta}D^{(+)}(x), \tag{11}$$

$$\langle 0|a(x)a(0)|0\rangle=-(i/\alpha)I^{(+)}(x)+i(\beta^2+2\beta\sqrt{\pi}/2)D^{(+)}(x), \tag{12}$$

$$\langle 0|\Sigma(x)\Sigma(0)|0\rangle=-i\Delta^{(+)}(x), \tag{13}$$

$$\langle 0|\rho(x)\rho(0)|0\rangle=\langle 0|\sigma(x)\sigma(0)|0\rangle=-iD^{(+)}(x), \tag{14}$$

$$\langle 0|\rho(x)\sigma(0)|0\rangle=-i\tilde{D}^{(+)}(x), \tag{15}$$

where

$$D^{(+)}(x)=(4\pi i)^{-1}\ln\mu^2(-x^2+i\epsilon x^0), \tag{16}$$

$$I^{(+)}(x)=(16\pi i)^{-1}x^2\ln\mu^2(-x^2+i\epsilon x^0). \tag{17}$$

$\Delta^{(+)}$ is the solution of $(\Box+e^2/\pi)\Delta^{(+)}=0$ with the normalization that yields

$$\partial_0\Delta^{(+)}(x)|_{x^0=0}=\delta(x^1)$$

and finally

$$\tilde{D}^{(+)}(x)=(4\pi i)^{-1}\ln[(x^0-i\epsilon+x^+)/(x^0-i\epsilon-x^1)]. \tag{18}$$

The finite normalization constant Z is given by

$$Z=(\sqrt{\pi}\mu/e)^{1/2}\exp[-\tfrac{1}{2}(\gamma-\ln 2)], \tag{19}$$

where γ is Euler's constant and μ is an arbitrary mass scale.

For further details, regarding properties of these solutions, the reader is referred to Ref. 1.

2. THE WIGHTMAN FUNCTIONS

Since the solution given in Ref. 1 conserves fermion number, the only nontrivial Wightman function involving only Fermi fields was already listed there and is given by

[a)] Permanent address: Istituto di Fisica, Università di Pisa, Piazza Torricelli 2, Pisa, Italy.

$$\langle 0|\phi(x_1)\cdots\phi(x_n)\phi^*(y_1)\cdots\phi^*(y_n)|0\rangle$$
$$= W_n(x,y)$$
$$= Z^{-n}\exp[\mathscr{F}^{(+)}(x,y)]w_0^{2n}(x,y), \tag{20}$$

where

$$\mathscr{F}^{+}(x,y) = \sum_{i,j=1}^{n} F^{(+)}(x_i,y_j) - \sum_{i<j=1}^{n}(F^{(+)}(x_i,x_j)+F^{(+)}(y_i,y_j)) \tag{21}$$

and

$$F^{(+)}(x,y) = e^2\left\{-\frac{i}{\alpha}I^{(+)}(x-y) - \frac{i\pi}{e^2}\gamma_x^5\gamma_y^5[\Delta^{(+)}(x-y)-D^{(+)}(x-y)]\right\}. \tag{22}$$

Here

$$w_0^2(x,y) = \frac{1}{2\pi i}\frac{\gamma\cdot(x-y)\gamma_0}{[(x-y)^2 - i\epsilon(x^0-y^0)]} = \langle 0|\psi(x)\psi^*(y)|0\rangle \tag{23}$$

is the free fermion two-point function and

$$w_0^{2n}(x,y) = \sum_{\text{Perm}}(-1)^{\delta p}\prod_{i=1}^{n} w_0^2(x_i,y_{ip}) = \det w_0^2(x_i,y_i) \tag{24}$$

is the free fermion $2n$-point function. It is worth noting that the two-point function can also be written as

$$\langle 0|\psi(x)\psi^*(y)|0\rangle = (\mu/2\pi)\exp 2\pi i \times\{\gamma^5\tilde{D}^{(+)}(x-y) - D^{(+)}(x-y)\}. \tag{25}$$

The vector potential A_μ is just a sum of free fields, as stated by Eq. (2). Thus all n-point functions of A_μ are just given in terms of the following two-point function:

$$\langle 0|A_\mu(x)A_\nu(y)|0\rangle = iH_{\mu\nu}^{(+)}(x-y) = (i/\alpha)\,\partial_\mu\partial_\nu I^{(+)}(x-y) + (i\pi/e^2)\,\partial_\mu\partial_\nu[\Delta^{(+)}(x-y) - D^{(+)}(x-y)] + ig_{\mu\nu}\Delta^{(+)}(x-y). \tag{26}$$

The result is

$$\langle 0|A_{\mu_1}(x_1)\cdots A_{\mu_{2n}}(x_{2n})|0\rangle = \sum\prod_{j_k}[iH_{\mu_{j_1}\mu_{j_2}}^{(+)}(x_{j_1}-x_{j_2})], \tag{27}$$

where the sum is over all partitions of $2n$ into n disjoint two-element subsets

$$(j_1j_2)(j_3j_4)\cdots(j_{2n-1}j_{2n}) \text{ with } j_{2k-1}<j_{2k}.$$

We next compute the simplest of the mixed Wightman functions, namely

$$\langle 0|A_\mu(z)\phi(x)\phi^*(y)|0\rangle = \langle 0|A_\mu^{(+)}(z)\phi(x)\phi^*(y)|0\rangle. \tag{28}$$

The computation is facilitated by using Eqs. (72)–(74) of Ref. 1, namely,

$$\phi(x) = \exp[-ie\Xi^{(-)}(x)]\zeta(x)\exp[-ie\Xi^{(+)}(x)] \tag{29}$$

with

$$\zeta(x) = \exp i\,P^{(-)}(x)\psi(x)\exp i\,P^{(+)}(x),$$
$$P(x) = \sqrt{\pi}(\rho(x) - \gamma^5\sigma(x)) \tag{30}$$

and

$$\Xi(x) = a(x) + \gamma^5(\sqrt{\pi}/2)\Sigma(x) - (\alpha\pi/e^2)\tilde{b}(x). \tag{31}$$

This is just a rewriting of the solution given by Eq. (2). We then obtain,

$$[A_\mu^{(+)}(z),\Xi^{(-)}(x)] = (i/e)G_\mu^{(+)}(z,x) = -i\{(1/\alpha)\partial_\mu I^{(+)}(z-x) + (\pi/e^2)\gamma^5\epsilon_{\mu\nu}\times\partial^\nu[\Delta^{(+)}(z-x) - D^{(+)}(z-x)]\}. \tag{32}$$

Now using the identity (for $[A,B]$ a c-number)

$$Ae^B = e^B(A + [A,B]),$$

we obtain

$$[A_\mu^{(+)}(z),\phi(x)] = G_\mu^{(+)}(z,x)\phi(x),$$
$$[A_\mu^{(+)}(z),\phi^*(x)] = -G_\mu^{(+)}(z,x)\phi^*(x), \tag{33}$$

where we used that $G^{(-)*}(x,y) = G^{(+)}(x,y)$.

Combining these results yields the desired Wightman function

$$\langle 0|A_\mu(z)\phi(x)\phi^*(y)|0\rangle = [-G_\mu^{(+)}(z,y) + G_\mu^{(+)}(z,x)]\langle 0|\phi(x)\phi^*(y)|0\rangle. \tag{34}$$

This generalizes immediately to

$$\langle 0|A_\mu(z)\phi(x_1)\cdots\phi(x_n)\phi^*(y_1)\cdots\phi^*(y_n)|0\rangle = \sum_{i=1}^{n}[G_\mu^{(+)}(z,x_i) - G_\mu^{(+)}(z,y_i)]W_n(x,y). \tag{35}$$

Further combining this result with Eq. (27), we find

$$\langle 0|A_{\mu_1}(z_1)\cdots A_{\mu_l}(z_l)\phi(x_1)\cdots\phi(x_n)\phi^*(y_1)\cdots\phi^*(y_n)|0\rangle = \sum_{r=1}^{n}\sum_{P(l)}\langle 0|A_{\mu_{i_{k+1}}}(z_{i_{k+1}})\cdots A_{\mu_{i_{k+l}}}(z_{i_{k+l}})\cdot|0\rangle \times\prod_{j=1}^{k}[G_{\mu_{i_j}}^{(+)}(z_{i_j},x_r) - G_{\mu_{i_j}}^{(+)}(z_{i_j},y_r)]W_n(x,y), \tag{36}$$

where the sum over $P(l)$ is over all partitions of l indices into two disjoint sets, with $i_j<i_k$ for $j<k$.

This completes the evaluation of all the Wightman functions. Before turning to the Hamiltonian, it is convenient to examine the operators $\zeta(x)$, $\zeta^*(x)$ given by Eq. (30). As we show later, they do not belong to the algebra of fields, but are nevertheless useful objects.

3. THE ζ-REPRESENTATION

We begin by considering the vacuum expectation value

$$Z(x,y) = \langle 0|\zeta(x_1)\cdots\zeta(x_n)\zeta^*(y_1)\cdots\zeta^*(y_n)|0\rangle. \tag{37}$$

To evaluate Z, we need

$$\zeta(x)\zeta(y)^* = \exp i[P^{(-)}(x) - P^{(-)}(y)]\psi(x)\psi^*(y) \times\exp i[P^{(+)}(x) - P^{(+)}(y)] \times\exp H^{(+)}(x,y), \tag{38}$$

$$\zeta(x_1)\zeta(x_2) = \exp i[P^{(-)}(x_1) + P^{(-)}(x_2)]\psi(x_1)\psi(x_2) \times\exp i[P^{(+)}(x_1) + P^{(+)}(x_2)] \times\exp[-H^{(+)}(x_1,x_2)], \tag{39}$$

$$\zeta^*(y_1)\zeta^*(y_2) = \exp - i[P^{(-)}(y_1) - P^{(-)}(y_2)]\psi^*(y_1)\psi^*(y_2)$$
$$\times\exp\{ - i[P^{(+)}(y_1) + P^{(+)}(y_2)]\}$$
$$\times\exp[- H^{(+)}(y_1, y_2)], \quad (40)$$

where

$$H^{(+)}(x, y) = i\pi[(1 + \gamma_x^5\gamma_y^5)D^{(+)}(x - y)$$
$$- (\gamma_x^5 + \gamma_y^5)\tilde{D}^{(+)}(x - y)]. \quad (41)$$

The subscripts x, y, etc. on γ_x^5 indicate on which side of a quantity γ^5 is to be multiplied. Thus $(\gamma_x^5 F(x, y))_{\alpha\beta}$ $= \gamma_{\alpha\gamma}^5 F_{\gamma\beta}(x, y)$ whereas $(\gamma_y^5 F(x, y))_{\alpha\beta} = F_{\alpha\gamma}(x, y)\gamma_{\gamma\beta}^5$. It then follows that

$$Z(x, y) = \exp \mathscr{H}(x, y) W_0^{2n}(x, y), \quad (42)$$

where

$$\mathscr{H}(x, y) = \sum_{i,j=1}^{n} H^{(+)}(x_i, y_j)$$
$$- \sum_{i<j=1}^{n} [H^{(+)}(x_i, x_j) + H^{(+)}(y_i, y_j)]. \quad (43)$$

To further evaluate this expression, we notice that both $\mathscr{H}(x, y)$ and $w_0^{2n}(x, y)$ are diagonal in the spinor indices. If we now consider all spinor indices to have the value 1, we find

$$w_{011}^2 = (2\pi i)^{-1}[x^0 - y^0 - (x^1 - y^1) - i\epsilon]^{-1} \quad (44)$$

and

$$w_{011\cdots11}^{2n}(x, y) = \det\{(2\pi i)^{-1}[x_i^- - y_j^- - i\epsilon]^{-1}\}$$
$$= (2\pi i)^{-n}\frac{\Pi_{i<j}(x_i^- - x_j^-)(y_i^- - y_j^-)}{\Pi_{i,j}(x_i^- - y_j^- - i\epsilon)}. \quad (45)$$

The last step above is proven in Refs. 2 and 3.

To complete the computation, we write out $\exp \mathscr{H}^{(+)}$ for $\gamma_x^5 = \gamma_y^5 = 1$ and use Eq. (25) to get

$$\exp \mathscr{H}^{(+)}(\gamma_x^5 = \gamma_y^5 = 1)$$
$$= \mu^n i^n \frac{\Pi_{i,j}(x_i^- - y_i^-)}{\Pi_{i<j}(x_i^- - x_j^- - i\epsilon)(y_i^- - y_j^- - i\epsilon)}, \quad (46)$$

and hence

$$\langle 0|\zeta_1(x_1)\cdots\zeta_1(x_n)\zeta_1^*(y_1)\cdots\zeta_1^*(y_n)|0\rangle = (\mu/2\pi)^n.$$

A similar computation for general spurion indices yields the following result:

$$\langle 0|\zeta_1(x_1)\cdots\zeta_1(x_n)\zeta_2(y_1)\cdots\zeta_2(y_m)\zeta_1^*(z_1)$$
$$\cdots\zeta_1^*(z_{n'})\zeta_2^*(w_1)\cdots\zeta_2^*(w_{m'})|0\rangle$$
$$= \delta_{n,n'}\delta_{m,m'}(-)^{n\cdot m}(\mu/2\pi)^{n+m}. \quad (47)$$

Thus we see that the algebra specified by Eq. (75) of Ref. 1 is represented on a Hilbert space with an orthonormal basis

$$|n,m\rangle = \left(\frac{\mu}{2\pi}\right)^{-(|n|+|m|)/2}(\zeta_1^*)^{(n+|n|)/2}\zeta_1^{(n-|n|)/2}$$
$$\times(\zeta_2^*)^{(m+|m|)/2}\zeta_2^{(m-|m|)/2}|0\rangle \quad (48)$$

for $n, m = 0, \pm 1, \pm 2, \pm 3, \ldots$ and $\langle n,m|n',m'\rangle$ $= \delta_{n,n'}\delta_{m,m'}$.

In this representation one has

$$\zeta_1\zeta_1^* = \zeta_2\zeta_2^* = \mu/2\pi, \quad (49)$$

and two charges q and q_5 can be defined by

$$[q, \zeta] = \zeta, \quad [q_5, \zeta] = \gamma^5\zeta, \quad (50)$$
$$q|0\rangle = q_5|0\rangle = 0.$$

It is worth noting that the algebra given by Eq. (75) of Ref. 1 has many representations, not just the one given above. The representations can even be finite if the charges q, q_5 are omitted from the algebra. An example of such a finite representation is

$$\zeta_1 = \zeta_1^* = \left(\frac{\mu}{2\pi}\right)^{1/2}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$$
$$\zeta_2 = \zeta_2^* = \left(\frac{\mu}{2\pi}\right)^{1/2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \quad (51)$$

4. THE HAMILTONIAN

The solutions of the Schwinger model in Ref. 1 were constructed to preserve the gauge invariance of the underlying formal Lagrangian equation (1). Thus, in regularizing the terms appearing in the Hamiltonian, we must respect this invariance. A straightforward calculation shows that the required condition is

$$\tau(\Lambda)P_\mu\tau^{-1}(\Lambda) = P_\mu + \alpha\int dx^1\, \partial\cdot A\, \ddot{\partial}_0\partial_\mu\Lambda, \quad (52)$$

where $\Box\Lambda = 0$ and P_μ is the generator of space-time translations. The operator $\tau(\Lambda)$ is explicitly given by

$$\tau(\Lambda) = \exp i\alpha\int dx^1\, \partial\cdot A\, \ddot{\partial}_0\Lambda \quad (53)$$

and has the properties that

$$\tau(\Lambda)A_\mu(x)\tau^{-1}(\Lambda) = A_\mu - \partial_\mu\Lambda,$$
$$\tau(\Lambda)\phi(x)\tau^{-1}(\Lambda) = \exp(ie\Lambda)\phi(x). \quad (54)$$

To see how the condition implied by Eq. (52) is implemented, we consider the *classical*, unsymmetrized energy–momentum tensor

$$K^{\mu\nu} = i\bar{\phi}\gamma^\mu\partial^\nu\phi - F^{\mu\rho}A_{\rho,\nu} - \alpha g^{\mu\rho}A_\rho{}^{,\nu}\,\partial\cdot A$$
$$- g^{\mu\nu}[- \tfrac{1}{4}F_{\rho\sigma}F^{\rho\sigma}$$
$$- \tfrac{1}{2}\alpha(\partial\cdot A)^2 + \bar{\phi}(i\gamma\cdot\partial - e\gamma\cdot A)\phi]. \quad (55)$$

The classical momentum operator is given by

$$P^\nu = \int dx^1\, K^{0\nu}. \quad (56)$$

Examining the individual terms (appearing in P^ν) under a gauge transformation, we find that Eq. (52) is valid if the following transformations hold:

$$\tau(\Lambda)\int(-\alpha A^{0,\nu}\,\partial\cdot A)\,dx^1\,\tau^{-1}(\Lambda)$$
$$= -\alpha\int A^{0,\nu}\,\partial\cdot A\,dx^1 + \alpha\int\partial\cdot A\,\partial_0\partial^\nu\Lambda\,dx^1, \quad (57)$$

$$\tau(\Lambda)i\int dx^1\,\bar{\phi}\gamma^0\gamma^\nu\phi\tau^{-1}(\Lambda)$$
$$= i\int dx^1\,\bar{\phi}\gamma^0\,\partial^\nu\phi - e\int dx^1\,\bar{\phi}\gamma^0\phi\,\partial^\nu\Lambda, \quad (58)$$

$$\tau(\Lambda)\int dx^1\, F^{0\rho}A_{\rho,}{}^{,\nu}\tau^{-1}(\Lambda)$$
$$=\int dx^1\, F^{0\rho}A_{\rho,}{}^{,\nu}-\int dx^1\,\partial^\rho F_{\rho 0}\,\partial^\nu\Lambda. \tag{59}$$

The most problematic term is the term $i\bar\phi\gamma^\mu\partial^\nu\phi$ encountered in Eq. (58). To define this term, we use the gauge-invariant point splitting given by

$$\exp\left\{-ie\int_x^y d\xi^\mu A_\mu^{(-)}(\xi)[i\bar\phi(y)\gamma^\mu\partial^\nu\phi(x)]\right\}$$
$$\times\exp\left[-ie\int_x^y d\xi^\mu A_\mu^{(+)}(\xi)\right]. \tag{60}$$

We expand this expression in a power series in $\eta=y-x$, subtract the singular parts, and verify that the result is compatible with Eq. (58) as well as the general properties of P^ν. The procedure is not covariant but can be made so by an averaging over all η_μ.[4] This averaging is discussed in Appendix A and has the effect of replacing any product of η_μ's as follows:

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_{2n}}}=\frac{(\eta^2)^n}{2n!!}\sum g_{\mu_i\mu_j}\cdots g_{\mu_k\mu_l}, \tag{61}$$

where the sum runs over all partitions of 1,2,...,2n into pairs such that $i<j$. Also

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_{2n+1}}}=0.$$

Now using

$$\phi^*(y)\phi(x)=Z^{-1}\exp F^{(+)}(x,y)$$
$$\times\exp ie[\Omega^{(-)}(y)-\Omega^{(-)}(x)]:\psi^*(y)\psi(x):$$
$$\times\exp ie[\Omega^{(+)}(y)-\Omega^{(+)}(x)]+Z^{-1}$$
$$\times\exp F^{(+)}(x,y)\exp ie[\Omega^{(-)}(y)-\Omega^{(-)}(x)]$$
$$\times\langle 0|\psi^*(y)\psi(x)|0\rangle$$
$$\times\exp ie[\Omega^{(+)}(y)-\Omega^{(+)}(x)]. \tag{62}$$

We find on inserting this result in the expression (60) that in the first term the limit $\eta\to 0$ can be taken immediately to yield

$$i:\bar\psi\gamma^\mu\partial^\nu\psi:(x)+e:\psi\gamma^\mu\partial^\nu\Omega\psi:(x)$$
$$=i:\bar\psi\gamma^\mu\partial^\nu\psi:(x)+m[:\gamma^\mu\rho\partial^\nu c:(x)-:\partial^\mu\sigma\partial^\nu d:(x)]. \tag{63}$$

To obtain this, we have used that $\gamma^\mu\gamma^5=-\epsilon^{\mu\nu}\gamma_\nu$ and that $\lim_{\eta\to 0}\exp F^{(+)}(x,x+\eta)=Z$. The second term requires more work and when inserted in (60) yields

$$iZ^{-1}\exp F(x,y)\exp ie\left[\Omega(y)-\Omega(x)-\int_x^y d\xi^\rho A_\rho(\xi)\right]^{(-)}$$
$$\times\{[\partial_x^\nu F(x,y)-ie\,\partial^\nu\Omega(x)]\langle 0|\psi^*(y)\gamma^0\gamma^\mu\psi(x)|0\rangle$$
$$+\langle 0|\psi^*(y)\gamma^0\gamma^\mu\partial^\nu\psi(x)|0\rangle\}$$
$$\times\exp ie\left[\Omega(y)-\Omega(x)-\int_x^y d\xi^\rho A_\rho(\xi)\right]^{(+)}. \tag{64}$$

Now

$$\Omega(y)-\Omega(x)-\int_x^y d\xi^\rho A_\rho(\xi)$$
$$=\gamma_x^5[d(y)-d(x)]-\int_x^y d\xi^\rho\,\xi_{\rho\sigma}\,\partial^\sigma d(\xi). \tag{65}$$

Therefore, we obtain, after expanding in η and retaining only the nonvanishing contributions for $\eta\to 0$, that

$$:\exp ie\left[\Omega(y)-\Omega(x)-\int_x^y d\xi^\rho A_\rho(\xi)\right]:$$
$$\to 1+ie(\gamma_x^5\eta^\rho\partial_\rho d-\eta^\rho\epsilon_{\rho\sigma}\partial^\sigma d)+\tfrac{1}{2}\eta^\alpha\eta^\beta$$
$$\times:-e^2(\gamma_x^5\partial_\alpha d-\epsilon_{\alpha\rho}\partial^\rho d)(\gamma_x^5\partial_\beta d-\epsilon_{\beta\sigma}\partial^\sigma d)$$
$$+ie(\gamma_x^5\partial_\alpha\partial_\beta d-\epsilon_{\alpha\sigma}\partial^\sigma\partial_\beta d):. \tag{66}$$

Combining this with the remaining expressions in (64), we find, after a straightforward computation using the averaging process described in Appendix A and Eq. (61) that the fermion kinetic term when properly defined yields

$$i\bar\phi\gamma^\mu\partial^\nu\phi\to\tfrac{1}{2}i[:\bar\psi\gamma^\mu\partial^\nu\psi:-\partial^\nu:\bar\psi\gamma^\mu\psi:]$$
$$+m:(\partial^\mu\rho-m\epsilon^{\mu\alpha}\partial_\alpha d)\partial^\nu c:$$
$$-m:\partial^\mu\sigma\partial^\nu d:$$
$$+\tfrac{1}{2}m^2g^{\mu\nu}:\partial^\rho d\,\partial_\rho d:, \tag{67}$$

where $m^2=e^2/\pi$. Evaluating the rest of the terms in $K^{\mu\nu}$ and combining all the results, we obtain

$$K^{\mu\nu}$$
$$=\tfrac{1}{2}i[:\bar\psi\gamma^\mu\partial^\nu\psi:-\partial^\nu:\bar\psi\gamma^\mu\psi:]+\alpha:\partial^\nu c\,\partial^\mu b-\partial^\mu c\partial^\nu b:$$
$$-m:\epsilon^{\mu\alpha}\partial_\alpha\Sigma\,\partial^\nu c+\Sigma\epsilon^{\mu\rho}\partial_\rho\partial^\nu c$$
$$+(\partial^\mu\sigma-m\,\partial^\mu d)\partial^\nu d:$$
$$+m^2:\tfrac{1}{2}g^{\mu\nu}\partial_\rho d\,\partial^\rho d-\partial^\mu d\,\partial^\nu d:$$
$$-m:\Sigma\,\partial^\mu\partial^\nu d+(\alpha/m)b\epsilon^{\mu\rho}\partial_\rho\partial^\nu d:$$
$$+g^{\mu\nu}:-\tfrac{1}{2}m^2\Sigma^2+\tfrac{1}{2}\alpha b^2:. \tag{68}$$

After some rewriting we then obtain the Hamiltonian

$$H=\int dx^1:\frac{i}{2}[\bar\psi\gamma^0\partial^0\psi-\partial^0\bar\psi\gamma^0\psi]$$
$$+\alpha\left(\partial^0a\,\partial^0b+\partial^1a\,\partial^1b-\frac{b^2}{2}\right)$$
$$+\alpha\beta(\partial^0\rho\,\partial^0b+\partial^1\rho\partial^1b)-\frac{1}{2}\left[\left(\partial^0\left(\sigma-\frac{\alpha}{m}\tilde b\right)\right)^2\right.$$
$$\left.+\left(\partial^1\left(\sigma-\frac{\alpha}{m}\tilde b\right)\right)^2\right]$$
$$+\frac{1}{2}[(\partial^0\Sigma)^2+(\partial^1\Sigma)^2+m^2\Sigma^2]:. \tag{69}$$

This can also be rewritten as

$$H=H_0+H_1 \tag{70}$$

with

$$H_0=\int dx^1:\frac{i}{2}[\bar\psi\gamma^0\partial^0\psi-\partial^0\bar\psi\gamma^0\psi]$$
$$+\alpha\left(\partial^0a\,\partial^0b+\partial^1a\,\partial^1b-\frac{b^2}{2}\right)$$
$$+\frac{\alpha^2}{2}\left(\beta^2+\frac{2\beta}{m}\right)[(\partial_0b)^2+(\partial_1b)^2]$$
$$+\frac{1}{2}[(\partial_0\Sigma)^2+(\partial_1\Sigma)^2+m^2\Sigma^2]: \tag{71}$$

and

$$H_1 = -\frac{1}{2}\int dx^1 : \left[\partial^0\left(\rho - \alpha\left(\beta + \frac{1}{m}\right)b\right)\right]^2 + \left[\partial^1\left(\rho - \alpha\left(\beta + \frac{1}{m}\right)b\right)\right]^2 : . \tag{72}$$

H_0 is clearly the Hamiltonian of the free building-block fields a, b, Σ, and ψ. Similarly starting from equation (68) we find that the momentum operator is given by

$$P = \int dx^1 K^{01} = P_0 + P_1, \tag{73}$$

where

$$P_0 = \int dx^1 : i\bar{\psi}\gamma^0\partial^1\psi + \alpha(\partial^1 a\partial^0 b + \partial^0 a\partial^1 b) + \alpha^2\left(\beta^2 + \frac{2\beta}{m}\right)\partial^0 b\partial^1 b + \partial^0\Sigma\partial^1\Sigma : \tag{74}$$

and

$$P_1 = \int dx^1 : -\partial^0\left[\rho - \alpha\left(\beta + \frac{1}{m}\right)b\right]\partial^1\left[\rho - \alpha\left(\beta + \frac{1}{m}\right)b\right] : . \tag{75}$$

Here again P_0 is the momentum operator for the free building-block fields a, b, Σ, and ψ. The Lagrangian corresponding to H_0 and P_0 is

$$\begin{aligned}\mathscr{L}_0 = {} & :\bar{\psi}i\gamma\cdot\partial\psi: + \alpha:\partial^\mu a\,\partial_\mu b - \tfrac{1}{2}b^2: \\ & + \tfrac{1}{2}\alpha(\beta^2 + 2\beta/m):(\partial_\mu b)^2: \\ & + \tfrac{1}{2}:(\partial_\mu\Sigma)^2 + m^2\Sigma^2:\end{aligned} \tag{76}$$

and involves only the free fields ψ, $\bar{\psi}$, a, b, and Σ.

A straightforward application of the equal-time commutation relations listed in Appendix B yields

$$\begin{aligned}&[H_0, \psi] = -i\dot{\psi}, \quad [H_0, \Sigma] = -i\dot{\Sigma}, \\ &[H_0, a] = -i\dot{a}, \quad [H_0, b] = -i\dot{b}, \\ &[H_0, \tilde{b}] = -i\dot{\tilde{b}}, \quad [H_0, \rho] = -i\dot{\rho}, \\ &[H_0, \sigma] = -i\dot{\sigma}.\end{aligned} \tag{77}$$

Using the equal-time commutation relations listed in Appendix B once more, we then find

$$\begin{aligned}&[H, \psi] = -i\dot{\psi} + \sqrt{\pi}\{[\dot{\rho} - \alpha(\beta + 1/m)\dot{b}] - [\dot{\sigma} - \alpha(\beta + 1/m)\dot{\tilde{b}}]\gamma^5\}\psi, \\ &[H, \psi] = -i\dot{\Sigma}, \\ &[H, a] = -i\dot{a} - i(\beta + 1/m)[\dot{\rho} - \alpha(\beta + 1/m)\dot{b}], \\ &[H, b] = -i\dot{b}, \quad [H, \tilde{b}] = -i\dot{\tilde{b}}, \\ &[H, \rho] = -i\alpha(\beta + 1/m)\dot{b}, \quad [H, \sigma] = -i\alpha(\beta + 1/m)\dot{\tilde{b}}.\end{aligned} \tag{78}$$

The sets of equations (77) clearly show that H_0 provides a time evolution operator for all the building-block fields and hence for the full algebra of fields $\mathfrak{A}(A_\mu, \phi, \bar{\phi})$. This is in fact what one would naively expect.

In addition to H_0, we have, however, the full Hamiltonian H, and, although it is not obvious from the set of equations (78), this Hamiltonian H is also a time evolution operator for the algebra of fields $\mathfrak{A}(A_\mu, \phi, \bar{\phi})$. This point will be clarified after we discuss the field algebra in the next section.

What is the role of these two Hamiltonians? H_0 provides the time evolution of $\mathfrak{A}$ but does not provide the full physical content of the theory. The subtle message obtained from the full operator H is that in addition to the obvious spectrum obtained from H_0 there are infinitely many Poincaré-invariant states so that we have infinitely many copies of the spectrum of H_0 (excluding the fermions) built up on these translation invariant states. The details of this will be expounded in Secs. 6 and 7 and will reveal just how subtle the confinement of fermions (exclusion from the spectrum of H) is.

5. THE ALGEBRA OF FIELDS

We consider those objects which are local relative to the fields ϕ, $\bar{\phi}$, and A_μ. Some of the useful properties of this algebra $\mathfrak{A}$ are

(i) $b = \partial^\mu A_\mu \in \mathfrak{A}$.

(ii) Since $\tilde{b}$ is not local relative to ϕ we have $\tilde{b} \notin \mathfrak{A}$. However, $\partial_\mu\tilde{b} = \epsilon_{\mu\nu}\partial^\nu b \in \mathfrak{A}$.

(iii) $\Sigma = -(1/m)\epsilon^{\mu\nu}F_{\mu\nu} \in \mathfrak{A}$.

(iv) The dipole field is not local relative to ϕ and hence $a \notin \mathfrak{A}$. However,

$$\partial_\mu[a + (\beta + 1/m)\rho] = A_\mu + (\alpha/m^2)\,\partial_\mu\partial\cdot A - (1/m^2)\,\partial^\nu F_{\mu\nu} \in \mathfrak{A}.$$

The next property of the algebra requires a proof and is thus stated as a lemma.

Lemma:

(v) $\rho, j_{f\mu} = (1/\sqrt{\pi})\partial_\mu\rho$, σ, and $\partial_\mu\sigma$ are *not* elements of $\mathfrak{A}$ if A_μ, ϕ, and $\bar{\phi}$ are irreducibly represented.

Proof: Suppose they belong to $\mathfrak{A}$ and choose $\beta = -1/m$; then they commute with ϕ, $\bar{\phi}$, and A_μ and should be c-numbers. This is contradicted by

$$\langle 0|\rho(x)\rho(y)|0\rangle = -iD^{(+)}(x-y),$$
$$\langle 0|\sigma(x)\sigma(y)|0\rangle = -iD^{(+)}(x-y),$$
$$\langle 0|\sigma(x)\rho(y)|0\rangle = -i\tilde{D}^{(+)}(x-y).$$

For a general value of β consider

$$\rho_0 = \rho - \alpha(\beta + 1/m)b,$$
$$\sigma_0 = \sigma - \alpha(\beta + 1/m)\tilde{b}.$$

Then ρ_0 and σ_0 again commute with ϕ, $\bar{\phi}$, and A_μ and the same argument applies.

(vi) Combining the results of (iv) and (v), we find that for $\beta \neq -1/m$

$$\partial_\mu a = \partial_\mu[a + (\beta + 1/m)\rho] - (\beta + 1/m)\,\partial_\mu\rho \notin \mathfrak{A}.$$

(vii) Since ζ_a and ζ_a^* are not local with respect to ϕ and $\bar{\phi}$, they also do not belong to $\mathfrak{A}$.

In view of these results, it is convenient to use instead of the original building-block fields a and ψ the compound fields

$$a_0 = a - (\alpha/m^2)b + (\beta + 1/m)\rho \tag{79}$$

and

$$\zeta = :\exp[i\sqrt{\pi}(\rho - \sigma\gamma^5)]\psi:. \tag{80}$$

These fields commute with each other and satisfy the same field equations as a and ψ, respectively. Furthermore, we can express A_μ and ϕ in terms of these fields

$$A_\mu = \partial_\mu a_0 + \frac{1}{m}\epsilon_{\mu\nu}\partial^\nu \Sigma, \tag{81}$$

$$\phi = :\exp\{-ie[a_0 + (\alpha/m^2)b + (1/m)(\Sigma - (\alpha/m)\tilde{b})\gamma^5]:\zeta\}, \tag{82}$$

and we see that the parameter β has completely disappeared. Next we find that

$$[H, \zeta] = -i\dot{\zeta}, \tag{83}$$

$$[H, a_0] = -i\dot{a}_0, \tag{84}$$

$$[H, b] = -i\dot{b}, \tag{85}$$

$$[H, \tilde{b}] = -i\dot{\tilde{b}}, \tag{86}$$

$$[H, \Sigma] = -i\dot{\Sigma}. \tag{87}$$

It is still true that ζ, a_0, and $\tilde{b}$ are not elements of $\mathfrak{A}$. If, however, we choose test functions $f_0 \in \mathscr{S}$ vanishing at $p_\mu = 0$, then both $a_0(f_0)$ and $\tilde{b}(f_0)$ belong to $\mathfrak{A}$, the algebra of fields. It would now be easy to read off the spectrum of H except that we find a host of Poincaré-invariant states in addition to the obvious vacuum. We examine these next.

6. POINCARÉ-INVARIANT STATES

To find translation-invariant states, we begin by "undressing" the fermion field ϕ. To do this requires exponentiating certain elements of the algebra $\mathfrak{A}$. We define such exponentials using the triple-dot-product. Thus for any free field $A \in \mathfrak{A}$, we define

$$:\exp A: \equiv \exp A^{(-)} \exp A^{(+)}. \tag{88}$$

For convenience we also choose the value $\beta = -1/m$ in this section. Since $\Sigma \in \mathfrak{A}$, we can "remove" Σ from ϕ and obtain

$$\begin{aligned}\phi_0(x) &\equiv \exp(ie/m)\gamma^5\Sigma^{(-)}(x)\phi(x)\exp(ie/m)\gamma^5\Sigma^{(+)}(x)\\ &= Z^{-1}\exp[-ie(a - (\alpha/m^2)\gamma^5\tilde{b})^{(-)}](x)\zeta(x)\\ &\quad\times\exp[-ie\{-(\alpha/m^2)\tilde{b}\gamma^5\}^{(+)}](x).\end{aligned} \tag{89}$$

Since $a \notin \mathfrak{A}$ but $\partial_\mu a$ is, we cannot "undress" ϕ_0 any further. For this reason we consider bilocal fields which can be undressed as far as the field a is concerned. Due to the presence of γ^5, the $\tilde{b}$ cannot be removed. Thus we consider

$$\phi_0(x)\phi_0^*(y)\exp\left[-ie\int_x^y d\xi^\mu\,\partial_\mu a(\xi)\right]$$

and multiply by the necessary c-number factors to obtain the bilocal field

$$\begin{aligned}B(x,y) &= \exp[(ie\alpha/m^2)(\tilde{b}(x)\gamma_x^5 - \tilde{b}(y)\gamma_y^5)^{(-)}]\zeta(x)\zeta^*(y)\\ &\quad\times\exp[(ie\alpha/m^2)(\tilde{b}(x)\gamma_x^5 - \tilde{b}(y)\gamma_y^5)^{(+)}],\end{aligned} \tag{90}$$

which belongs to the algebra $\mathfrak{A}$. This field has the following local properties:

$$\begin{aligned}[A_\mu(z), B(x,y)] &= (e/m^2)[\partial_\mu\tilde{D}(z-x)\gamma_x^5\\ &\quad - \partial_\mu\tilde{D}(z-y)\gamma_y^5]B(x,y),\end{aligned} \tag{91}$$

$$\begin{aligned}&[\phi(z), B(x,y)]\\ &\quad= \{\exp i\pi[\tilde{D}(z-x) - \tilde{D}(z-y)] - 1\}B(x,y)\phi(z)\\ &\qquad\text{for } \zeta(x) = \zeta(y)\\ &\quad= -\{\exp i\pi[\tilde{D}(z-x) + \tilde{D}(z-y)] + 1\}B(x,y)\phi(z)\\ &\qquad\text{for } \zeta(x) \neq \zeta(y),\end{aligned} \tag{92}$$

and we see that both commutators vanish whenever z is spacelike with respect to both x and y. Thus $B(x,y)$ is truly bilocal.

We next consider the vacuum expectation value

$$\langle 0|\phi(x)\phi^*(y)B(z,w)|0\rangle.$$

Then (keeping always $\beta = -1/m$) using the commutator

$$\begin{aligned}K^{(+)}(x,y) &\equiv [\Xi^{(+)}(x), \Xi^{(-)}(y)]\\ &= -(i/\alpha)I^{(+)}(x,y) - (i/m^2)D^{(+)}(x-y)\\ &\quad + (i/m^2)(\gamma_x^5 + \gamma_y^5)\tilde{D}^{(+)}(x-y)\\ &\quad - (i/m)\gamma_x^5\gamma_y^5\Delta^{(+)}(x-y)\end{aligned} \tag{93}$$

and the identity given by Eq. (25), we find

$$\begin{aligned}&\langle 0|\phi(x)\phi^*(y)B(z,w)|0\rangle\\ &\quad= Z^{-1}\exp[e^2K^{(+)}(x,y)](\mu/2\pi)^2\\ &\qquad\times\{\delta xw\,\delta xy\exp i\pi\gamma_w^5\\ &\qquad\times[\tilde{D}^{(+)}(y-z) - \tilde{D}^{(+)}(x-z) + \tilde{D}^{(+)}(x-w)\\ &\qquad - \tilde{D}^{(+)}(y-w)] + \delta xw\,\delta yz[1 - \delta wz]\\ &\qquad\times\exp i\pi\gamma_w^5[-\tilde{D}(y-z) + \tilde{D}^{(+)}(x-z)\\ &\qquad + \tilde{D}^{(+)}(x-w) - \tilde{D}^{(+)}(y-w)]\},\end{aligned} \tag{94}$$

where the Kronecker delta refers to the Lorentz indices.

Next we take the limit $w \to z$ and consider the three components B_{11}, B_{22}, and B_{12} separately to find

$$\begin{aligned}&\lim_{w\to z}\langle 0|\phi(x)\phi^*(y)B_{11}(z,w)|0\rangle\\ &\quad= \lim_{w\to z}\langle 0|\phi(x)\phi^*(y)B_{22}(z,w)|0\rangle\\ &\quad= Z^{-1}\exp e^2[K^{(+)}(x,y)](\mu/2\pi)^2\end{aligned} \tag{95}$$

so that

$$\lim_{w\to z}B_{11}(z,w) = \lim_{w\to z}B_{22}(z,w) = \mu/2\pi. \tag{96}$$

On the other hand we obtain

$$\begin{aligned}&\lim_{w\to z}\langle 0|\phi(x)\phi^*(y)B_{12}(z,w)|0\rangle\\ &\quad= Z^{-1}(\mu/2\pi)^2\exp e^2K^{(+)}(x,y)\\ &\qquad\times\exp 2\pi i[\tilde{D}^{(+)}(y-z) - \tilde{D}^{(+)}(x-z)]\\ &\quad= -(\mu/2\pi)\langle 0|\phi(x)\phi^*(y)\sigma_+(z)|0\rangle.\end{aligned} \tag{97}$$

We now look for translation-invariant states by Fourier transforming the relevant part of Eq. (97) with respect to z. For convenience we also choose $x = 0$. The relevant expression is

$$\begin{aligned}&\int d^2z\,e^{-ip\cdot z}\exp 2\pi i[\tilde{D}^{(+)}(y-z) - \tilde{D}^{(+)}(-z)]\\ &\quad= \frac{1}{2}\int dz^+\,dz^-\exp\left[-\frac{i}{2}(p^+z^- + p^-z^+)\right]\\ &\qquad\times\left(\frac{y^+ - z^+ - i\epsilon}{y^- - z^- - i\epsilon}\right)^{1/2}\left(\frac{-z^- - i\epsilon}{-z^+ - i\epsilon}\right)^{1/2}.\end{aligned} \tag{98}$$

If we consider the integral over z^+ we find

$$\int_{-\infty}^{\infty} dz^+ \, e^{-ip^- z^+/2}\left[\left(\frac{y^+ - z^+ - i\epsilon}{-z^+ - i\epsilon}\right)^{1/2} - 1 + 1\right]$$
$$= 2\pi\delta\left(\frac{p^-}{2}\right) + 2\pi i \cdot 2\theta(p^-)$$
$$\times \int_0^{y^+} \left(\frac{y^+ - z^+}{z^+}\right)^{1/2} e^{-ip^- z^+/2}\, dz^+. \tag{99}$$

The second term is an analytic function of p^- since it is the Fourier transform of a function with compact support.

Thus Eq. (98) becomes

$$\int d^2 z \, e^{-ip\cdot z} \exp 2\pi i[\tilde{D}^{(+)}(y - z) - \tilde{D}^{(+)}(-z)]$$
$$= (2\pi)^2 \delta^{(2)}(p) + \text{terms in } \theta(p^-)\delta(p^+),\ \theta(p^+)\delta(p^-)$$
and $\theta(p^-)\theta(p^+)$ multiplied by analytic functions in p^+ and p^-. (100)

From this we conclude that the state $(2\pi/\mu)\xi_2^* \xi_1|0\rangle$ is a normalized, Poincaré-invariant state. With these preliminaries out of the way we can finally discuss the spectrum of the Hamiltonian H.

7. THE SPECTRUM OF THE HAMILTONIAN

Using the results of the previous section, we see that we have the normalized Poincaré-invariant states

$$|n\rangle = ((2\pi/\mu)\xi_2^* \xi_1)^{(|n| + n)/2}((2\pi/\mu)\xi_1^* \xi_2)^{(|n| - n)/2}|0\rangle,$$
$$n = 0, \pm 1, \pm 2, \cdots. \tag{101}$$

Each of these states can be used as a cyclic vacuum with regard to the fields Σ, a_0, and b, where the field a_0 is not allowed to carry zero frequencies. In this way we build up a Fock space G_n of $\Sigma(f)$, $a_0(f_0)$, $b(f)$, where $f \in \mathscr{S}(\mathbb{R}^2)$ and $f_0 \in \mathscr{S}_0(\mathbb{R}^2) \subset \mathscr{S}(\mathbb{R}^2)$ is the space of test functions whose support excludes the origin $p_\mu = 0$. The Hilbert space G of asymptotic states is then the direct sum over the individual Fock spaces G_n:

$$G = \bigoplus_{n=-\infty}^{\infty} G_n. \tag{102}$$

It is now clear that each space G_n contains the same spectrum as H_0 if the fermion term is dropped from H_0. Thus in each space G_n no vestige of the fermions remains. The fermions are confined. Nevertheless, a hint of their existence is manifested by the infinite degeneracy of the spectrum.

Another comment is in order. Since a_0 is a dipole field (except for the Landau gauge, $\alpha = 0$), neither H_0 nor H can be diagonalized.[5,6]

8. RENORMALIZATION

In defining the electromagnetic current in Ref. 1 we used a split-point regularization and gauge invariance. The current was then defined by

$$j_\mu(x) = \lim_{\substack{\epsilon \to 0 \\ \epsilon^2 \neq 0}} \left\{\bar{\phi}(x + \epsilon)\gamma_\mu \phi(x) \right.$$
$$\left. \times \exp\left[-ie\int_0^\epsilon A_\nu(x + \xi)\, d\xi^\nu - \langle\ \rangle_0\right]\right\}. \tag{103}$$

It is, however, possible to maintain gauge invariance by using a different definition. Thus we now consider the current $\bar{j}_\mu$ defined by

$$\bar{j}_\mu(x) = \lim_{\substack{\epsilon \to 0 \\ \epsilon^2 \neq 0}} \left\{\bar{\phi}(x + \epsilon)\gamma_\mu \phi(x) \right.$$
$$\left. \times \exp\left[-ie\int_0^\epsilon V_\nu(x + \xi)\, d\xi^\nu - \langle\ \rangle_0\right]\right\}, \tag{104}$$

where

$$V_\nu(x) = A_\nu(x) + u\partial_\nu \partial\cdot A(x) + v\partial^\lambda F_{\lambda\nu}(x) \tag{105}$$

and u, v are two real parameters. This definition also leads to a viable current. As we now show, the net effect of replacing j_μ by $\bar{j}_\mu$ is to renormalize the Lagrangian (1).

Using the same procedure as in Ref. 1 to triple dot order the terms in (104), we find for small ϵ that

$$\bar{j}_\mu(x) = j_\mu(x) - (eu/2\pi)\,\partial^\mu \partial\cdot A(x) - (ev/2\pi)\,\partial^\lambda F_{\lambda\nu}(x), \tag{106}$$

where we have also made extensive use of the Dirac equation

$$(i\partial\!\!\!/ - eA\!\!\!/)\phi = 0. \tag{107}$$

With the above result we find that the equation of motion for A_ν is

$$(1 + e^2 v/2\pi)\,\partial^\lambda F_{\lambda\mu} + (\alpha + e^2 u/2\pi)\,\partial_\mu \partial\cdot A = e j_\mu. \tag{108}$$

Equations (107) and (108) describe the new equations of motion due to using $\bar{j}$ instead of j. They can be considered to be the equations of motion arising from the renormalized formal Lagrangian

$$\mathscr{L}_R = -\tfrac{1}{4} Z_3 (F_{\mu\nu})^2 - \tfrac{1}{2}\alpha Z_\alpha (\partial\cdot A)^2 + \bar{\phi}(i\partial\!\!\!/ - eA\!\!\!/)\phi \tag{109}$$

with

$$Z_3 = 1 + e^2 v/2\pi, \quad Z_\alpha = 1 + e^2 u/2\pi\alpha. \tag{110}$$

By rescaling the fields we can rewrite this Lagrangian as

$$\mathscr{L}_0 = -\tfrac{1}{4}(F_{\mu\nu})^2 - \tfrac{1}{2}\bar{\alpha}(\partial\cdot A)^2 - \bar{\phi}(i\partial\!\!\!/ - \bar{e}A\!\!\!/)\phi, \tag{111}$$

which is of the same form as the original Lagrangian (1) except that we have replaced α by

$$\bar{\alpha} = \alpha \frac{Z_\alpha}{Z_3} = \alpha \frac{1 + e^2 u/2\pi\alpha}{1 + e^2 v/2\pi} \tag{112}$$

and e by

$$\bar{e} = eZ_3^{-1/2} = e(1 + e^2 v/2\pi)^{-1/2}. \tag{113}$$

The mass arising from $\mathscr{L}_0$ is

$$\bar{m}^2 = \frac{\bar{e}^2}{\pi} = \frac{e^2}{\pi}\left(1 + \frac{e^2 v}{2\pi}\right)^{-1} = m^2\left(1 + \frac{e^2 v}{2\pi}\right)^{-1}. \tag{114}$$

Thus various choices of the parameters u, v lead to equivalent theories.

As a final item we consider the analytic properties of the Wightman functions with respect to the coupling constant.

9. ANALYTICITY IN THE COUPLING CONSTANT

Schwinger's[7] original solution of the Schwinger model was given in terms of perturbation theory. Since then there have been other perturbation theoretic considerations of this model.[8,9] As we now have all the Wightman functions of this model explicitly displayed, it is feasible to examine their

analyticity properties with respect to the coupling constant e.

We begin by considering the $e\to 0$ limit of the various Wightman functions. To accomplish this, we need only consider the two-point function for A_μ, the fermion $2n$-point function, and the mixed three-point function. From Eq. (26) we see that due to the presence of the term $(i\pi/e^2)\,\partial_\mu\partial_\nu \times[\Delta^{(+)} - D^{(+)}]$ the limit $e\to 0$ exists only if we take test functions which vanish for $p_\mu = 0$. In that case we obtain

$$\lim_{e\to 0}\langle 0|A_\lambda(x)A_\nu(0)|0\rangle = \lim_{m\to 0} ig_{\mu\nu}\Delta^{(+)}(m^2, x). \tag{115}$$

On the other hand, using Eqs. (15) and (20) of Ref. 1, we find that

$$\lim_{k\to 0}[\Delta^{(+)}(m^2, x) + (1/\pi i)\ln Z\,] = D^{(+)}(x) \tag{116}$$

so this limit exists.

We next consider the $e\to 0$ limit for the $2n$-point fermion functions given by Eq. (20), namely,

$$W_n(x, y) = Z^{-n}\exp[\mathscr{F}^{(1)}(x, y)]\,w_0^{2n}(x, y).$$

Using Eq. (116), we obtain

$$\lim_{e\to 0} Z^{-n}\exp\mathscr{F}^{+}(x, y)$$
$$= \exp\Bigg\{ -n\ln Z + \ln Z\Bigg[\sum_{i,j=1}^{n}\gamma_{x_i}^5\phi_{y_j}^5 - \sum_{i<j=1}^{n}(\gamma_{x_i}^5\gamma_{x_j}^5 + \gamma_{y_i}^5\gamma_{y_j}^5)\Bigg]\Bigg\}. \tag{117}$$

If we now take the spinor indices of the first $k\leqslant n$ fermion fields to be 1 and the spinor indices of the remaining $n-k$ fermion fields to be 2, then we can evaluate the sums over the γ^5 matrices to get

$$\sum_{i,j=1}^{n}\gamma_{x_i}^5\gamma_{y_j}^5 = k^2 + (n-k)^2 - 2k(n-k) = (n-2k)^2, \tag{118}$$

$$\sum_{i,j=1}^{n}\gamma_{x_i}^5\gamma_{y_j}^5 = \tfrac{1}{2}k(k-1) + \tfrac{1}{2}(n-k)(n-k-1) + k(n-k). \tag{119}$$

Combining these results, Eq. (117) becomes

$$\exp\ln Z\{-n + (n-2k)^2 - k(k-1) - (n-k)(n-k-1) - 2k(n-k)\} = 1. \tag{120}$$

Thus

$$\lim_{e\to 0} W_n(x, y) = w_0^{2n}(x, y). \tag{121}$$

Finally we consider the limit $e\to 0$ for the mixed three-point function given by Eq. (34). To obtain this limit, we must simply consider the limit of the function $G_\mu^{(+)}(x, y)$ given by Eq. (32). Again using Eq. (116), we easily obtain that

$$\lim_{e\to 0} G_\mu^{(+)}(x, y) = 0. \tag{122}$$

Thus the limits of all these Wightman functions exist in the sense of distributions in $\mathscr{S}_0'$ whose test functions are Fourier transforms of functions in $\mathscr{S}$ with their support excluding the origin $p_\mu = 0$. In spite of this, the Wightman functions are not analytic in e. To see this, consider the fermion two-point function

$$\langle 0|\phi(x)\phi^*(y)|0\rangle = Z^{-1}\exp e^2\{-(i/\alpha)I^{(+)}(x-y) - (i\pi/e^2) \times[\Delta^{(+)}(x-y) - D^{(+)}(x-y)]\}w_0^2(x, y), \tag{123}$$

where we have used that $\gamma_x^5\gamma_y^5 = 1$ for this case.

Now for small $m^2 = e^2/\pi$ we have

$$\Delta^{(+)}(m, x) = -\frac{1}{4}H^{(1)}(im(-x^2 + i\epsilon x^0)^{1/2})$$
$$\equiv -\frac{1}{4}H^{(1)}(y)$$
$$= -\frac{1}{4}\Bigg\{J_0(y)\Bigg[1 + \frac{2i}{\pi}\Bigg(\gamma + \ln\frac{y}{2}\Bigg)\Bigg] - \frac{2i}{\pi} \times\sum_{k=0}^{\infty}\frac{(-1)^n(y/2)^{2k}}{(k!)^2}\Bigg(1 + \frac{1}{2} + \cdots + \frac{1}{k}\Bigg)\Bigg\}. \tag{124}$$

This clearly shows that

$$\exp\{-i\pi[\Delta^{(+)}(m, x) + (1/i\pi)\ln Z\,]\} \underset{m\to 0}{\simeq} \exp[-\tfrac{1}{2}J_0(y)\ln(y/2) - \ln Z\,]$$

and has a cut in m.

Thus we find that the coupling constant e is not a suitable expansion parameter around zero. In spite of this, when such an expansion is summed, the correct analytic properties in e are obtained.

10. CONCLUSIONS

We have studied certain properties of the Schwinger model. In particular, we have obtained all the Wightman functions for this model. We have also studied the algebra of fields and representations of this algebra. A particularly interesting object of this model turns out to be the Hamiltonian. It does not consist simply of the Hamiltonian H_0 for the building block fields, although this one does yield the correct time evolution for the algebra of fields. The full Hamiltonian H reflects the "confinement" of the quarks in that the only vestige of the fermions that remains are zero-energy (actually Poincaré-invariant) states in its spectrum.

We also briefly discuss renormalization of the theory and analyticity of the amplitudes in terms of the coupling constant.

ACKNOWLEDGMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. One of us (R. F.) would also like to thank the theoretical Physics Institute of the University of Alberta for support during his visit there.

APPENDIX A: AVERAGING OF POINT SPLITTING

In computing the energy–momentum tensor regularized by point splitting, one obtains expressions of the form

$\eta_{\mu_1}\cdots\eta_{\mu_n}$ in the point splitting parameter η_μ. We prescribe an averaging method over "all directions" of n_μ.

Since the Lorentz group is noncompact, we go to the Euclidean region $\eta_0 \to i\eta_0$ to perform our averaging. In this case keeping the length of η_μ fixed is no problem.

Thus we define the average in the Euclidean region by

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_n}} = \frac{(i)^{\delta_{\mu_1 0}+\cdots+\delta_{\mu_n 0}}}{\pi}\int d^2\eta\,\delta(\eta^2-a^2)\,\eta_{\mu_1}\cdots\eta_{\mu_n}. \tag{A1}$$

Letting $\eta_0 = R\cos\theta$, $\eta_1 = R\sin\theta$, the integral becomes

$$I_n = \int_0^{2\pi} d\theta\,\frac{1}{2}\int R^n\,dR^2\,\delta(R^2-a^2)\,(\cos\theta)^k(\sin\theta)^{(n-k)}, \tag{A2}$$

where we have assumed that k of the μ_i values have the value 0 and the rest have the value 1. It is easy to see that unless n and k are even I_n vanishes. For n, k even we obtain

$$I_n = \frac{a^n}{2}\cdot 2\pi\,\frac{(k-1)!!(n-k-1)!!}{n!!}. \tag{A3}$$

These results now immediately yield:

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_{2n}}}_{\text{Euclidean}} = \frac{a^{2n}(i)^{\delta_{\mu_1 0}+\cdots+\delta_{\mu_{2n}0}}}{(2n)!!}\sum_{\substack{\text{partitions}\\ \text{in } n \text{ pairs}}}\delta_{\mu_i\mu_j}\cdots\delta_{\mu_k\mu_l}, \tag{A4}$$

which in Minkowsky space becomes

$$\overline{\eta_{\mu_1}\cdots\eta_{\mu_{2n}}} = \frac{(\eta^2)^n}{(2n)!!} + \sum_{\text{partitions}} g_{\mu_i\mu_j}\cdots g_{\mu_k\mu_l}, \tag{A5}$$

where the sum is over the partitions of the $2n$ indices into pairs (μ_i, μ_j) with $i<j$. Furthermore, we immediately find that the "average" of an odd product of η_μ's vanishes.

APPENDIX B: EQUAL-TIME COMMUTATORS FOR BUILDING-BLOCK FIELDS

Using the various commutators for the building-block fields, one easily finds the following useful equal-time (anti-) commutation relations:

$$[\dot a(x), b(0)]_{x^0=0} = -i/\alpha\delta(x^1), \tag{B1}$$

$$[\dot a(x), a(0)]_{x^0=0} = i(\beta^2 + 2\beta/m)\delta(x^1), \tag{B2}$$

$$[\dot b(x), a(0)]_{x^0=0} = -(i/\alpha)\delta(x^1), \tag{B3}$$

$$[\Sigma(x), \Sigma(0)]_{x^0=0} = -i\delta(x^1), \tag{B4}$$

$$[\dot\rho(x), \rho(0)]_{x^0=0} = -i\delta(x^1), \tag{B5}$$

$$[\partial^1\sigma(x), \rho(0)]_{x^0=0} = i\delta(x^1), \tag{B6}$$

$$[\partial^1\rho(x), \sigma(0)]_{x^0=0} = i\delta(x^1), \tag{B7}$$

$$[\dot\sigma(x), \sigma(0)]_{x^0=0} = -i\delta(x^1), \tag{B8}$$

$$\{\psi(x), \bar\psi(0)\}_{x^0=0} = \gamma^0\delta(x^1), \tag{B9}$$

$$\{\bar\psi(x), \psi(0)\}_{x^0=0} = \gamma^0\delta(x^1), \tag{B10}$$

$$[\dot\rho(x), \psi(y)]_{x^0=y^0} = -\sqrt{\pi}\psi(y)\delta(x^1), \tag{B11}$$

$$[\partial^1\rho(x), \psi(y)]_{x^0=y^0} = -\sqrt{\pi}\,\gamma^5\psi(y)\delta(x^1), \tag{B12}$$

$$[\dot\sigma(x), \psi(y)]_{x^0=y^0} = \sqrt{\pi}\,\gamma^5\psi(y)\delta(x^1), \tag{B13}$$

$$[\partial^1\sigma(x), \psi(y)]_{x^0=y^0} = \sqrt{\pi}\,\psi(y)\delta(x^1). \tag{B14}$$

Moreover, we find that

$$[:\bar\psi\gamma^1\partial^1\psi:(y), \rho(x)]_{y^0=x^0} = -\dot\rho(x)\delta(x^1-y^1) \tag{B15}$$

and

$$[:\bar\psi\gamma^1\partial^1\psi:(y), \sigma(x)]_{y^0=x^0} = -\dot\sigma(x)\delta(x^1-y^1). \tag{B16}$$

APPENDIX C: THE LANDAU GAUGE

We briefly consider the Landau gauge here. It is obtained as the $\alpha\to\infty$ limit of our solutions if one keeps

$$b_0(x) \equiv \alpha b(x) \tag{C1}$$

fixed.

It then follows from Eq. (79) that

$$a_0(x) = a(x) - b_0(x)/m^2 + (\beta + 1/m)\rho(x) \tag{C2}$$

and

$$\langle 0|a_0(x)a_0(0)|0\rangle = (i/m^2)D^{(+)}(x), \tag{C3}$$

whereas

$$\langle 0|a_0(x)b_0(0)|0\rangle = -iD^{(+)}(x) \tag{C4}$$

and

$$\langle 0|b_0(x)b_0(0)|0\rangle = 0, \tag{C5}$$

$$\langle 0|a_0(x)\tilde b_0(0)|0\rangle = -i\tilde D^{(+)}(x). \tag{C6}$$

Moreover, both a_0 and b_0 are scalar fields:

$$\Box a_0 = 0, \quad \Box b_0 = 0. \tag{C7}$$

The fields ϕ and A_μ are now given by

$$\phi(x) = :\exp\Big\{-ie\Big[a_0(x) + \frac{b_0(x)}{m^2} + \frac{1}{m}\Big(\Sigma(x) - \frac{\tilde b_0(x)}{m}\Big)\gamma_x^5\Big]:\zeta(x)\Big\}, \tag{C8}$$

$$A_\mu = \partial_\mu a_0 + (1/m)\epsilon_{\mu\nu}\partial^\nu\Sigma, \tag{C9}$$

where ζ is still given by Eq. (80). The content of these equations is clarified if instead of the two massless scalar fields a_0, b_0 we introduce two commuting *massless* scalar fields

$$a_1 = ma_0 + (1/m)b_0, \quad a_2 = ma_0, \tag{C10}$$

$$b_0 = m(a_1 - a_2), \quad a_0 = (1/m)a_2. \tag{C11}$$

We then find

$$\langle 0|a_1(x)a_1(0)|0\rangle = -iD^{(+)}(x), \tag{C12}$$

$$\langle 0|a_1(x)a_2(0)|0\rangle = 0, \tag{C13}$$

$$\langle 0|a_2(x)a_2(0)|0\rangle = +iD^{(+)}(x). \tag{C14}$$

Thus the field a_2 carries a *negative* norm.

[1] A. Z. Capri and R. Ferrari, Nuovo Cimento A **62**, 273 (1981).
[2] B. Klaiber, *Boulder Lectures in Theoretical Physics* (Gordon and Breach, New York, 1967), Vol. XA, p. 141.
[3] N. Nakanishi, Prog. Theor. Phys. **57**, 1025 (1977).
[4] C. M. Sommerfield, Ann. Phys. (N. Y.) **26**, 1 (1964).
[5] R. Ferrari, Nuovo Cimento A **19**, 204 (1974).
[6] A. Z. Capri, G. Grübl, and R. Kobes, Ann. Phys. (N. Y.) **147**, 140 (1983).
[7] J. Schwinger, Phys. Rev. **128**, 2425 (1962).
[8] P. Becher and H. Joos, "(1 + 1)-Dimensional quantum electrodynamics," DESY Preprint 77/43, 1977.
[9] I. O. Stamatescu and T. T. Wu, Nucl. Phys. B **143**, 503 (1978).

A novel mass-eigenvalue problem for spinors in deSitter space

Edward H. Kerner
Sharp Physics Laboratory, University of Delaware, Newark, Delaware 19711

(Received 6 May 1983; accepted for publication 26 August 1983)

It is shown that an unambiguous quantum theory of spinors in positively curved deSitter space, based on distinguished coordinates in a Hamiltonian framework, leads to a set of spinors corresponding to unsharp energy but sharp mass defined in a family of novel eigenvalue problems. An example is given in which partly real and partly complex discrete mass spectra come forth.

PACS numbers: 11.10.Qr, 04.90. + e

Spinors in spaces of constant curvature [deSitter spaces of O(3,2) or O(4,1) symmetry] have received continuing attention[1] for nearly fifty years. Their structure is of interest not only in its own right since deSitter spaces are the physically distinguished ones having maximal (tenfold) symmetry, but also because they are local osculating spaces[2] (rather than mere tangent spaces) to more generally curved Riemann spaces, attaining thereby a prototypical role. Further, they form background spaces for supersymmetry,[3] and have been broached[4] as closed up "microuniverses" for considering particle confinement at a basic geometrical level.

In the present paper an unusual family of eigenvalue problems is brought out for the mass of a spinning particle running along a geodesic of O(3,2) deSitter space. This results from a well-set and essentially unique Hamiltonian formulation of the motion developed in recent years,[5] in contrast to the formal spinor theories usually invoked.[1]

In the latter, governed by general convariance considerations, Klein–Gordon equations are typically factorized to curved-space Dirac equations $(\gamma^{\mu}(x)\nabla_{\mu} + m)\psi = 0$ as a matter of formal prescription (∇_{μ} = covariant derivative). The coordinates remain ambiguous, and of course commutation rules are renounced. The Hamitonian formulation, on the other hand, relies on distinguished coordinates and proceeds through clear commutation rules to a quite unambiguous statement of quantum theory. The basis here is a specialized subgroup of the projective transformations $x_i' = \Lambda_i(x,a)/\Delta(x,a) \equiv \Gamma_i(x)$, with Λ_i and Δ inhomogeneous linear functions of space Cartesians $x_1,x_2,x_3 = \mathbf{r}$ and time $x_0 = t$, and a = a universal length. These are isomorphic to the deSitter group of pseudorotations O(3,2) in the five-space of homogeneous coordinates $X_i, U(x_i \equiv X_i/U)$. What is notable is that x' and x are in the relationship of coordinates of inertial frames, since $d^2\mathbf{r}'/dt'^2 = 0$ is sent into $d^2\mathbf{r}/dt^2 = 0$ and conversely, making these coordinates clearly distinguished above all others. While the appropriate invariant line element indeed describes constant curvature $1/a^2$, the geodesics one and all are the global free-particle motions $d^2\mathbf{r}/dt^2 = 0$. Given this order of simplicity, general covariance is rendered irrelevant, and only the automorphism of space-time under $x' = \Gamma(x)$ is consequential, as with the automorphism of Minkowski space under the Poincaré group. Coordinate ambiguities and equivocal quantization recipes may then be set aside, and instead the usual commutation rules $(x_i, p_j^0) = i\hbar\delta_{ij}$, etc. $(i,j = 1,2,3)$ tenably introduced as the primary physical hypothesis for the quantum dynamics of a free particle, in accord with all physical experience.

Useful coordinate transformations can now (*post* settlement of the physical basis) be performed, such as $\boldsymbol{\rho}(\mathbf{r},t)$ and $\tau(t)$ described earlier,[5] that rephrases the straights $d^2\mathbf{r}/dt^2 = 0$ as the harmonic-oscillator geodesics $d^2\boldsymbol{\rho}/d\tau^2 + (c^2/a^2)\boldsymbol{\rho} = 0$ otherwise familiar in deSitter space, and that gives a ladder spectrum of Klein–Gordon energy eigenvalues. The further transformation $\mathbf{R} = \boldsymbol{\rho}/(1 - \rho^2/a^2)^{1/2}$ brings the Hamiltonian-squared

$$\frac{H^2}{c^2} = \left[\mathbf{P}^2 + \frac{\mathbf{L}^2}{a^2} + \frac{\hbar^2}{a^2}\right] + \kappa^2\frac{\hbar^2}{a^2}\left[1 + \frac{\mathbf{R}^2}{a^2}\right] \equiv H_1^2 + \kappa^2 H_2^2, \tag{1}$$

$$\kappa^2 = m^2c^2a^2/\hbar^2 - \tfrac{1}{4}, \quad \mathbf{P} \equiv \tfrac{1}{2}(\hat{I} + \mathbf{RR}/a^2)\cdot\mathbf{P}_c + \text{h.c.},$$

where $\mathbf{P}_c$ is canonical mate $-i\hbar\nabla_R$ to $\mathbf{R}$, and $\mathbf{L}$ is $\mathbf{R}\times\mathbf{P}$, with $\hat{I}$ the unit dyadic.

This reduction forces into particularly clear view the issue of linearization to determine H upon the primary physical basis, an issue distinct from generally covariant factorization of $\nabla^{\mu}\nabla_{\mu} + m^2$. As has been remarked,[6] there does not exist any ordinary matrix squareroot of $H_1^2 + \kappa^2 H_2^2$ in Dirac matrices or otherwise (except for $\kappa = 0$). Since this point is central to any consideration of spinor theory on a Hamiltonian base, the proof will be briefly reviewed.

Taking $\hbar,c,a = 1$ from here on, the one-dimensional form of Eq. (1) already reveals the difficulty:

$$H^2 = P^2 + \kappa^2(1 + X^2),$$

(where both the terms $\mathbf{L}^2/a^2$ and $\hbar^2/a^2$ are to be dropped in one dimension). If H is $F(x)P + G(X)$, it is then required that

$$F^2 = 1,$$

$$FG + GF = iFF',$$

$$G^2 - iFG' = \kappa^2(1 + X^2),$$

be identically satisfied in X, where F' means $(1 + X^2)dF/dX$ and similarly for G'. Multiply the second, right and left, by F. This brings $FF' = F'F$, while the first states that $FF' + F'F = 0$. Hence $FF' = 0 = F'F$, so that $F' = 0$ and $FG + GF = 0$. Now multiply the third, right and left, by F, producing $FG' = G'F$. But $(FG + GF)' = FG' + G'F = 0$, whence $G'F = 0 = FG'$. Consequently $G' = 0$, and then G = const, cannot satisfy the third (except for $\kappa = 0$).

In short, while H_1^2 and H_2^2 are separately Dirac linearizable,[5] for example as

$$H_1 = \boldsymbol{\alpha}\cdot\mathbf{P} - \boldsymbol{\sigma}\cdot\mathbf{L} - 1,$$

$$H_2 = \beta + i\beta\boldsymbol{\alpha}\cdot\mathbf{R}, \tag{2}$$

 0022-2488/84/010150-04$02.50

with standard Dirac matrices β,α,σ, the pieces H_1 and H_2 are *incompatible* in that they cannot, in general, be brought together to give a general overall linear Hamiltonian. The choices for H_1 and H_2 above are not unique but are here selected for simplicity. (A second possibility for H_1 is $\alpha\cdot\mathbf{P}+\alpha\cdot\mathbf{L}-i\gamma_5$, while the roots of $1+R^2$ for H_2 are very numerous; but in all cases a single general Hamiltonian is ruled out.)

To hold to the Hamiltonian framework, and accord to the Hamiltonian its master dynamical role of generator of time (τ) translations, is nevertheless achievable (notwithstanding the incompatibility of H_1 and H_2), provided the spinors to be considered are suitably restricted, and as well the value of the mass parameter $\kappa=(m^2-\frac{1}{4})^{1/2}$.

Clearly, if ψ is a spinor such that

$$H_1\psi=\lambda H_2\psi, \tag{3}$$

then for these spinors an overall linearization of H becomes possible,

$$i\frac{\partial\psi}{\partial\tau}=H\psi=(\alpha_1H_1+\alpha_2H_2)\psi, \tag{4}$$

($\lambda,\alpha_1,\alpha_2$ numerical parameters) since

$$H^2\psi=[\alpha_1^2H_1^2+\alpha_1\alpha_2(H_1H_2+H_2H_1)+\alpha_2^2H_2^2]\psi$$
$$=[(\alpha_1^2+\alpha_1\alpha_2/\lambda)H_1^2+(\alpha_2^2+\lambda\alpha_1\alpha_2)H_2^2]\psi.$$

This requires only that

$$\alpha_1^2+\alpha_1\alpha_2/\lambda=1,\quad \alpha_2^2+\lambda\alpha_1\alpha_2=\kappa^2,$$

or that

$$\alpha_1=(1+\kappa^2/\lambda^2)^{-1/2},\quad \alpha_2=(\kappa^2/\lambda)(1+\kappa^2/\lambda^2)^{-1/2},$$

bringing Eq. (4) to

$$i\frac{\partial\psi}{\partial\tau}=\left(1+\frac{\kappa^2}{\lambda^2}\right)^{1/2}H_1\psi \tag{5}$$
$$=\lambda\left(1+\frac{\kappa^2}{\lambda^2}\right)^{1/2}H_2\psi. \tag{6}$$

If ϕ is some initial spinor, one gets [$\zeta\equiv(1+\kappa^2/\lambda^2)^{1/2}$]

$$\psi=[\exp(-i\zeta H_1\tau)\phi=[\exp(-i\lambda\zeta H_2\tau)]\phi,$$

so that this initial state is constrained to satisfy

$$H_1\phi=\lambda H_2\phi. \tag{7}$$

Stationary states are here ruled out.

As will be shown below, Eq. (7) does not allow arbitrary λ or arbitrary ϕ; rather a discrete spectrum of eigenvalues λ_j and eigenstates ϕ_j is demanded. But then in Eqs. (5) and (6) the operators $(1+\kappa^2/\lambda_j^2)^{1/2}H_1$ or $\lambda_j(1+\kappa^2/\lambda_j^2)^{1/2}H_2$ are not uniquely valued [i.e., are not independent of the index j labeling the eigensolutions of Eq. (7)] unless κ is restricted. Taking uniquely valued spinor wave equations as a basic requirement, two mutually exclusive restrictions on κ stand forth, which may be called cases (A) and (B). These correspond to

$$(1+\kappa^2/\lambda_j^2)^{1/2}=\beta_1 \qquad \text{(A)}$$

or

$$\lambda_j(1+\kappa^2/\lambda_j^2)^{1/2}=\beta_2, \qquad \text{(B)}$$

where β_1, β_2 are arbitrary real numbers independent of the label j. Not both of (A) and (B) can be allowed simultaneously since $\beta_2/\beta_1=\lambda_j$ is ruled out. Then

$$\kappa_j^2=(\beta_1^2-1)\lambda_j^2 \quad \text{(A)}$$

or (8)

$$\kappa_j^2=\beta_2^2-\lambda_j^2, \qquad \text{(B)},$$

prescribe the allowed mass spectra, while the uniquely valued spinor wave equations are

$$i\frac{\partial\psi}{\partial\tau}=\beta_1H_1\psi, \qquad \text{(A)}$$

or

$$i\frac{\partial\psi}{\partial\tau}=\beta_2H_2\psi, \qquad \text{(B)}$$

with β_iH_i remaining Hermitian when H_i are Hermitian (β_1,β_2 may be absorbed into scale changes in τ if desired). It is easily demonstrated that $\mathbf{J}=\mathbf{L}+\frac{1}{2}\sigma$ commutes with both H_1, H_2 of Eq. (2), so that ψ may be an eigenstate of total angular momentum, but it clearly cannot be an eigenstate of energy (either H_1 or H_2).

We may summarize as follows: *Within the Hamiltonian framework in deSitter space, spinors exist which are not eigenstates of the Hamiltonian but rather are eigenstates of a "mass-generating operator" $H_2^{-1}H_1$* [Eq. (7)] *whose eigenvalues prescribe a family of allowed masses (Eq. 8) and whose elements H_1, H_2 are Dirac square roots of well-defined operators within that framework.* In a word, these particular states are unsharp in energy but sharp in mass. To the extent that one may regard the parameters β_1, β_2 as running freely over their real values, the mass spectra are of the nature of bands, with individual bands labeled discretely according to the eigenvalues of the $H_2^{-1}H_1$ operator.

A further perspective on the reduction given above is sketched in the Appendix, where a novel square root process[6] for $H_1^2+\kappa^2H_2^2$ in total is reviewed, and the case where $\lambda=\kappa$ is particularly obtained.

Turning to the eigenvalue problem of λ, Eq. (7), we may use H_1 and H_2 from Eq. (2) as an example. In view of the many possible choices for H_i noted before, this will be understood to be primarily illustrative rather than exhaustive or definitive, demonstrating the principal point that λ has a discrete spectrum. Since the H_1, H_2 of Eq. (2) do not commute, the mass generator $H_2^{-1}H_1$ in $H_2^{-1}H_1\phi=\lambda\phi$ is not Hermitian, so λ cannot be expected to have a completely real spectrum in the present example.

Eq. (7) is readily analyzed upon recognizing certain structural similarities to the classical Dirac–Coulomb problem as set forth particularly by Foldy.[7] First it is convenient to return to the harmonic-oscillator coordinate $\boldsymbol{\rho}$ or ρ,θ,ϕ in polar coordinates ($0\leqslant\rho\leqslant1$) with corresponding momentum $\mathbf{p}=-i\nabla_\rho$. Then employing Foldy's operators

$$\hat{k}=\beta(\sigma\cdot\mathbf{L}+1),$$
$$\alpha_\rho=\alpha\cdot\boldsymbol{\rho}/\rho,$$
$$P_\rho=(1/\rho)(\boldsymbol{\rho}\cdot\mathbf{p}-i),$$

the operators H_1, H_2 are

$$H_1=(1-\rho^2)^{1/2}(\alpha_\rho\,p_\rho+(i/\rho)\alpha_\rho\,\beta\hat{k})$$
$$+\tfrac{1}{2}i\rho\alpha_\rho/(1-\rho^2)^{1/2}-\beta\hat{k},$$
$$H_2=\beta+i\beta\alpha_\rho\,\rho/(1-\rho^2)^{1/2}.$$

The operators $\beta, \hat{k}, L_2, J_z$ are intercommuting and their common eigenvector, which depends only on θ and ϕ, may be designated as ξ, belonging respectively to the eigenvalues $1, k, l(l+1), m_j$. A second angular spin function $\eta \equiv i\alpha_\rho \xi$ is also an eigenvector of $\hat{k}$ and J_z with the same eigenvalues k and m_j as ξ [though it is not an eigenvector of L^2 belonging to $l(l+1)$]. Since η is an eigenvector of β belonging to the eigenvalues -1, it is orthogonal to ξ. Hence when one introduces

$$\phi = (f(\rho)/\rho)\xi + (g(\rho)/\rho)\eta,$$

into $H_1\phi = \lambda H_2\phi$, one obtains terms only in ξ and η, and thence by their orthogonality, the pair of coupled radial equations

$$\frac{df}{d\rho} + \left(-\frac{k}{\rho} - \frac{1}{2}\frac{\rho}{1-\rho^2} - \lambda\frac{\rho}{1-\rho^2}\right)f + \frac{k+\lambda}{(1-\rho^2)^{1/2}}g = 0,$$

$$\frac{dg}{d\rho} + \left(\frac{k}{\rho} - \frac{1}{2}\frac{\rho}{1-\rho^2} + \lambda\frac{\rho}{1-\rho^2}\right)g + \frac{k+\lambda}{(1-\rho^2)^{1/2}}f = 0. \tag{9}$$

Here k is an eigenvalue of $\hat{k}$, namely $k^2 = (j+\frac{1}{2})^2$ with $j = \frac{1}{2}, \frac{3}{2}, \ldots$, that is, $k = \pm 1, \pm 2, \ldots$ or $|k| \equiv s = 1,2,\ldots$.

The normalization of ϕ is defined by

$$\int_0^1 \frac{|f|^2 + |g|^2}{\rho^2}\frac{\rho^2 d\rho}{(1-\rho^2)^{5/2}} = 1, \tag{10}$$

when ξ and η are normalized according to

$$\int \xi^+\xi \sin\theta\, d\theta\, d\phi = 1 = \int \eta^+\eta \sin\theta\, d\theta\, d\phi,$$

where the factor $(1-\rho^2)^{-5/2}$ comes from the invariant line element in ρ,τ variables that prescribe the invariant volume element $(1-\rho^2)^{-5/2}d\rho\, d\tau$ in deSitter space. Consequently f and g must be regular at $\rho = 0$ and vanish sufficiently fast at $\rho = 1$.

One very simple solution to Eqs. (9) stands out at once in the case $k + \lambda = 0$,

$$f = \rho^k(1-\rho^2)^{-1/4-(1/2)\lambda},$$
$$g = \rho^{-k}(1-\rho^2)^{-1/4+(1/2)\lambda}.$$

Not both of these may be retained, but only

$$f = 0 \quad g = \rho^s(1-\rho^2)^{(1/2)s-1/4}$$

or

$$g = 0 \quad f = \rho^s(1-\rho^2)^{(1/2)s-1/4}$$

with eigenvalues

$$\lambda^2(s) = s^2 = 9,16,\ldots$$

for $s = 3,4,\ldots$ in view of Eq. (10).

Proceeding to the general situation, write

$$f = (1-\rho^2)^{1/4}F, \quad g = (1-\rho^2)^{-1/4}G$$

to get rid of roots of $1-\rho^2$,

$$F' - \left(\frac{k}{\rho} + (1+\lambda)\frac{\rho}{1-\rho^2}\right)F + \frac{k+\lambda}{1-\rho^2}G = 0,$$

$$G' + \left(\frac{k}{\rho} + \lambda\frac{\rho}{1-\rho^2}\right)G + (k+\lambda)F = 0,$$

and decouple to obtain a second-order equation in G alone,

$$G'' - \frac{\rho}{1-\rho^2}G' + \left[\frac{-k(k+1)}{\rho^2} - \frac{(k+\lambda)^2 + 2k\lambda + k}{1-\rho^2} + \frac{\lambda - \lambda^2\rho^2}{(1-\rho^2)^2}\right]G = 0.$$

Now extract the characteristic behavior at $\rho = 0$ and $\rho^2 = 1$ through

$$G = \rho^\alpha(1-\rho^2)^\beta S$$

to obtain the indicial roots

$$\alpha = -k,\ k+1,$$
$$\beta = \tfrac{1}{2}\lambda,\ \tfrac{1}{2}(1-\lambda),$$

with S satisfying the differential equation of essentially hypergeometric type

$$S'' + \left(\frac{2\alpha}{\rho} - \frac{(1+4\beta)\rho}{1-\rho^2}\right)S' - \frac{\gamma}{1-\rho^2}S = 0$$
$$\gamma \equiv (k+\lambda)^2 + 2k\lambda + k + \alpha + 2\beta + 4\alpha\beta - \lambda.$$

In the customary way, the series solution $S = \Sigma a_\nu \rho^\nu$ produces the recursion

$$\frac{a_{\nu+2}}{a_\nu} = \frac{(\nu+\alpha+2\beta+q)(\nu+\alpha+2\beta-q)}{(\nu+2)(\nu+1+2\alpha)}, \tag{11}$$

$$q^2 = (\alpha+2\beta)^2 - \gamma = -4k\lambda.$$

The even and odd solutions here are then

$$S_e = {}_2F_1[(\alpha+2\beta+q)/2,(\alpha+2\beta-q)/2;\alpha+\tfrac{1}{2};\rho^2),$$
$$S_o = \rho_3F_2(1,(1+\alpha+2\beta+q)/2,(1+\alpha+2\beta-q)/2;$$
$$\tfrac{3}{2},1+\alpha;\rho^2).$$

The recursion relation Eq. (11) shows that S behaves like $(1-\rho^2)^{1/2-2\beta}$ near $\rho = 1$. This overwhelms the factor $(1-\rho^2)^\beta$ in G when at the outset $\mathrm{Re}(\beta)$ is taken as positive to ensure that G vanishes appropriately at $\rho = 1$. Hence the S series must be broken off in a polynomial,

$$n + \alpha + 2\beta \pm q = 0,$$
$$n = 0,1,2,\ldots .$$

Therefore when $\alpha = -k$ (k negative) $= s$ and $\beta = \lambda/2$ one obtains the λ spectrum

$$\lambda^2 + 2\lambda(n-s) + (n+s)^2 = 0,$$
$$\lambda(s,n) = s - n \pm 2i\sqrt{sn},$$

requiring $s \geqslant n+3$ for satisfactory $\mathrm{Re}(\beta) > 0$ [the root $\beta = (1-\lambda)/2$ of the indicial equation is ruled out].

Similarly, when $\alpha = k+1$ (k positive) and $\beta = (1-\lambda)/2$, the λ spectrum is

$$\lambda(k,n) = n - k + 2 \pm 2i\sqrt{k(n+2)},$$

with $k \geqslant n+4$ for suitable $\mathrm{Re}(\beta)$ (the indicial root $\beta = \lambda/2$ being ruled out here). The case $\alpha = 0$ ($k = -1$) is not allowed.

This concludes the illustration of how the mass generator $H_2^{-1}H_1$ eventuates in a spectrum of eigenvalues $\lambda(s)$,

$\lambda(s,n)$, $\lambda(k,n)$ and corresponding spinors belonging to sharp masses. In the case of the real eigenvalue $\lambda(s)$, the mass spectra according to Eq. (8) are

$$m_j^2 = (\beta_1^2 - 1)(j+\tfrac{1}{2})^2 + \tfrac{1}{4} \qquad \text{(A)},$$

or

$$m_j^2 = \beta_2^2 + \tfrac{1}{4} - (j+\tfrac{1}{2})^2, \qquad \text{(B)}$$
$$j = \tfrac{5}{2},\tfrac{7}{2},\ldots,$$

where (A) describes an infinite real discrete spectrum for $\beta_1^2 > 1$ and a finite real spectrum for $\beta_1^2 - 1$ small and negative; while (B) describes a finite real spectrum for adequately large β_2. Corresponding mass *bands* are defined when β_1, β_2 are allowed to range freely. The complex eigenvalues $\lambda(s,n)$, $\lambda(k,n)$ of course do not admit ready interpretation [though perhaps hinting to a later discrete spectrum of (composite) particle decay times accompanying discrete masses]. Indeed the meaning of *mass* altogether in such totally closed up or 'interior' geometry as that of O(3,2) remains in issue until that geometry is clarified as a locale of an 'exterior' large-scale geometry suited to physical observation.

ACKNOWLEDGMENT

My thanks go to the U.S. Department of Energy for its partial support of this work.

APPENDIX

The fundamental eigenvalue problem $H_1\phi = \lambda H_2\phi$ of the present work also occurs, for $\lambda = \kappa$, upon introducing[6] a novel square root process for

$$P_\tau^2\psi = (H_1^2 + \kappa^2 H_2^2)\psi,$$

($P_\tau = i\partial/\partial\tau$). Namely the linearization

$$I \otimes N_0\, P_\tau\psi = (H_1 \otimes N_1 + \kappa H_2 \otimes N_2)\psi$$

is feasible in that iteration produces

$$(I \otimes N_0)^2\, P_\tau^2\psi = (I \otimes N_0)^2(H_1^2 + \kappa^2 H_2^2)\psi,$$

when $N_0^2 = N_1^2 = N_2^2$, and (to overcome the incompatibility of H_1, H_2) $N_1N_2 = 0 = N_2N_1$. That is, the N_i are suitable *singular* matrices which are nilpotent like $N_i^3 = 0$. The analysis shows[6] that N_i must be at least 4×4 and then of typical structure $N_i = n_i T$ (upon enforcing n_i Hermitian and T unitary)

$$n_1 = \begin{array}{|cccc} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot \\ \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{array} = 0 \oplus \lambda_i,$$

$$n_2 = \begin{array}{|cccc} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & 1 & \cdot \end{array} = 0 \oplus \lambda_6.$$

Here dots stand for zeroes, $\otimes$ for direct product, $\oplus$ for direct sum, and $N_0 = (N_1 + N_2)/\sqrt{2}$ while $T_{13}, T_{22}, T_{31}, T_{41} = 1$ ($T_{ij} = 0$ otherwise) and λ_1, λ_6 are two of the conventional generators of SU(3) [other SU(3) generators and other unitary T are possible, as are higher-dimensional $n_i = 0 \oplus$ SU(N) for all $N \geqslant 3$ but $N < 3$ is ruled out]. In short unitary spin comes forth quite directly in a fusion with Dirac spin, and here is not an ad hoc appendage.

The unitary transform $\Phi = I \otimes T\psi$ brings the linearized wave equation

$$I \otimes n_0\, P_\tau\Phi = (H_1 \otimes n_1 + H_2 \otimes n_2)\Phi, \qquad (12)$$

with $n_0 = (n_1 + n_2)/\sqrt{2}$. Introducing Φ as col$(\Phi_a, \Phi_b, \Phi_c, \Phi_d)$ with indices tied to the n-matrices, Φ_a of course falls aside, leaving Eq. (12) as

$$i\Phi_c' = H_1\Phi_c,$$
$$i(\Phi_b' + \Phi_d') = H_1\Phi_b + \kappa H_2\Phi_d,$$
$$i\Phi_c' = \kappa H_2\Phi_c,$$

where $\Phi' = \partial\Phi/\partial\tau'$ ($\tau' = \tau\sqrt{2}$). It is sufficient to span unitary-spin space, to take $i\Phi_b' = H_1\Phi_b$ and $i\Phi_d' = \kappa H_2\Phi_d$, leaving

$$\Phi_c = [\exp(-iH_1\tau')]\phi = [\exp(-i\kappa H_2\tau')]\phi,$$

and requiring

$$H_1\phi = \kappa H_2\phi, \qquad (13)$$

as in Eq. (7).

Hence in the present spin $\otimes$ unitary-spin scheme the mass parameter κ is directly fixed as eigenvalue of Eq. (13), for example $\kappa = s = |j + \tfrac{1}{2}|$ for $j = \tfrac{5}{2},\tfrac{7}{2},\ldots$ as before. In the latter case,

$$m_j^2 = j(j+1) + \tfrac{1}{2} = \tfrac{37}{4},\tfrac{65}{4},\ldots .$$

This result resembles that of Barut and Böhm[8] for a so-called deSitter "rotator," which, however, stems not from O(3,2) but from O(4,1), and refers not to a particle but to a composite system.

[1] P. A. M. Dirac, Ann. of Math. **36**, 657 (1935); F. Gürsey and T. D. Lee, Proc. Natl. Acad. Sci. **49**, 179 (1963); O. Nachtmann, Comm. Math. Phys. **6**, 1 (1967); G. Börner and H. P. Dürr, Nuovo Cimento A **64**, 669 (1969); M. S. Drew, Ann. Phys. (N.Y.) **103**, 469 (1977).
[2] M. D. Maia, J. Math. Phys. **22**, 538 (1981).
[3] S. Deser and B. Zumino, Phys. Rev. Lett. **38**, 1433 (1977).
[4] A. Salam and J. Strathdee, Phys. Rev. D **18**, 4596 (1978); C. Sivaram and K. P. Sinha, Phys. Rep. **51**, 111 (1979).
[5] E. H. Kerner, Phys. Rev. D **22**, 280 (1980).
[6] E. H. Kerner, Phys. Rev. D **26**, 390 (1982).
[7] L. L. Foldy, in *Quantum Theory III*, edited by D. R. Bates (Academic, New York, 1962), p. 29.
[8] A. O. Barut and A. Böhm, Phys. Rev. **139**, B1107 (1965).

Vortex properties in first- and second-order formulations of abelian gauge theories

John van der Hoek
Department of Pure Mathematics, The University of Adelaide, Adelaide, South Australia, 5000

M. A. Lohe
The Flinders University of South Australia, School of Mathematical Sciences, Bedford Park, South Australia, 5042

(Received 22 October 1982; accepted for publication 7 January 1983)

Properties of noninteracting vortices in a class of models which generalize the Ginzburg–Landau model of superconductivity are described. Previous results of existence and uniqueness for solutions to the first-order equations are extended to cover the case in which the gauge photon and the scalar meson become massless, when long range interactions exist. Several properties of the solutions are also discussed. With some assumptions, and with restrictions on the class of models, all finite-energy solutions of the second-order equations are shown to be solutions of the first-order equations. The second-order equations are formulated in a gauge invariant way, resulting in a second-order elliptic system of two coupled nonlinear equations, which completely determine all gauge invariant quantities.

PACS numbers: 11.15. – q, 74.20.De

I. INTRODUCTION

Finite-energy solutions in field theories are of importance because they serve as good starting approximations for the quantum field theory. For nonabelian gauge theories in three space dimensions these solutions are magnetic monopoles, and detailed properties of these monopoles and their interactions are obtained from a study of the relevant field equations. The simplest of the gauge theories with nontrivial finite energy solutions is the abelian Higgs model in two dimensions, for which the static equations are the Ginzburg–Landau equations of superconductivity. A detailed study of the static solutions (vortices) has been undertaken in Refs. 1–3. Of particular interest is the noninteracting case when the coupling constant λ assumes a critical value ($\lambda = 1$); for this value, static solutions exist which describe vortices located at arbitrary positions in the plane. Evidently, the opposing forces due to the massive gauge photon and the scalar (Higgs) meson cancel exactly.

In Refs. 4 and 5 a model has been described which generalizes the Ginzburg–Landau equations by incorporating into the model an arbitrary nonnegative function $F(|\phi|)$ of the scalar field ϕ. This generalization is of interest because it preserves the noninteracting nature of the vortices; properties of the Ginzburg–Landau equations are revealed to be special cases of similar properties for the general system. Solutions can be found by solving three first order equations, and in Ref. 5 solutions were not shown to exist which describe, as for the Ginzburg–Landau equations, vortices located at arbitrary positions in the plane.

In this paper we extend our previous analysis of the generalized system. First, we strengthen results[5] on the existence and uniqueness to include a class of solutions of particular interest. As mentioned above, in general the class of models we consider share features similar to those of the Ginzburg–Landau theory, which appears as the special case $F(|\phi|)\equiv 1$. An exception arises when $F(|\phi|)$ assumes an asymptotic value $F(1)$, which is zero. The masses of the photon and the scalar meson, which are equal for the noninteracting theory, are given by the value of $F(1)$ so that for $F(1) = 0$ we have massless particles. Instead of the short-range interaction experienced by the massive particles, we now have long-range interactions, with the fields decaying to their asymptotic values according to an inverse power law. In Sec. III we demonstrate the existence and uniqueness of solutions to the first-order equations under very general circumstances, including also the massless case, and dispensing with the assumptions of Ref. 5, excepting, of course, the finite-energy condition. Here we draw on the results of Benilan, Brezis, and Crandall[6] and recent work by Vazquez,[7] which investigates equations of the form

$$-\Delta u + \beta(u) \quad \ni \quad g \quad \text{on } \mathbf{R}^N, \tag{1.1}$$

where $\beta(u)$ is a maximal monotone graph and g is a measure. This equation is precisely of the type which appears in Sec. III. Also discussed in Sec. III are several properties of the solutions, including asymptotic estimates.

Now, we turn attention to the full second-order equations obtained by varying the Lagrangian for the generalized model. We pose the question as to whether all finite-energy solutions of the second-order equations are also solutions of the first-order equations. For the Ginzburg–Landau theory the answer is in the affirmative,[2,3] and we extend this result, using maximum principle type arguments, to the general case provided some assumptions are made on $F(|\phi|)$. One assumption is a growth condition on F, which enables us to conclude that $|\phi|$ is bounded, and another assumption, $F > 0$, is also necessary but excludes the massless case. A convenient feature of the abelian gauge theory under consideration is that the gauge covariant equations are readily expressible in gauge invariant form; we can write a closed second-order system of equations for the two gauge invariant quantities $|\phi|$ and f, where f is the Maxwell field tensor. From the solu-

tions for f and $|\phi|$ the gauge potential A can be constructed in a suitable gauge using Maxwell's equations. The gauge invariant system is derived in Sec. IV and the equivalence of the first- and second-order equations demonstrated in Sec. V. The proofs follow the same strategy as in Refs. 2 and 3 but require modification, particularly with the application of the maximum principle. The difficulty in generalizing the proofs is the appearance in the field equations of a term which lies in $L^1(\mathbb{R}^2)$ [see Eq. (2.6)], and for which *a priori* estimates are difficult to obtain. However, first we discuss in Sec. II some properties of the model.

II. THE MODEL

Define the energy functional[4,5]

$$E = \int [\tfrac{1}{4}(F_{ij})^2 + \tfrac{1}{2}F(|\phi|)|D_i\phi|^2 + \tfrac{1}{2}w^2], \tag{2.1}$$

where the integral is understood to be over $\mathbb{R}^2$, $F(|\phi|)$ is nonnegative, and w is defined for each F according to

$$w(|\phi|) = \int_{|\phi|}^{1} sF(s)\,ds. \tag{2.2}$$

The field tensor F_A is given in terms of the gauge potential $A = A_i(x)\,dx^i$ as follows (for notation see Jaffe and Taubes[3]):

$$F_A = dA = \tfrac{1}{2}F_{ij}\,dx^i \wedge dx^j = \tfrac{1}{2}(\nabla_i A_j - \nabla_j A_i)\,dx^i \wedge dx^j, \tag{2.3}$$

and the covariant derivative by

$$D_A\phi = D_i\phi\,dx^i = (\nabla_i\phi - iA_i\phi)\,dx^i, \tag{2.4}$$

where ϕ is a complex valued function on $\mathbb{R}^2$. The Ginzburg–Landau energy functional is recovered by choosing $F \equiv 1$, in which case the potential $\frac{1}{2}w^2$ reduces to the usual ϕ^4 interaction. Notice that we have set the electric field potential A_0, equal to zero. This follows in fact from the requirement of finite energy, $E < \infty$, provided that $F(1) > 0$ (see also Julia and Zee[8]). The particle masses m can be determined heuristically by identifying the coefficients of the quadratic terms in the fields with m^2, and we find $m^2 = F(1)$, where m is the mass of both the gauge photon and the Higgs meson; these masses are equal provided the coupling constant λ in the interaction $\lambda w^2/2$ is equal to 1, as in Eq. (2.1). For $F(1) = 0$, then, the photon and meson are massless; this is verified by the asymptotic estimates of Sec. III (see Proposition 3.7).

The variational equations which follow from (2.1) are

$$df + |\phi|FJ = 0, \tag{2.5}$$

$$*D_A*(FD_A\phi) + wF\phi - \tfrac{1}{2}F'\hat{\phi}\,|D_A\phi|^2 = 0, \tag{2.6}$$

where $|D_A\phi|^2 = *(D_A\phi \wedge *\overline{D_A\phi})$,

$$f = -*F_A = F_{21}, \tag{2.7}$$

$\hat{\phi} = \phi/|\phi|$ and J is the dual of the Noether current:

$$J = *\mathrm{Im}(\hat{\phi}\,\overline{D_A\phi}). \tag{2.8}$$

Equations (2.5) constitute Maxwell's equations, coupled to a complex scalar field ϕ determined by (2.6). Notice that the generalization of (2.1), by including the arbitrary function $F(|\phi|)$, has not changed the form of Maxwell's equations; by putting $\psi = \phi\sqrt{F}$, Eqs. (2.5) take the usual form

$$*df = \mathrm{Im}(\psi\,\overline{D_A\psi}). \tag{2.9}$$

Observe that when $F(1) = 0$, ψ will attain an asymptotic value of zero, and that in this case there is no symmetry breaking if we regard ψ as the fundamental field. However, (2.6) is of a form different to that when $F \equiv 1$, in particular the term $F'\hat{\phi}\,|D_A\phi|^2$ on the right-hand side is new.

The space of continuous gauge potentials with finite energy separates into disjoint sectors labelled by the vortex number n,[3,9] where

$$n = \frac{1}{2\pi}\int f, \tag{2.10}$$

and is an integer. In each such sector the energy is bounded below,

$$E \geqslant 2\pi w(0)|n|. \tag{2.11}$$

This follows from the decomposition, valid for sufficiently smooth fields, following Bogomol'nyi,[4,10]

$$E = \tfrac{1}{2}\int \{(f \pm w)^2 + F|J \pm d|\phi|\,|^2\} \pm 2\pi w(0)n. \tag{2.12}$$

The lower bound is therefore attained if and only if

$$f = w, \quad J = d|\phi| \quad \text{for } n > 0, \tag{2.13a}$$

or

$$f = -w, \quad J = -d|\phi| \quad \text{for } n < 0. \tag{2.13b}$$

These equations can be reduced to a single equation for $|\phi|$, by eliminating the potential A (see Refs. 2–4):

$$\Delta \log|\phi| + w(|\phi|) = 2\pi\sum_{i=1}^{|n|}\delta(x - a^i), \tag{2.14}$$

where the $2n$ parameters (a^i) are the locations of the n vortices in $\mathbb{R}^2$. The gauge fields are constructed from

$$A = -d\alpha + *d(\log|\phi|), \tag{2.15}$$

where $\alpha(x)$ is a gauge parameter. Therefore, from a solution of (2.14), supplemented by the requirement of finite energy, we obtain a solution of Eqs. (2.5) and (2.6).

Let us also make the following observations. Since solutions of (2.14) satisfy

$$E = 2\pi w(0)|n|, \tag{2.16}$$

we must demand that $w(0) < \infty$. This excludes functions F with behavior that is too singular at $|\phi| = 0$, as is evident from (2.2). This includes $F = |\phi|^{-2}$, i.e., $w = -\log|\phi|$, for which (2.14) is linear. Evidently this corresponds to the free field case for theories of the type in Eq. (2.1), in which the kinetic and potential terms are related by the definition (2.2). This is made manifest by defining a new field $u = -\log|\phi|$, and the fields A and u are then seen to be decoupled in a suitable gauge.

Note also that the Hamiltonian (2.1) retains its form under the transformation

$$|\phi| \to |\phi|^{-1}, \tag{2.17}$$

together with the redefinition $|\phi|^{-4}F(|\phi|^{-1}) \to F(|\phi|)$. This provides a way of defining finite-energy vortices in a model with singular behavior at $|\phi| = 0$. For example, $F = |\phi|^{-4}$ violates $w(0) < \infty$ but under (2.17) the Hamiltonian (2.1) is transformed into the Ginzburg–Landau model, with $F \equiv 1$.

III. EXISTENCE AND UNIQUENESS OF VORTEX SOLUTIONS

We have seen that vortex solutions for the models under consideration can always be constructed from solutions of Eq. (2.14). Let

$$u = -\log|\phi|, \quad \beta(u) = w(e^{-u}). \tag{3.1}$$

From the definition (2.3) for w, the condition $F \geqslant 0$, and assuming local integrability for $sF(s)$, β is continuous monotone nondecreasing on $\mathbb{R}$ and hence maximal monotone. Equation (2.4) becomes

$$-\Delta u + \beta(u) = 2\pi \sum_{i=1}^{|n|} \delta(x - a^i). \tag{3.2}$$

This equation is of the form

$$-\Delta u + \beta(u) \quad \ni \quad g, \tag{3.3}$$

which is studied in Refs. 6 and 7, where β is a maximal monotone graph in $\mathbb{R}$. In Ref. 6, $g \in L^1(\mathbb{R}^2)$, and in Ref. 7 results are extended to the case where $g \in \mathcal{M}(\mathbb{R}^2)$, the space of bounded Radon measures in $\mathbb{R}^2$. This latter result is obtained by approximating $g \in \mathcal{M}(\mathbb{R}^2)$ with a sequence $\{g_n\}$ such that $g_n \in C^\infty(\mathbb{R}^2) \cap L^1(\mathbb{R}^2)$ and using the results of Ref. 6. In order to state the existence results, we define first the Marcinkiewicz space $M^p(\mathbb{R}^2)$ and then the exponential orders of growth of β:

Definition 3.1: Let u be a measurable function on $\mathbb{R}^2$, $1 < p < \infty$ and $1/p' + 1/p = 1$. Then $\|u\|_{M^p} = \min\{c \in [0,\infty] \mid \int_\Omega |u(x)| < c(\text{meas}\,\Omega)^{1/p'}$ for all measurable $\Omega \subset \mathbb{R}^2\}$. $M^p(\mathbb{R}^2)$ is the set of measurable functions u on $\mathbb{R}^2$ satisfying $\|u\|_{M^p} < \infty$.

Definition 3.2: The exponential orders of growth of a maximal monotone graph β at infinity are defined as

$$a^+(\beta) = \begin{cases} \sup\left\{a \mid \int_0^\infty \beta(s)e^{-as}\,ds = \infty\right\} & \text{if } \sup D(\beta) = \infty \\ \infty \quad \text{otherwise,} \end{cases}$$

$$a^-(\beta) = \begin{cases} \sup\left\{a \mid -\int_0^\infty \beta(-s)e^{-as}\,ds = \infty\right\} & \text{if } \inf D(\beta) = -\infty \\ \infty \quad \text{otherwise,} \end{cases}$$

where $D(\beta)$ is the domain of β.

It is assumed for (3.3) that

$$0 \in \beta(0) \cap \text{Int}\,\beta(\mathbb{R}). \tag{3.4}$$

Observe that the condition $0 \in \text{Int}\,\beta(\mathbb{R})$ implies $a^\pm \geqslant 0$. Define also the Sobolev spaces $W^{k,p}(\Omega)$, $W^{k,p}_{\text{loc}}(\Omega)$ in the usual way. We need to consider only $g \in \mathcal{M}(\mathbb{R}^2)$ of the form $g = \Sigma_{i=1}^\infty c_i \delta(x - a^i)$, $a^i \in \mathbb{R}^2$, where the $c_i \in \mathbb{R}$ are the point mass coefficients. We can now state:

Theorem 3.3 (Vazquez[7]): Let β have finite exponential orders and let $g \in \mathcal{M}(\mathbb{R}^2)$. There exists a $u \in W^{1,1}_{\text{loc}}(\mathbb{R}^2)$ with $|\nabla u| \in M^2(\mathbb{R}^2)$ and a $w \in L^2(\mathbb{R}^2)$ such that $w \in \beta(u)$ a. e. and $\Delta u = w - g$ if and only if every point mass coefficient of g, c_i, is such that $c^- \leqslant c_i \leqslant c^+$, where the critical values are defined by $c^\pm = \pm 4\pi/a^\pm$. In addition, the solution is unique of $\beta^{-1}(0) = \{0\}$, or $\int g \neq 0$.

This theorem enables us to generalize the results of Ref. 5; we can now include the case $F(1) = 0$ of massless particles and dispense with other assumptions as well. In order to apply the theorem and its further consequences, we note first from (3.1) that

$$\beta(0) = 0. \tag{3.5}$$

We also demand

$$0 < \beta(\infty), \tag{3.6}$$

and, because of finite energy [see (2.16)],

$$\beta(\infty) < \infty. \tag{3.7}$$

A further natural requirement is

$$\beta^{-1}(0) = \{0\}, \tag{3.8}$$

for this is equivalent to demanding that the potential term $w^2/2$ in the expression (2.1) should have a unique minimum, which will lie at $|\phi| = 1$. This ensures that the symmetry breaking, and the asymptotic value of $|\phi|$, are uniquely defined, and excludes functions F which are identically zero in a neighborhood of $|\phi| = 1$. However, solutions still exist and are unique even if (3.8) is violated, and the asymptotic value of $|\phi|$ is then smallest $|\phi|$ for which $w(|\phi|) = 0$.

Since in our application $g \geqslant 0$, it follows that any solution u satisfies $u \geqslant 0$ (Ref. 7, Proposition 2). Together with (3.5) and (3.6) this fact ensures that (3.4) is satisfied. Furthermore, (3.7) implies that the exponential order a^+ takes the value 0. a^- takes a value which depends on F; but, since $a^- \geqslant 0$, $c^- = -4\pi/a^- \leqslant 0$, and we find the conditions $c^- \leqslant c_i \leqslant c^+$ of the theorem always to be satisfied. We conclude therefore that a solution to Eq. (3.2) exists, and is unique.

Remarks 3.4: (i) The unique solution has finite energy. Given $|\phi|$, we construct the gauge potential according to (2.15) and the vortex energy (2.1) is given by [using $f^2 = w^2, |D_A\phi|^2 = 2(\nabla|\phi|)^2$],

$$E = \int (F(\nabla|\phi|)^2 + w^2). \tag{3.9}$$

In order to demonstrate that $E < \infty$, we apply Lemma A.1 of Ref. 7, which extends Lemma A.13 of Ref. 6. Since $\beta(u) \in L^1(\mathbb{R}^2)$ there is a $k > 0$ such that $\text{meas}[u > k] < \infty$. Provided $\beta \in C^1(\mathbb{R})$, at least on $[0,\infty)$, we can choose $p(u) = \beta(u)/\beta(\infty)$; then $p \in C^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ is nondecreasing, and satisfies $|p| \leqslant 1$. The equation

$$\int p'(u)|\nabla u|^2 + \int p(u)\beta(u) \leqslant 2\pi|n| \tag{3.10}$$

from Lemma A.1, Ref. 7, then shows that $E < \infty$. In addition, $\int \Delta u = 0$ shows that $\int w = 2\pi|n|$, i.e., the solution describes n vortices.

(ii) The regularity of the solution depends on the properties of F. Since $|\phi| \leqslant 1$ ($u \geqslant 0$), the regularity of the solution depends only on that of $F(|\phi|)$ on $[0,1]$:

Proposition 3.5[5]: If the first k derivatives of $F(|\phi|)$ are bounded on the interval $[0,1]$, then $|\phi| \in C^{k+1}(\mathbb{R}^2)$.

(iii) If $|\phi| \in C^2(\mathbb{R}^2)$, an application of the strong maximum principle using $|\phi| \leqslant 1$ shows that $|\phi| < 1$ ($u > 0$).

(iv) Define $\beta_+^{-1}(s) = \sup\{t : \beta(t) \ni s\}$. Then we have:

Proposition 3.6 (Vazquez,[7] Lemma 4): Let $g \in \mathscr{M}(\mathbb{R}^2)$ have support in $B_R(0)$, $R>0$ (choose $R > \max_i\{|a^i|\}$). Then u is locally bounded outside $B_R(0)$, and we have the estimate

$$u(x) \leqslant -2|n|\log(1 - R/|x|) + \beta_+^{-1}(2|n|/R(2|x| - R)). \tag{3.11}$$

Thus, if $\beta^{-1}(0) = \{0\}$, u converges uniformly to 0 at infinity.

It is of interest to improve the estimate (3.11), in particular to demonstrate the different behavior of the massive $[F(1) \neq 0]$ and massless $[F(1) = 0]$ models. The former will have an asymptotic dependence $u \sim \exp(-m|x|)$, where m is the mass, while for the latter u will decay more slowly, $u \sim |x|^{-p}$ for some exponent p, as is shown in the following estimates. Let us note that more precise asymptotic estimates, for $\beta(u)$ of the form $\beta(u) = u|u|^{q-1}$ have been given by Veron.[11]

Proposition 3.7: (i) Suppose F is continuous on $[\delta,1]$; F' exists on $[\delta,1]$ for some $0<\delta<1$, and $F(1) \neq 0$. Then for any $\epsilon > 0$ there exists $M < \infty$, $R(\epsilon) > 0$ such that

$$0 < u(x) < M \exp[-|x|(\sqrt{F(1)} - \epsilon)], \quad |x| > R(\epsilon). \tag{3.12}$$

(ii) Suppose $F^{(n-1)}$, $n \geqslant 1$, is continuous on $[\delta,1]$, and $F^{(n)}$ exists on $[\delta,1]$ for some $\delta > 0$, with $F^{(i-1)}(1) = 0$, $i = 1,\ldots,n$, $F^{(n)}(1) \neq 0$. Then there exist $0 < M_1 \leqslant M_2 < \infty$, $R > 0$ such that

$$M_1|x|^{-2/n} \leqslant u(x) \leqslant M_2|x|^{-2/n}, \quad |x| > R. \tag{3.13}$$

Proof: (i) From Proposition 3.6, for sufficiently small $\delta > 0$ there exists $R(\delta) > 0$ such that $0 < u < \delta$, for $|x| > R$. Using Taylor's theorem for $\beta(u)$ on $[0,\delta]$, there exists $\xi \in [0,\delta]$ with

$$\begin{aligned}\beta(u) &= \beta(0) + u\beta'(\xi) \\ &= uF(e^{-\xi})e^{-2\xi} \\ &\geqslant u(F(1) - \epsilon)\end{aligned}$$

by continuity of F. Hence, for $|x| > R$,

$$\Delta u \geqslant u(F(1) - \epsilon). \tag{3.14}$$

Now, since $u \in C^2(\mathbb{R}^2)$ we can apply Proposition 7.2 of Ref. 3 to obtain the result.

(ii) As in (i), apply Taylor's theorem to $\beta(u)$ for $u \in [0,\delta]$:

$$\beta(u) = [u^{n+1}/(n+1)!]\,\beta^{n+1}(\xi), \quad \xi \in [0,\delta]. \tag{3.15}$$

Hence $C_1 u^{n+1} \leqslant \beta(u) \leqslant C_2 u^{n+1}$, for constants $0 < C_1 \leqslant C_2$. Define, for $|x| \geqslant R$, $v = M|x|^{-2/n}$, satisfying

$$\Delta v = (4M^{-n}/n^2)v^{n+1}. \tag{3.16}$$

Now apply the strong maximum principle to $u - v$, to obtain upper and lower bounds on $u(x)$, $|x| > R$. For example, choosing $4M^{-n}/n^2 \leqslant C_1$,

$$\begin{aligned}\Delta(v-u) &\leqslant C_1(v^{n+1} - u^{n+1}) \\ &= C(x)(v-u),\end{aligned} \tag{3.17}$$

where

$$0 \leqslant C(x) = C_1\left(\frac{v^{n+1} - u^{n+1}}{v-u}\right) \in L^\infty(\mathbb{R}^2).$$

Apply the maximum principle to (3.17) on $\{|x| > R\}$, noting that we can choose M sufficiently large to ensure that $v - u|_{|x|=R} \geqslant 0$, to obtain $v - u \geqslant 0$, for $|x| > R$. Similarly we obtain the lower bound.

For the massless case, it is not difficult to find examples which allow explicit solutions. A simple example is the following, in which the polynomial decay for the massless fields is evident:

Example 3.8:

$$F = 8|1 - |\phi|^2|. \tag{3.18}$$

The unique solution to (2.14), for $n = 1$, is

$$|\phi| = |x|/\sqrt{1 + |x|^2}. \tag{3.19}$$

The gauge potential A (in the Coulomb gauge), the field f, and the vortex mass E are readily calculated using formulas such as (2.15) and (2.16), and we find

$$\begin{aligned}A &= -[|x|^2/(1+|x|^2)]\,d\theta, \\ f &= w = 2/(1+|x|^2), \\ E &= 4\pi.\end{aligned} \tag{3.20}$$

IV. SECOND-ORDER EQUATIONS

Following the existence of solutions which achieve the lower energy bound shown in (2.11), a natural question arises as to whether these solutions exhaust all finite-energy solutions. To answer this, we need to return to the second-order equations (2.5) and (2.6). By using maximum principle type arguments, and by modifying the proofs in Ref. 3, we find that, with some assumptions, no new solutions exist. First we simplify Eqs. (2.5) and (2.6), casting them into a gauge invariant form which requires us to solve only two coupled equations, for f and $|\phi|$. The gauge covariance of Eqs. (2.5) and (2.6) implies that there are only three independent equations, for $|\phi|$ and for the two components of A. The equation for $w(|\phi|)$, which follows directly from (2.6), is

$$\Delta w = \rho w - \gamma F^2|\phi|^2|D_A\phi|^2, \tag{4.1}$$

where

$$\begin{aligned}\rho &= F|\phi|^2, \\ \gamma &= \frac{(F|\phi|^2)'}{2F^2|\phi|^3} = \frac{F'}{2F^2|\phi|} + \frac{1}{F|\phi|^2}.\end{aligned} \tag{4.2}$$

From Eqs. (2.5), which are second order in the potential A, we can derive a second-order equation for f by differentiation. We find [using the definition (2.8) for J]

$$\Delta f = \rho f - \gamma F^2|\phi|^2 i^*(D_A\phi \wedge \overline{D_A\phi}). \tag{4.3}$$

By squaring (2.5) and using

$$|J|^2 = |D_A\phi|^2 - (\nabla|\phi|)^2, \tag{4.4}$$

we find

$$|\nabla f|^2 = F^2|\phi|^2|D_A\phi|^2 - (\nabla w)^2. \tag{4.5}$$

Again, using the definition of J,

$$(J, d|\phi|) = \tfrac{1}{2} i^*(D_A\phi \wedge \overline{D_A\phi}), \tag{4.6}$$

and we obtain the following gauge invariant system, involving only the unknowns f and $|\phi|$:

$$\begin{aligned}&\Delta f - \rho f + 2\gamma\nabla f\cdot\nabla w = 0, \\ &\Delta w - \rho w + \gamma[(\nabla f)^2 + (\nabla w)^2] = 0.\end{aligned} \tag{4.7}$$

The boundary conditions for (4.7) are determined by the fin-

ite-energy requirements, which can be written as follows, again using (4.5):

$$\int f^2 < \infty, \quad \int w^2 < \infty,$$
$$\int \frac{(\nabla f)^2}{F|\phi|^2} < \infty, \quad \int \frac{(\nabla w)^2}{F|\phi|^2} < \infty. \tag{4.8}$$

The system (4.7), (4.8) forms a closed elliptic system for f and $|\phi|$, and our aim is to find all solutions of this system. Evidently, solutions can always be obtained by putting $f = \pm w$, with w determined by (2.14). With the solutions of (4.7),(4.8) we can construct the gauge fields using Maxwell's equations (2.5). In order to see this, put

$$\phi = |\phi| e^{i\alpha}, \tag{4.9}$$

where $\alpha(x)$ is a gauge parameter, necessarily multivalued for nontrivial solutions.[3] Equation (2.5) can be written

$$A = -d\alpha - *df/F|\phi|^2. \tag{4.10}$$

Therefore, given f and $|\phi|$ as determined by (4.7),(4.8), we need only to choose a gauge to be able to write down the solution for A. If we can determine that all solutions satisfy $f = \pm w$, we recover Eqs. (2.15); that is, $f = \pm w$ together with Maxwell's equations imply the remaining first-order equations $J = \pm d|\phi|$, which appear in Eqs. (2.13).

Using (4.10), the equation for f can be cast into a useful divergence form:

$$\nabla(\nabla f/F|\phi|^2) = f - g, \tag{4.11}$$

where $g(x) = [\nabla_1, \nabla_2]\alpha(x)$ is singular, being nonzero only at points where $|\phi| = 0$. This is evident from Eqs. (4.9) and (4.10) since, in order that (A,ϕ) be sufficiently smooth, the zeros of $|\phi|$ should coincide with the points where α is discontinuous. In the next section (Proposition 5.2) we demonstrate, following Ref. 2, that we can always choose a gauge in which A is smooth, provided F is sufficiently smooth and assuming local regularity properties of (A,ϕ). It is worth remarking that Eqs. (4.7) and (4.11) for f and $|\phi|$ can be obtained as the Euler equations of the following functional $\mathscr{A}(f,|\phi|)$:

$$\mathscr{A}(f,|\phi|) = \int \left[\frac{(\nabla f)^2 - (\nabla w)^2}{F|\phi|^2} + f^2 - w^2 - 2fg \right]. \tag{4.12}$$

Next we describe a virial theorem, following Ref. 3. Define the Maxwell stress tensor

$$T_{ij} = \{\nabla_i w \nabla_j w - \nabla_i f \nabla_j f + \tfrac{1}{2}\delta_{ij}[(\nabla f)^2 + (\Delta w)^2]\}/F|\phi|^2 + \tfrac{1}{2}\delta_{ij}(f^2 - w^2). \tag{4.13}$$

It follows from (4.7) that

$$\nabla_j T_{ij} = 0, \tag{4.14}$$

and from (4.8) that

$$\int |T_{ij}| < \infty. \tag{4.15}$$

Proposition 4.1: Let (f,w) be a solution to Eqs. (4.7),(4.8). Then the stress tensor (4.13) satisfies

$$\int T_{ij} = 0. \tag{4.16}$$

Proof: See Jaffee and Taubes,[3] p. 31.

As a consequence, we have the following useful relation:

$$\int f^2 = \int w^2. \tag{4.17}$$

V. EQUIVALENCE OF FIRST- AND SECOND-ORDER EQUATIONS

We now require several assumptions on the behavior of F, and also assume local regularity of (A,ϕ). We show then that $|\phi|$ is bounded, and, following Taubes,[2] show that, with a suitable choice of gauge, (A,ϕ) is smooth. This will imply that f and w are continuous, and from (4.7),(4.8) we then show that $w \geqslant |f|$; combined with (4.17) this implies $f = w$ or $f = -w$ and, as explained above, this is sufficient to demonstrate the equivalence of the first- and second-order equations. The assumptions on F are

(i) $F > 0$, (5.1)

(ii) there exists a constant $K \geqslant 1$ such that for all $s > K$,

$$F(s) + \tfrac{1}{2}sF'(s) \geqslant 0, \tag{5.2}$$

(iii) $F \in C^1[0,\infty)$. (5.3)

The first condition is used to obtain a lower bound on F, although it excludes the massless case. The second condition is used solely to show that $\|\phi\|_\infty \leqslant K$; it means that $F(s)s^2$ is a nondecreasing function of s, for $s > K$, and is satisfied by any positive polynomial F and by any function F which increases for $s > K$. Using (5.3), $|\phi| \leqslant K$ implies that $F(|\phi|)$ is bounded above and below:

$$0 < k_1 \leqslant F(|\phi|) \leqslant k_2, \tag{5.4}$$

for finite constants k_1 and k_2. Similarly, because F' is continuous,

$$|F'(|\phi|)| \leqslant k_3 < \infty. \tag{5.5}$$

The third condition (5.3) also ensures that the solutions f, $w \in C^2(\mathbb{R}^2)$, and so in fact are classical solutions (see Proposition 3.5).

We also assume that the components of A belong to $W^{1,2}_{\text{loc}}(\mathbb{R}^2)$, and that $|\phi| \in W^{2,2}_{\text{loc}}(\mathbb{R}^2)$. This last assumption is stronger than that used by Taubes[2] and has been necessary, in order to ensure that f and w are sufficiently smooth, because of the difficulty posed by the extra L^1 term in the field equations (2.6). This assumption implies that $|\phi|$ is continuous.

Proposition 5.1: With the above assumptions, $|\phi| \leqslant K$.

Proof: Let

$$v = \int_{|\phi|}^{1} ds\, F(s). \tag{5.6}$$

v satisfies the distributional equation

$$\Delta v = |\phi| F w - (F/|\phi| + \tfrac{1}{2}F')|D_A \phi|^2 + (F/|\phi|)(\nabla|\phi|)^2. \tag{5.7}$$

Define $b_R(x) = b(|x|/R)$, where $0 \leqslant b(|x|) \leqslant 1$ is a C_0^∞ monotonically decreasing function with

$$b(|x|) = \begin{cases} 1, & |x| \leqslant 1, \\ 0, & |x| \geqslant 2. \end{cases} \tag{5.8}$$

Define $\eta \in W_0^{1,2}(B_{2R}(0))$ by

$$\eta = b_R \max(0, |\phi| - K). \tag{5.9}$$

Equation (5.7) implies

$$\int_{\Omega_{2R}} [\nabla\eta\cdot\nabla v + F\cdot|\phi|w\eta - (\eta/|\phi|)(F + \tfrac{1}{2}|\phi|F')|D_A\phi|^2 + (\eta F/|\phi|)(\nabla|\phi|)^2] = 0,$$

where $\Omega_{2R} = \{ x \in \mathbb{R}^2 | \, |\phi|(x) > K \} \cap B_{2R}(0)$. Observe that all terms are finite, due to the local regularity assumptions and finite energy. Using definitions (5.6) and (5.9) and collecting terms,

$$\int_{\Omega_{2R}} b_R \{ [(|\phi| - K)/|\phi|](F + \tfrac{1}{2}|\phi|F')|D_A\phi|^2 + (KF/|\phi|)\cdot(\nabla|\phi|)^2 - F|\phi|w\cdot(|\phi| - K)\} = -\int_{\Omega_{2R}} [F\cdot(|\phi| - K)\nabla|\phi|\cdot\nabla b_R]. \tag{5.10}$$

Let

$$G(|\phi|) = \int_{|\phi|}^{1} F(s)(s - K)\, ds. \tag{5.11}$$

For $|\phi| > K \geqslant 1$,

$$|G| \leqslant \int_1^{|\phi|} F(s)(s + K)\, ds \leqslant (K + 1)\int_1^{|\phi|} F(s)s\, ds = (K + 1)|w|. \tag{5.12}$$

The integral of the left-hand side of Eq. (5.10) is nonnegative [using (5.2)], and with the definition (5.11) we obtain

$$\int_{\Omega_R} \{ [(|\phi| - K)/|\phi|](F + \tfrac{1}{2}|\phi|F')|D_A\phi|^2 + (KF/|\phi|)\cdot(\nabla|\phi|)^2 - Fw\cdot|\phi|(|\phi| - K)\} \leqslant \int_{\Omega_{2R}} \nabla b_R\cdot\nabla G \leqslant \left[\int_{\Omega_{2R}} G^2\right]^{1/2} \|\Delta b_R\|_{L^2} \leqslant [(K + 1)^2/R]\|\Delta b\|_{L^2}\|w\|_{L^2}, \tag{5.13}$$

where we have integrated by parts, used Hölder's inequality, the estimate (5.12), and the scaling properties of b_R. Since $\Omega_R \subseteq \Omega_{R'}$ for $R' \geqslant R$ we conclude that Ω_∞ has zero measure and hence $\|\phi\|_\infty \leqslant K$.

Next, with the above assumptions, we prove (following Taubes[2]) that it is always possible to choose a gauge in which the potential A is smooth.

Proposition 5.2 (Taubes[2]): Let (A,ϕ) be a weak solution of Eqs. (2.5) and (2.6). Then there exists a pair $(\tilde{A},\tilde{\phi})$ related to (A,ϕ) by $(\tilde{A},\tilde{\phi}) = A - d\alpha, \phi e^{i\alpha})$, where the components of $\tilde{A} \in C^1(\mathbb{R}^2)$, $\tilde{\phi} \in C^0(\mathbb{R}^2)$ and $\alpha \in W^{2,2}(\Omega)$ for all open sets $\Omega \subset \mathbb{R}^2$ with compact closure.

Proof: We need only outline the proof, which is to be found in Ref. 2. By a weak solution A of Eqs. (2.5) we mean a potential A with locally integrable components, and locally integrable first derivatives, satisfying

$$\int [db \wedge *F_A + b \wedge *\mathrm{Im}(F\phi\ \overline{D_A\phi})] = 0, \tag{5.14}$$

where b has components in $W^{1,2}(\mathbb{R}^2)$ and $|\phi| \in W^{2,2}_{\mathrm{loc}}(\mathbb{R}^2)$. We determine the gauge parameter $\alpha(x)$ which transforms A into the Coulomb gauge, in $B = B_2(0)$; that is, we choose $\alpha \in W^{2,2}(B)$ as the unique solution of

$$\Delta\alpha = *d*A, \quad \alpha|_{\partial B} = 0. \tag{5.15}$$

Then, using $|\phi| \leqslant K$, the standard regularity estimates (Morrey,[12] Chap. 6) and the Sobolev imbedding theorem,[13] we find that $\tilde{A} = A - d\alpha$ is continuous in B. Since we have assumed that $|\phi| \in W^{2,2}(B)$ we can iterate, using $F \in C^1[0,\infty)$, to obtain that A and its first derivatives are continuous in B. This means that $f = -*dA$ is continuous in B. Further iterations, using also Eq. (2.6) for ϕ, are possible if extra smoothness is assumed for F. Since the origin was chosen arbitrarily, we find that f, and by assumption $|\phi|$, are continuous in $\mathbb{R}^2$. By a patching procedure we can also construct α such that $\alpha \in W^{2,2}(\Omega)$ for any bounded set $\Omega \subset \mathbb{R}^2$.

Let us now return to the gauge invariant formulation of the second-order equations (4.7). By adding and subtracting these equations, we obtain

$$\Delta u - \rho u + \gamma(\nabla u)^2 = 0, \tag{5.16}$$

which holds for each of $u = w + f$, $u = w - f$. Using $|\phi| \leqslant K$ we find that $F|\phi|^2$ is bounded above and hence, from (4.8), $\|\nabla F\|_{L^2} < \infty$, $\|\nabla w\|_{L^2} < \infty$. This implies that f, $w \in W^{1,2}(\mathbb{R}^2)$, i.e., $u \in W^{1,2}(\mathbb{R}^2)$. A consequence of this and (5.16) is that $u \geqslant 0$. This is straightforward to prove if F is such that $\gamma \geqslant 0$, by application of the maximum principle,[14] as in Refs. 1 and 2. For more general γ we note:

Lemma 5.3: With the above assumptions on F, $\gamma(|\phi|)$ is bounded below.

Proof: From (4.2), for any $\epsilon > 0$,

$$\gamma \geqslant [(F')^2/16F^4|\phi|^2][16F^3/(F')^2 - \epsilon] - 1/\epsilon.$$

Now,

$$16F^3/(F')^2 \geqslant k > 0,$$

for some positive constant k, since by (5.4) and (5.5) $|F'|$ is bounded above, and $F \geqslant k_1$ for some $k_1 > 0$. Hence, by choosing ϵ sufficiently small,

$$\gamma \geqslant -c, \tag{5.17}$$

for some $c > 0$. ■

Lemma 5.4: The function $(e^v - 1)$ for $v \in W^{1,2}(\mathbb{R}^2)$ is square-integrable on $L^2(\mathbb{R}^2)$.

Proof: See Taubes,[1] Lemma 4.6.

Using Lemma 5.3, we obtain

$$\Delta u - c(\nabla u)^2 - \rho u \leqslant 0. \tag{5.18}$$

Proposition 5.5: For $u \in W^{1,2}(\mathbb{R}^2) \cap C^0(\mathbb{R})$, $c > 0$, $\rho(x) \geqslant 0$, and bounded, (5.18) implies that $u \geqslant 0$.

Proof: Define the test function $v \in W_0^{1,2}(B_R(0))$ by

$$v = \begin{cases} (e^{-cu} - 1)b_R & \text{for } u < 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.19}$$

where b_R is the cutoff function defined above [see Eq. (5.8)]. Since v is compactly supported and $v \geqslant 0$, we can multiply (5.18) by v and integrate by parts:

$$-\int \nabla v\cdot\nabla u - c\int v|\nabla u|^2 - \int \rho uv \leqslant 0. \tag{5.20}$$

Using (5.19) and collecting terms,

$$c\int_{\Omega_R}(\nabla u)^2-\int_{\Omega_R}\rho u(e^{-cu}-1)\leqslant\int_{\Omega_R}(e^{-cu}-1)\nabla u\cdot\nabla b_R, \tag{5.21}$$

where $\Omega_R=\{x\in\mathbb{R}^2|u(x)<0\}\cap B_R(0)$. A bound for the right-hand side of (5.21), using Hölder's inequality, is

$$\left|\int_{\Omega_R}(e^{-cu}-1)\nabla u\cdot\nabla b_R\right|\leqslant(\|\nabla b\|_\infty/R)\|\nabla u\|_{L^2}\|e^{-cu}-1\|_{L^2}. \tag{5.22}$$

Since $u\in W^{1,2}(\mathbb{R}^2)$, we have $\|\nabla u\|_{L^2}<\infty$, $\|e^{-cu}-1\|_{L^2}<\infty$ by Lemma 5.4. Taking lim inf $R\to\infty$, we find that Ω_∞ has zero measure, i.e., $u\geqslant 0$.

Since u can be either $w+f$ or $w-f$, we find $w\geqslant|f|\geqslant 0$. Equation (4.17) implies, using continuity, $f^2=w^2$, or $f(x)=\pm w(x)$. Substituting into Eq. (4.11), we find

$$\Delta\log|\phi|+w=0,\quad|\phi|\neq 0. \tag{5.23}$$

Lemma 5.6: Either $w\equiv 0$ or $w>0$.

Proof: Since we have assumed $F\in C^1[0,\infty)$, $|\phi|\in C^2(\mathbb{R}^2)$ (see Ref. 5, Proposition 3.5); also $w\geqslant|f|$ implies $|\phi|\leqslant 1$. Now apply the strong maximum principle to (5.23) on the set $\{x||\phi|(x)>0\}$ to complete the proof (for details, see Ref. 5, Lemma 5.2).

Finally, using Lemma 5.6 and the continuity properties of f and w as in Ref. 3, we deduce that $f(x)=\pm w(x)$ holds with the same sign everywhere, this sign depending on the sign of n by (2.10):

$$f=w\quad\text{if } n>0,\qquad f=-w\quad\text{if } n<0. \tag{5.24}$$

As explained in Sec. IV, Eqs. (5.24) imply the first-order relations (2.15), which together with an analysis of the zeros of $|\phi|$ (see Refs. 3, Chap. III) imply Eq. (3.2), which was investigated in Sec. III.

ACKNOWLEDGMENTS

We wish to thank Professor H. Brezis for bringing the work of Vazquez[7] to our notice and Professor L. C. Evans for helpful comments concerning Sec. V.

[1] C. H. Taubes, Commun. Math. Phys. **72**, 277 (1980).
[2] C. H. Taubes, Commun. Math. Phys. **75**, 207 (1980).
[3] A. Jaffe and C. Taubes, *Vortices and Monopoles* (Birkhauser, Boston, 1980).
[4] M. A. Lohe, Phys. Rev. D **23**, 2335 (1981).
[5] M. A. Lohe and John van der Hoek, "Existence and uniqueness of generalized vortices," J. Math. Phys. **24**, 148 (1983).
[6] Ph. Benilan, H. Brezis, and M. Crandall, Ann. Scuola Norm. Sup. Pisa II **4**, 523 (1975).
[7] J. L. Vazquez, "On a Semilinear Equation in $\mathbb{R}^2$ Involving Bounded Measures," preprint, Universidad Complutense, Madrid, 1981.
[8] B. Julia and A. Zee, Phys. Rev. D **11**, 2227 (1975).
[9] E. Weinberg, Phys. Rev. D **19**, 3008 (1979).
[10] E. B. Bogomol'nyi, Yad. Fiz. **24**, 861 (1976) [Sov. J. Nucl. Phys. **24**, 449 (1976)].
[11] L. Veron, "Asymptotic behaviour of the solutions of some nonlinear elliptic equations," in *Nonlinear Problems of Analysis in Geometry and Mechanics*," edited by M. Atteia, D. Bancel and I. Gumowski (Pitman, Boston, 1981).
[12] C. Morrey, *Multiple Integrals in the Calculus of Variations* (Springer-Verlag, Berlin, Heidelberg, New York, 1966).
[13] R. Adams, *Sobolev Spaces* (Academic, New York, 1975).
[14] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order* (Springer-Verlag, Berlin, Heidelberg, New York, 1977).

A gravitational Poincaré gauge theory and Higgs mechanism

R. J. McKellar
Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

(Received 14 June 1983; accepted for publication 11 August 1983)

In this paper we shall construct the Lagrangian of a gravitational Poincaré gauge theory using degeneracy in the Euler–Lagrange expressions as a primary restriction. Such a generalization of a Lorentz gauge theory requires the addition of not only a translation gauge connection, but also a Goldstone field. The intractability of the field equations is lessened somewhat by means of a particular choice of gauge which acts like a Higgs mechanism. With one further assumption a complete reduction to the corresponding Lorentz theory can be made, and the Einstein vacuum field equations with cosmological term are recovered.

PACS numbers: 11.30.Cp, 11.15.Kc, 12.25. + e, 04.20.Fy

1. INTRODUCTION

Several authors[1] have sought to show how their gravitational field equations can be characterized as those of a unique Poincaré theory. In most instances the Poincaré transformations involved are actually coordinate transformations with parameters from the Poincaré group and not true internal gauge transformations.

It was shown in an earlier paper[2] how the Einstein vacuum field equations with cosmological term could be derived as a consequence of the Euler–Lagrange equations of a Lorentz gauge theory which is in some sense unique. Since the Lorentz group is a subgroup of the Poincaré group, we could also say we have a Poincaré gauge theory. Nonetheless, the absence of any reference to the translation subgroup in the determined Lagrangian should stop us from using this terminology. The aim of this paper is to construct a true Poincaré gauge theory by generalizing the Lorentz theory.

We shall make use of the formalism developed in two previous papers.[2,3] Thus, a Poincaré gauge transformation is characterized by associating at each point of the space-time manifold M (local coordinates x^i, $i = 1,...,4$) an element $u = u(x^i)$ of the connected component of the identity of the Poincaré group. The coordinates of $u(x^i)$ relative to a canonical chart of the first kind[4] are $u^{\alpha\beta}(x^i) = -u^{\beta\alpha}(x^i)$ and $u^{\alpha}(x^i)$, $\alpha,\beta = 1,...,4$.

To generalize the Lorentz gauge theory to a true Poincaré gauge theory, we shall introduce not only a translation gauge connection $A_i^{\ \alpha}$, but also what turns out to be a Goldstone field[5] Φ^{α}. As was shown in Ref. 2, the inclusion of $A_i^{\ \alpha}$ in the formulation of the variational principle without Φ^{α} is futile since the invariance identities eliminate $A_i^{\ \alpha}$ when the Lagrangian is actually constructed. The insertion of Φ^{α} leads to only one additional term to the Lorentz Lagrangian, viz.,

$$de^{ijkh}\eta_{\alpha\beta}f_i^{\ \alpha}{}_j f_k^{\ \beta}{}_h,$$

where d is an arbitrary constant, ϵ^{ijkh} is the four-dimensional Levi-Civita symbol,

$$\eta_{\alpha\beta} \equiv \mathrm{diag}(-1,-1,-1,1)$$

and $f_i^{\ \alpha}{}_j$ is defined in terms of the Poincaré gauge curvatures[2] $F_i^{\ \alpha\beta}{}_j$ and $F_i^{\ \alpha}{}_j$ as

$$f_i^{\ \alpha}{}_j \equiv F_i^{\ \alpha\beta}{}_j\,\eta_{\beta\gamma}\Phi^{\gamma} + F_i^{\ \alpha}{}_j.$$

A simplification of the resulting field equations is obtained by means of a particular choice of gauge which acts like a Higgs mechanism.[5] In this gauge Φ^{α} vanishes and $A_i^{\ \alpha}$ is no longer regarded as a translation gauge connection but as a set of vectors.

To check the validity of the theory, we find that we can reduce it to the Lorentz theory by imposing

$$\Phi^{\alpha}{}_{\|i} = \kappa h_i^{\ \alpha},$$

where a double bar signifies the double covariant derivative,[2,3,6] κ is an arbitrary constant, and the $h_i^{\ \alpha}$ are the components of the orthonormal tetrad (or vierbein).

2. PRELIMINARIES

With a true gauge theory the gauge potential should be a connection in a principal fiber bundle.[7] In particular, the group acts freely on the fiber, i.e., only the action of the identity leaves each element of the fiber invariant. Thus we violate this condition when the action of the Poincaré group is restricted to being

$$h_i^{\ \beta} = a_{\alpha}^{\beta} h'^{\alpha}_{\ i}, \tag{2.1a}$$

where a_{α}^{β} is a Lorentz matrix and a prime indicates the gauge-transformed quantity.

We need to introduce an additional object in the manner of Pilch[8] whose components Φ^{α} undergo the Poincaré gauge transformation

$$\Phi^{\beta} = a_{\alpha}^{\beta}\,\Phi'^{\alpha} + a^{\beta}, \tag{2.1b}$$

where a^{β} characterizes a translation. A coordinate transformation leaves Φ^{α} invariant. When a canonical chart of the first kind is used, the gauge transformation laws (2.1) can be expressed[2] as

$$h'^{\alpha}_{\ i} = \mathscr{L}_{\beta}^{\alpha}\,h_i^{\beta}$$
and
$$\Phi'^{\alpha} = \mathscr{L}_{\beta}^{\alpha}\,\Phi^{\beta} - \mathscr{L}_{\beta}^{\alpha} l_{\gamma}^{\beta}\,u^{\gamma}, \tag{2.2}$$

where

$$\mathscr{L}_{\beta}^{\alpha} \equiv \exp(-u^{\alpha\gamma}\eta_{\gamma\beta})$$

and

$$l_{\beta}^{\alpha} \equiv \delta_{\beta}^{\alpha} + (1/2!)u^{\alpha\gamma}\eta_{\gamma\beta} + (1/3!)u^{\alpha\gamma}\eta_{\gamma\omega}u^{\omega\nu}\eta_{\nu\beta} + \cdots.$$

 0022-2488/84/010161-06$02.50

In addition to Φ^{α}, we shall also make use of the object with components

$$\Phi^{i} \equiv h^{i}_{\alpha}\, \Phi^{\alpha},$$

which enables us to put the transformation laws (2.2) into the form

$$\begin{bmatrix} h^{i}_{\alpha} \\ \Phi^{i} \end{bmatrix}' = \begin{bmatrix} \delta^{i}_{j} \hat{\mathscr{L}}^{\beta}_{\alpha} & 0 \\ -\delta^{i}_{j}\, l^{\beta}_{\gamma} u^{\gamma} & \delta^{i}_{j} \end{bmatrix} \begin{bmatrix} h^{j}_{\beta} \\ \Phi^{j} \end{bmatrix}, \tag{2.3}$$

where h^{i}_{α} is the inverse of h^{α}_{i} and $\hat{}$ denotes the inverse. The purpose of this is to take advantage of the formalism introduced in a previous paper[3] where we now make the identification

$$\rho^{A} = \begin{bmatrix} h^{i}_{\alpha} \\ \Phi^{i} \end{bmatrix}.$$

Under a coordinate transformation $\bar{x}^{i} = \bar{x}^{i}(x^{j})$ with

$$J^{i}_{j} \equiv \frac{\partial x^{i}}{\partial \bar{x}^{j}}$$

and

$$J \equiv \det J^{i}_{j} > 0,$$

we have

$$\begin{bmatrix} \overline{h^{i}_{\alpha}} \\ \Phi^{i} \end{bmatrix} = \begin{bmatrix} \hat{J}^{i}_{j}\, \delta^{\beta}_{\alpha} & 0 \\ 0 & \hat{J}^{i}_{j} \end{bmatrix} \begin{bmatrix} h^{j}_{\beta} \\ \Phi^{j} \end{bmatrix},$$

where we have used a horizontal bar to denote the corresponding quantity in the new coordinate system.

Since ρ^{A} transforms linearly and homogeneously under both Poincaré and coordinate transformations, it is possible to take its double covariant derivative[2,3,6] and obtain

$$h^{i}_{\alpha \| a} = h^{i}_{\alpha,a} + \{_{j}{}^{i}{}_{a}\} h^{j}_{\alpha} - A^{\beta\gamma}_{a} h^{i}_{\beta}\, \eta_{\gamma\alpha}$$

and

$$\Phi^{i}_{\|a} = \Phi^{i}{}_{,a} + \{_{j}{}^{i}{}_{a}\} \Phi^{j} + A^{\beta}_{a} h^{i}_{\beta},$$

where $\{_{j}{}^{i}{}_{a}\}$ is the Christoffel symbol of the second kind and $A^{\beta\gamma}_{a}$ is the Lorentz gauge connection. The corresponding commutation laws[3] for the second derivatives are then

$$h^{i}_{\alpha \| ab} - h^{i}_{\alpha \| ba} = R_{j}{}^{i}{}_{ab}\, h^{j}_{\alpha} - F_{a}{}^{\beta\gamma}{}_{b} h^{i}_{\beta} \eta_{\gamma\alpha}$$

and

$$\Phi^{i}{}_{\|ab} - \Phi^{i}{}_{\|ba} = R_{j}{}^{i}{}_{ab}\, \Phi^{j} + F_{a}{}^{\beta}{}_{b} h^{i}_{\beta},$$

where $R_{j}{}^{i}{}_{ab}$ is the Riemann curvature tensor. It is also possible to show that

$$\Phi^{\gamma}{}_{\|a} = \Phi^{\gamma}{}_{,a} + A^{\gamma}_{a} \eta_{\beta\omega}\, \Phi^{\omega} + A^{\gamma}_{a}$$

and

$$\Phi^{\gamma}{}_{\|ab} - \Phi^{\gamma}{}_{\|ba} = f_{a}{}^{\gamma}{}_{b} \equiv F_{a}{}^{\gamma\beta}{}_{b}\, \eta_{\beta\omega}\, \Phi^{\omega} + F_{a}{}^{\gamma}{}_{b}. \tag{2.4}$$

Note that the gauge transformation law for $\Phi^{\gamma}{}_{\|a}$ is the same as for h^{γ}_{a}, i.e.,

$$\Phi'^{\gamma}{}_{\|a} = \mathscr{L}^{\gamma}_{\omega} \Phi^{\omega}{}_{\|a},$$

and we also have

$$f'_{a}{}^{\gamma}{}_{b} = \mathscr{L}^{\gamma}_{\omega} f_{a}{}^{\omega}{}_{b}.$$

3. DEGENERACY

In Ref. 2 it was found that

$$\begin{aligned} L = {} & a_1 \epsilon^{ijkh} \epsilon_{\alpha\beta\gamma\omega}\, F_{i}{}^{\alpha\beta}{}_{j}\, F_{k}{}^{\gamma\omega}{}_{h} + a_2 \epsilon^{ijkh} \eta_{\alpha\gamma}\, \eta_{\beta\omega}\, F_{i}{}^{\alpha\beta}{}_{j}\, F_{k}{}^{\gamma\omega}{}_{h} \\ & + b_1 h h^{i}_{\mu} h^{j}_{\nu} \eta^{\mu\gamma} \eta^{\nu\omega} \epsilon_{\alpha\beta\gamma\omega}\, F_{i}{}^{\alpha\beta}{}_{j} \\ & + b_2 h h^{i}_{\alpha} h^{j}_{\beta}\, F_{i}{}^{\alpha\beta}{}_{j} + ch, \end{aligned}$$

where a_1, a_2, b_1, b_2, and c are arbitrary constants, $\epsilon_{\alpha\beta\gamma\omega}$ is a four-dimensional Levi-Civita symbol, and

$$h \equiv \det h^{\alpha}_{i},$$

is the most general Lagrangian of the form

$$L = L(h^{\alpha}_{i}; A^{\alpha\beta}_{i}; A^{\alpha\beta}_{i,j}; A^{\alpha}_{i}; A^{\alpha}_{i,j}),$$

which has the transformation laws

$$\bar{L} = JL$$

and

$$L' = L, \tag{3.1}$$

and is degenerate in the sense that its Euler–Lagrange expressions

$$E^{k}_{\sigma\tau} \equiv \frac{\partial L}{\partial A^{\sigma\tau}_{k}} - \frac{\partial}{\partial x^{h}} \left(\frac{\partial L}{\partial A^{\sigma\tau}_{k,h}} \right)$$

and

$$E^{k}_{\sigma} \equiv \frac{\partial L}{\partial A^{\sigma}_{k}} - \frac{\partial}{\partial x^{h}} \left(\frac{\partial L}{\partial A^{\sigma}_{k,h}} \right)$$

are such that

$$\begin{aligned} & \frac{\partial E^{k}_{\sigma\tau}}{\partial A^{\alpha\beta}_{i,jh}} \equiv 0, \quad \frac{\partial E^{k}_{\sigma\tau}}{\partial A^{\alpha}_{i,jh}} \equiv 0, \\ & \frac{\partial E^{k}_{\sigma}}{\partial A^{\alpha\beta}_{i,jh}} \equiv 0, \quad \text{and} \quad \frac{\partial E^{k}_{\sigma}}{\partial A^{\alpha}_{i,jh}} \equiv 0. \end{aligned} \tag{3.2}$$

We shall now generalize this result to a Lagrangian which includes Φ^{i}, i.e.,

$$L = L(h^{\alpha}_{i}; \Phi^{i}; A^{\alpha\beta}_{i}; A^{\alpha\beta}_{i,j}; A^{\alpha}_{i}; A^{\alpha}_{i,j})$$

and demand the same transformation laws (3.1) and degeneracy (3.2). The construction of the Lagrangian follows closely that of Ref. 2 to which the reader should refer constantly. Also, several lemmas were proved in Ref. 2 which are required here and are listed in the Appendix.

To simplify our calculations, we shall use upper case Greek letters to represent all ten gauge indices, so that, for example, A^{Σ}_{i}, $\Sigma = 1,\ldots,10$, signifies the ordered pair $(A^{\alpha\beta}_{i}, A^{\alpha}_{i})$. The degeneracy condition (3.2) can then be expressed as

$$\frac{\partial E^{k}_{\Sigma}}{\partial A^{\Lambda}_{i,jh}} \equiv 0.$$

As in Ref. 2, this condition, together with the invariance identity corresponding to (4.5) in Ref. 3, implies that $\partial^{2} L / \partial A^{\Lambda}_{i,j}\, \partial A^{\Sigma}_{k,h}$ is totally antisymmetric in its Latin indices. Thus,

$$\frac{\partial^{2} L}{\partial A^{\Lambda}_{i,j}\, \partial A^{\Sigma}_{k,h}} = \epsilon^{ijkh} L_{\Lambda\Sigma}(h^{\mu}_{a}; \Phi^{a}), \tag{3.3}$$

where we have made use of the transformation laws of $L_{\Lambda\Sigma}$ inherited from $\partial^2 L/\partial A^{\Lambda}_{i,j}\,\partial A^{\Sigma}_{k,h}$ and the invariance identity corresponding to (4.6) in Ref. 3. Upon integrating (3.3) twice with respect to $A^{\Lambda}_{i,j}$ while noting the appropriate invariance identities we obtain

$$L = \tfrac{1}{8}\epsilon^{ijkh}L_{\Lambda\Sigma}F_k{}^{\Sigma}{}_h F_i{}^{\Lambda}{}_j + \tfrac{1}{2}L^{ij}_{\Lambda}(h^{\mu}_a;\Phi^a)F_i{}^{\Lambda}{}_j + L_0(h^{\mu}_a;\Phi^a),$$

where L^{ij}_{Λ} and L_0 transform in the same way as $\partial L/\partial A^{\Lambda}_{i,j}$ and L, respectively. When we return to lower case Greek indices, we can express the above as

$$\begin{aligned}L = {}&\epsilon^{ijkh}\mathscr{L}_{a\beta\gamma\omega}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma\omega}{}_h\\ &+\epsilon^{ijkh}\mathscr{L}_{\alpha\beta\gamma}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma}{}_h\\ &+\epsilon^{ijkh}L_{\alpha\beta}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha}{}_j F_k{}^{\beta}{}_h\\ &+\mathscr{L}^{ij}_{\alpha\beta}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j + L^{ij}_{\alpha}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha}{}_j + L_0(h^{\mu}_a;\Phi^a).\end{aligned}$$

It is actually more convenient to express L in terms of $f_i{}^{\alpha}{}_j$ rather than $F_i{}^{\alpha}{}_j$, whereby the Lagrangian becomes

$$\begin{aligned}L = {}&\epsilon^{ijkh}L_{a\beta\gamma\omega}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma\omega}{}_h\\ &+\epsilon^{ijkh}L_{\alpha\beta\gamma}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j f_k{}^{\gamma}{}_h\\ &+\epsilon^{ijkh}L_{\alpha\beta}(h^{\mu}_a;\Phi_a)f_i{}^{\alpha}{}_j f_k{}^{\beta}{}_h\\ &+L^{ij}_{\alpha\beta}(h^{\mu}_a;\Phi^a)F_i{}^{\alpha\beta}{}_j + L^{ij}_{\alpha}(h^{\mu}_a;\Phi^a)f_i{}^{\alpha}{}_j + L_0(h^{\mu}_a;\Phi^a).\end{aligned} \tag{3.4}$$

All that remains in the construction is to determine the structure of the various concomitants of h^{μ}_a and Φ^a as a consequence of their symmetry properties and transformation laws, viz.:

(i) $L_{\alpha\beta\gamma\omega} = -L_{\beta\alpha\gamma\omega} = -L_{\alpha\beta\omega\gamma}$,
$\overline{L}_{\alpha\beta\gamma\omega} = L_{\alpha\beta\gamma\omega}$,
$L'_{\mu\nu\sigma\tau}\mathscr{L}^{\mu}_{\alpha}\mathscr{L}^{\nu}_{\beta}\mathscr{L}^{\sigma}_{\gamma}\mathscr{L}^{\tau}_{\omega} = L_{\alpha\beta\gamma\omega}$;

(ii) $L_{\alpha\beta\gamma} = -L_{\beta\alpha\gamma}$,
$\overline{L}_{\alpha\beta\gamma} = L_{\alpha\beta\gamma}$,
$L'_{\mu\nu\sigma}\mathscr{L}^{\mu}_{\alpha}\mathscr{L}^{\nu}_{\beta}\mathscr{L}^{\sigma}_{\gamma} = L_{\alpha\beta\gamma}$;

(iii) $\overline{L}_{\alpha\beta} = L_{\alpha\beta}$,
$L'_{\mu\nu}\mathscr{L}^{\mu}_{\alpha}\mathscr{L}^{\nu}_{\beta} = L_{\alpha\beta}$;

(iv) $L^{ij}_{\alpha\beta} = -L^{ji}_{\alpha\beta} = -L^{ij}_{\beta\alpha}$,
$\overline{L}^{ij}_{\alpha\beta}J^a_i J^b_j = JL^{ab}_{\alpha\beta}$,
$L'^{ij}_{\mu\nu}\mathscr{L}^{\mu}_{\alpha}\mathscr{L}^{\nu}_{\beta} = L^{ij}_{\alpha\beta}$;

(v) $L^{ij}_{\alpha} = -L^{ji}_{\alpha}$,
$\overline{L}^{ij}_{\alpha}J^a_i J^b_j = JL^{ab}_{\alpha}$,
$L'^{ij}_{\mu}\mathscr{L}^{\mu}_{\alpha} = L^{ij}_{\alpha}$;

(vi) $\overline{L}_0 = JL_0$,
$L'_0 = L_0$.

We begin by considering the quantity

$$B_0 = B_0(h^{\mu}_a;\Phi^a) = L_0/h,$$

which has the transformation laws

$$\overline{B}_0 = B_0$$

and

$$B'_0 = B_0. \tag{3.5}$$

Expansion of (3.5) gives

$$B'_0(\mathscr{L}^{\mu}_{\beta}h^{\beta}_a;\Phi^a - h^a_{\beta}l^{\beta}_{\phi}u^{\phi}) = B_0(h^{\mu}_a;\Phi^a).$$

By taking the derivative with respect to u^{γ} and evaluating at the identity transformation, we obtain

$$-\frac{\partial B_0}{\partial\Phi^a}h^a_{\gamma} = 0$$

and thus

$$\frac{\partial B_0}{\partial\Phi^a} = 0.$$

Lemma A1 of the Appendix then yields

$$B_0 = c,$$

where c is an arbitrary constant and hence

$$L_0 = ch.$$

In a similar manner the remaining quantities are all independent of Φ^a, and we have:

(i) $L_{\alpha\beta\gamma\omega} = a_1\epsilon_{\alpha\beta\gamma\omega} + \tfrac{1}{2}a_2(\eta_{\alpha\gamma}\eta_{\beta\omega} - \eta_{\alpha\omega}\eta_{\beta\gamma})$,

by Lemma A2 of the Appendix;

(ii) $L_{\alpha\beta\gamma} = 0$,

by Lemma A3 of the Appendix;

(iii) $L_{\alpha\beta} = d\eta_{\alpha\beta}$,

by Lemma A4 of the Appendix;

(iv) $L^{ij}_{\alpha\beta} = hh^i_{\mu}h^j_{\nu}\eta^{\mu\gamma}\eta^{\nu\omega}[b_1\epsilon_{\alpha\beta\gamma\omega} + \tfrac{1}{2}b_2(\eta_{\alpha\gamma}\eta_{\beta\omega} - \eta_{\alpha\omega}\eta_{\beta\gamma})]$,

by applying Lemma A2 of the Appendix to

$$D_{\alpha\beta\gamma\omega} \equiv (1/h)\eta_{\gamma\mu}h^{\mu}_i\eta_{\omega\nu}h^{\nu}_j L^{ij}_{\alpha\beta};$$

(v) $L^{ij}_{\alpha} = 0$,

by applying Lemma A3 of the Appendix to

$$D_{\alpha\beta\gamma} \equiv (1/h)\eta_{\alpha\mu}h^{\mu}_i\eta_{\beta\nu}h^{\nu}_j L^{ij}_{\gamma};$$

where a_1, a_2, b_1, b_2, and d are all arbitrary constants.

We have thus established the following:

Theorem 3.1: If a Lagrangian of the form

$$L = L(h^{\alpha}_i;\Phi^i;A^{\alpha\beta}_i;A^{\alpha\beta}_{i,j};A^{\alpha}_i;A^{\alpha}_{i,j})$$

has the transformation laws

$$\overline{L} = JL$$

and

$$L' = L,$$

and is degenerate in the sense that its Euler–Lagrange expressions satisfy

$$E^k_{\sigma\tau} = E^k_{\sigma\tau}(h^{\alpha}_i;h^{\alpha}_{i,j};\Phi^i;\Phi^i{}_{,j};A^{\alpha\beta}_i;A^{\alpha\beta}_{i,j};A^{\alpha}_i;A^{\alpha}_{i,j})$$

and

$$E^k_{\sigma} = E^k_{\sigma}(h^{\alpha}_i;h^{\alpha}_{i,j};\Phi^i;\Phi^i{}_{,j};A^{\alpha\beta}_i;A^{\alpha\beta}_{i,j};A^{\alpha}_i;A^{\alpha}_{i,j})$$

then L is restricted to being

$$\begin{aligned}L = {}&a_1\epsilon^{ijkh}\epsilon_{\alpha\beta\gamma\omega}F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma\omega}{}_h + a_2\epsilon^{ijkh}\eta_{\alpha\gamma}\eta_{\beta\omega}F_i{}^{\alpha\beta}{}_j F_k{}^{\gamma\omega}{}_h\\ &+b_1hh^i_{\mu}h^j_{\nu}\eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}F_i{}^{\alpha\beta}{}_j + b_2hh^i_{\alpha}h^j_{\beta}F_i{}^{\alpha\beta}{}_j + ch\\ &+d\epsilon^{ijkh}\eta_{\alpha\beta}f_i{}^{\alpha}{}_j f_k{}^{\beta}{}_h,\end{aligned} \tag{3.6}$$

where a_1, a_2, b_1, b_2, c and d are arbitrary constants.

Remark 1: There is only one additional term due to the

presence of Φ^i in the Lagrangian, viz., the coefficient of d.

Remark 2: It was shown in Ref. 2 that the coefficients of a_1 and a_2 are divergences, and thus their Euler–Lagrange expressions are identically zero.

The Euler–Lagrange expressions for the Lagrangian (3.6) take the form

$$\begin{aligned}\mathscr{E}^s_\phi \equiv \frac{\partial L}{\partial h^\phi_s} &= b_1 h\,(h^s_\phi h^i_\mu h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}F_i^{\alpha\beta}{}_j \\ &\quad - 2h^s_\mu\, h^i_\phi h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}\, F_i^{\alpha\beta}{}_j) \\ &\quad + b_2 h\,(h^s_\phi\, h^i_\alpha h^j_\beta\, F_i^{\alpha\beta}{}_j - 2h^s_\alpha\, h^i_\phi h^j_\beta\, F_i^{\alpha\beta}{}_j) \\ &\quad + chh^s_\phi + 2d\,{}^*f^i{}_\alpha{}^j F_i^{\alpha\beta}{}_j \eta_{\beta\phi}\Phi^s, \end{aligned} \tag{3.7}$$

$$E_j \equiv \frac{\partial L}{\partial \Phi^j} = 4d\,{}^*f^a{}_\mu{}^b{}_{\|ba}\, h^\mu_j, \tag{3.8}$$

$$E^s_\alpha = -4d\,{}^*f^s{}_\alpha{}^t{}_{\|t}, \tag{3.9}$$

and

$$\begin{aligned}E^s_{\alpha\beta} &= b_1\,{}^\dagger K^s_{\alpha\beta} + b_2 K^s_{\alpha\beta} - 2d\,(\eta_{\beta\gamma}\,\Phi^\gamma\,{}^*f^s{}_\alpha{}^t)_{\|t} \\ &\quad + 2d\,(\eta_{\alpha\gamma}\,\Phi^\gamma\,{}^*f^s{}_\beta{}^t)_{\|t}, \end{aligned} \tag{3.10}$$

where

$$\begin{aligned}&{}^*f^i{}_\alpha{}^j \equiv \epsilon^{ijkh}\eta_{\alpha\beta}f_k{}^\beta{}_h, \\ &K^s_{\alpha\beta} \equiv -2h\,(h^s_{[\alpha}h^t_{\beta]})_{\|t}, \\ &{}^\dagger K^s_{\alpha\beta} \equiv K^s_{\mu\nu}\,\eta^{\mu\gamma}\,\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}, \end{aligned}$$

and square brackets around indices denotes antisymmetrization. It should be noted that E_j and E^s_α are not independent. In fact, even for a Lagrangian which is not degenerate, one of the conservation laws corresponding to (4.8) in Ref. 3 is

$$E_j = -h^\mu_j E^a_{\mu\|a}.$$

4. A CHOICE OF GAUGE

The lack of independence of the Euler–Lagrange expressions suggests that perhaps a particular gauge transformation could simplify the field equations while reducing the degrees of freedom. Such a transformation is given by

$$u^{\alpha\beta} = 0 \tag{4.1}$$

and

$$u^\alpha = \Phi^\alpha,$$

in which case the transformed field variables (signified by a dot), are

$$\dot{h}^i_\alpha = h^i_\alpha$$

and

$$\dot{\Phi}^i = 0.$$

Thus, Φ^i can be thought of as a Goldstone field.

Even though

$$\dot{\Phi}^\alpha = 0$$

and

$$\dot{F}_i^{\alpha\beta}{}_j = F_i^{\alpha\beta}{}_j,$$

and hence

$$\dot{f}_i{}^\alpha{}_j = \dot{F}_i{}^\alpha{}_j = f_i{}^\alpha{}_j,$$

the double covariant derivative of $\dot{\Phi}^\alpha$ does not vanish; in fact,

$$\dot{\Phi}^\alpha{}_{\|i} = \dot{A}^\alpha_i. \tag{4.2}$$

Thus any reference to Φ^α and its derivatives can be eliminated from both the Lagrangian and the field equations.

We still have the freedom to perform any Lorentz gauge transformation. It is then possible to say that we have obtained a Lorentz gauge theory from a Poincaré gauge theory by means of a Higgs mechanism. The Lagrangian is of the form

$$\dot{L} = \dot{L}\,(\dot{h}^\alpha_i; \dot{A}^{\alpha\beta}_i; \dot{A}^{\alpha\beta}_{i,j}; \dot{A}^\alpha_i; \dot{A}^\alpha_{i,j}),$$

and $\dot{A}^\alpha_i$ is no longer regarded as the translation gauge connection, but as a set of vector fields which transform in the same way as $\dot{h}^\alpha_i$. The Lagrangian (3.6) can then be expressed in this gauge as

$$\begin{aligned}\dot{L} &= a_1\,\epsilon^{ijkh}\epsilon_{\alpha\beta\gamma\omega}\,F_i^{\alpha\beta}{}_j\,F_k^{\gamma\omega}{}_h + a_2\epsilon^{ijkh}\eta_{\alpha\gamma}\eta_{\beta\omega}\,F_i^{\alpha\beta}{}_j\,F_k^{\gamma\omega}{}_h \\ &\quad + b_1 hh^i_\mu h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}\,F_i^{\alpha\beta}{}_j + b_2 hh^i_\alpha h^j_\beta\,F_i^{\alpha\beta}{}_j + ch \\ &\quad + 4d\epsilon^{ijkh}\eta_{\alpha\beta}\dot{A}^\alpha_{i\|j}\dot{A}^\beta_{k\|h}, \end{aligned} \tag{4.3}$$

where we have made use of (2.4) and (4.2) and it is now legitimate to consider the double covariant derivative of $\dot{A}^\alpha_i$.

Corresponding to (3.7)–(3.10), we have the Euler–Lagrange expressions

$$\begin{aligned}\dot{\mathscr{E}}^s_\phi &= b_1 h\,(h^s_\phi h^i_\mu h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}F_i^{\alpha\beta}{}_j \\ &\quad - 2h^s_\mu h^i_\phi h^j_\nu \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega}F_i^{\alpha\beta}{}_j) \\ &\quad + b_2 h\,(h^s_\phi h^i_\alpha h^j_\beta\, F_i^{\alpha\beta}{}_j - 2h^s_\alpha h^i_\phi h^j_\beta\, F_i^{\alpha\beta}{}_j) + chh^s_\phi, \end{aligned} \tag{4.4}$$

$$\begin{aligned}\dot{E}^s_\alpha &= -8d\epsilon^{stij}\eta_{\alpha\beta}\,\dot{A}^\beta_{i\|jt} \\ &= -4d\epsilon^{stij}\eta_{\alpha\beta}\eta_{\gamma\omega}\,\dot{A}^\gamma_i F_j^{\beta\omega}{}_t, \end{aligned} \tag{4.5}$$

and

$$\dot{E}^s_{\alpha\beta} = b_1\,{}^\dagger K^s_{\alpha\beta} + b_2 K^s_{\alpha\beta} + 4d\eta_{\gamma[\alpha}\eta_{\beta]\omega}\epsilon^{stij}\dot{A}^\omega_{i\|j}\,\dot{A}^\gamma_t. \tag{4.6}$$

Note that we have no Euler–Lagrange expression corresponding to (3.8) due to the elimination of Φ^i.

It should be stressed that we do not have a true Lorentz gauge theory here, but one that has been obtained from a Poincaré gauge theory through symmetry breaking involving a Higgs mechanism. The fields $\dot{A}^\alpha_i$ do not arise in a true Lorentz gauge theory without sources.

5. COMPLETE REDUCTION TO LORENTZ

In the particular gauge (4.1) where only subsequent Lorentz transformations are allowed, the ordered pair $(\dot{A}^{\alpha\beta}_i, \dot{A}^\alpha_i)$ can be regarded as the restriction to the Poincaré subgroup of a generalized affine connection as defined by Kobayashi and Nomizu.[9] Furthermore, if we assume

$$\dot{A}^\alpha_i = h^\alpha_i, \tag{5.1}$$

then we have a Poincaré restriction of their affine connection. In doing so, we have completed a reduction[9,10] of the Poincaré theory to a Lorentz theory by means of soldering[7] in addition to the use of a Higgs mechanism. The fields $\dot{A}^\alpha_i$ have now been eliminated.

Some authors[8,11] regard the assumption (5.1) as essential, while others[12] feel that it is not absolutely necessary to

perform such a reduction in all Poincaré gauge theories. When discussing this point few authors stress the fact that it is possible to identify the translation connection and the vierbein only under this choice of gauge where Φ^{α} vanishes and just Lorentz transformations are then allowed. The two quantities transform differently under a general Poincaré transformation, and it does not make sense to take the double covariant derivative of A_i^{α} except under this choice of gauge when we consider that we have just a Lorentz theory.

The above difficulties are overcome by assuming

$$\Phi^{\alpha}{}_{\|i} = h_i^{\alpha} \tag{5.2}$$

instead,[8,11] which reduces to (5.1) under our particular choice of gauge. This effectively completes the reduction by combining the Higgs mechanism and the soldering into one process. A particular choice of gauge is not required.

To see what effect a complete reduction has on our Poincaré gauge theory, we shall generalize (5.2) to

$$\Phi^{\alpha}{}_{\|i} = \kappa h_i^{\alpha}, \tag{5.3}$$

where κ is a constant. This yields the more useful relation

$$f_i{}^{\alpha}{}_j = \Phi^{\alpha}{}_{\|ij} - \Phi^{\alpha}{}_{\|ji} = \kappa(h^{\alpha}_{i\|j} - h^{\alpha}_{j\|i}). \tag{5.4}$$

There are actually two ways to impose (5.3). *A priori* we can substitute (5.3) and (5.4) into the Lagrangian (3.6) and thereby reduce the number of field variables. *A posteriori* it is possible to adjoin (5.3) to the Euler–Lagrange equations corresponding to (3.7)–(3.10). The results are not always the same.[13]

When (5.4) is substituted into (3.6), the coefficient of d becomes

$$4\kappa^2\epsilon^{ijkh}\eta_{\alpha\beta}\, h^{\alpha}_{i\|j} h^{\beta}_{k\|h},$$

which can be expressed as

$$(4\kappa^2\epsilon^{ijkh}\eta_{\alpha\beta}\, h_i^{\alpha} h^{\beta}_{k\|h})_{\|j} - 4\kappa^2\epsilon^{ijkh}\eta_{\alpha\beta} h_i^{\alpha} h^{\beta}_{k\|hj}.$$

By virtue of the commutation law

$$h^{\beta}_{k\|hj} - h^{\beta}_{k\|jh} = -R_k{}^a{}_{hj} h_a^{\beta} + h_k^{\omega} F_h{}^{\beta\gamma}{}_j\, \eta_{\gamma\omega}$$

and the identities

$$R_k{}^a{}_{hj} + R_h{}^a{}_{jk} + R_j{}^a{}_{kh} = 0$$

and

$$\epsilon^{ijkh} = -hh^i_{\alpha} h^j_{\beta} h^k_{\gamma} h^h_{\omega} \eta^{\alpha\mu}\eta^{\beta\nu}\eta^{\gamma\sigma}\eta^{\omega\tau}\epsilon_{\mu\nu\sigma\tau},$$

the coefficient of d in (3.6) takes the form

$$(4\kappa^2\epsilon^{ijkh}\eta_{\alpha\beta} h_i^{\alpha} h^{\beta}_{k\|h})_{\|j} + 2\kappa^2 hh^i_{\mu} h^j_{\nu}\, \eta^{\mu\gamma}\eta^{\nu\omega}\epsilon_{\alpha\beta\gamma\omega} F_i{}^{\alpha\beta}{}_j.$$

Since the first term is a divergence and the second term is proportional to the coefficient of b_1 in (3.6), the effective reduced Lagrangian is

$$L = (b_1 + 2\kappa^2 d)hh^i_{\mu} h^j_{\nu}\, \eta^{\mu\gamma}\eta^{\alpha\omega}\epsilon_{\alpha\beta\gamma\omega} F_i{}^{\alpha\beta}{}_j + b_2 hh^i_{\alpha} h^j_{\beta} F_i{}^{\alpha\beta}{}_j + ch.$$

Therefore, the Euler–Lagrange equations yield[2] the Einstein vacuum field equations with cosmological term, i.e.,

$$b_2 R_{ij} = \tfrac{1}{2} c g_{ij},$$

provided

$$(b_1 + 2\kappa^2 d)^2 + b_2{}^2 \neq 0.$$

In a similar, but more tedious, manner, the *a posteriori* imposition of (5.3) in addition to the Euler–Lagrange equations of (3.6) also yields the Einstein vacuum field equations with cosmological term, subject to the same restriction on the constants.

6. DISCUSSION

We have constructed the Lagrangian of a true Poincaré gauge theory whose Euler–Lagrange equations can be simplified by means of a Higgs mechanism. In this form the translation subgroup is manifested only in the translation connection $\dot{A}_i^{\alpha}$. The usual interpretation of such $\dot{A}_i^{\alpha}$ in a gauge theory using a Higgs mechanism is that they are regarded as a set of vector bosons.[5] Thus the generalization of the theory from Lorentz to Poincaré gives rise to an interaction of a set of vector bosons with the gravitational field. An interesting feature of the Lagrangian (4.3) is that minimal coupling arose without having to impose it.

In this paper complete reduction of the Poincaré theory to the Lorentz theory is regarded merely as a check that the Einstein vacuum field equations can be obtained in some sort of limit. Complete reduction eliminates all aspects of the translation subgroup, and thus we no longer have a Poincaré gauge theory. Therefore, complete reduction should not be required.

ACKNOWLEDGMENT

I would like to thank the Natural Sciences and Engineering Research Council of Canada for its award of an operating grant to conduct this research.

APPENDIX

The following lemmas which are used in the body of the paper were proved in Ref. 2:

Lemma A1: If a quantity $B_0 = B_0(h_i^{\alpha})$ is a scalar under both coordinate and Poincaré gauge transformations, i.e., $\overline{B}_0 = B_0$ and $B'_0 = B_0$, then

$$B_0 = c,$$

where c is an arbitrary constant.

Lemma A2: If a quantity $B_{\alpha\beta\gamma\omega} = B_{\alpha\beta\gamma\omega}(h_i^{\mu})$ has the antisymmetries

$$B_{\alpha\beta\gamma\omega} = -B_{\beta\alpha\gamma\omega} = -B_{\alpha\beta\omega\gamma}$$

and the transformation laws

$$\overline{B}_{\alpha\beta\gamma\omega} = B_{\alpha\beta\gamma\omega}$$

and

$$B'_{\rho\nu\sigma\tau}\, \mathscr{L}^{\rho}_{\mu} \mathscr{L}^{\nu}_{\beta} \mathscr{L}^{\sigma}_{\gamma} \mathscr{L}^{\tau}_{\omega} = B_{\mu\beta\gamma\omega},$$

then

$$B_{\alpha\beta\gamma\omega} = a\epsilon_{\alpha\beta\gamma\omega} + b(\eta_{\alpha\gamma}\, \eta_{\beta\omega} - \eta_{\alpha\omega}\, \eta_{\beta\gamma}),$$

where a and b are arbitrary constants.

Lemma A3: If a quantity $B_{\alpha\beta\gamma} = B_{\alpha\beta\gamma}(h_i^{\mu})$ has the antisymmetry

$$B_{\beta\alpha\gamma} = -B_{\alpha\beta\gamma}$$

and the transformation laws

$$\overline{B}_{\alpha\beta\gamma} = B_{\alpha\beta\gamma}$$

and

$$B'_{\rho\nu\sigma}\mathcal{L}^{\rho}_{\mu}\mathcal{L}^{\nu}_{\beta}\mathcal{L}^{\sigma}_{\gamma} = B_{\mu\beta\gamma},$$

then

$$B_{\alpha\beta\gamma} \equiv 0.$$

Lemma A4: If a quantity $B_{\alpha\beta} = B_{\alpha\beta}(h_i^{\mu})$ has the transformation laws

$$\overline{B}_{\alpha\beta} = B_{\alpha\beta}$$

and

$$B'_{\rho\nu}\mathcal{L}^{\rho}_{\mu}\mathcal{L}^{\nu}_{\beta} = B_{\mu\beta},$$

then

$$B_{\alpha\beta} = b\eta_{\alpha\beta},$$

where b is an arbitrary constant.

[1] Y. M. Cho, Phys. Rev. D **14**, 3335 (1976); F. W. Hehl, P. von der Heyde, G. D. Kerlick, and J. M. Nester, Rev. Mod. Phys. **48**, 393 (1976); T. W. B. Kibble, J. Math. Phys. **2**, 212 (1961).
[2] R. J. McKellar, J. Math. Phys. **22**, 2934 (1981).
[3] R. J. McKellar, J. Math. Phys. **22**, 862 (1981).
[4] F. Brickell and R. S. Clarke, *Differentiable Manifolds* (Van Nostrand-Reinhold, London, 1970).
[5] E. S. Abers and B. W. Lee, Phys. Rep. **9**, 1 (1973).
[6] C. N. Yang, "Gauge Fields," in *Proceedings of the Sixth Hawaii Topical Conference on Particle Physics*, edited by P. N. Dobson, Jr., S. Paksava, V. Z. Peterson, and S. F. Tuan (University of Hawaii Press, Honolulu, 1976).
[7] W. Drechsler and M. E. Mayer, *Fibre Bundle Techniques in Gauge Theories* (Springer-Verlag, New York, 1975).
[8] K. A. Pilch, Lett. Math. Phys. **4**, 49 (1980).
[9] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, (Interscience, New York, 1963), Vol. I.
[10] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds and Physics* (Elsevier, New York, 1982).
[11] R. Giachetti, R. Ricci, and E. Sorace, Lett. Math. Phys. **5**, 85 (1981).
[12] L. K. Norris, R. O. Fulp, and W. R. Davis, Phys. Lett. A **79**, 278 (1980); Y. N. Obukov, Phys. Lett. A **90**, 13 (1982).
[13] See, e.g., J. L. Safko, M. Tsamparlis, and F. Elston, Phys. Lett. A **60**, 1 (1977).

Unified geometrical approach to relativistic particle dynamics

A. P. Balachandran,[a] G. Marmo,[b] N. Mukunda,[c] J. S. Nilsson,[d] A. Simoni,[b] E. C. G. Sudarshan,[e] and F. Zaccaria[b]

Center for Particle Theory and Department of Physics, The University of Texas at Austin, Austin, Texas 78712

(Received 15 June 1982; accepted for publication 10 December 1982)

Models for systems of relativistic particle dynamics are reviewed in terms of a geometrical setting for constraint dynamics. They are derived from the same grand abstract space by means of a common reduction procedure and are put in correspondence with invariant subgroups of the Poincaré group. A new model corresponding to the identity subgroup is also discussed.

PACS numbers: 11.80. – m, 11.30.Cp, 02.20.Rt

I. INTRODUCTION: ON THE DESCRIPTION OF BECOMING

Dynamics is the expression of flow by stringing together sequences of configurations together each labelled by a time evolution parameter according to an explicit rule. The collections of configurations so strung together in a well-ordered sequence constitute trajectories of the system, and each trajectory has certain configurational functionals characterizing them. These would be the constants of motion. In this account the configurations are the conventional coordinate space together with the velocity fibers: whatever constitutes the initial specification to make use of Newton's formulation of the equations of motion.

When such ideas are to be implemented for a relativistic system, we do encounter some new problems. Traditionally, we consider clock time as the time evolution parameter, and a configuration is defined by considering simulataneous specification of coordinates and velocities. In relativistic theory this poses a problem since distant simultaniety is not relativistically invariant. If we insist, nevertheless, on using clock time and a canonical formalism, the no-interaction theorem tells us that the only relativistically invariant descriptions could be for noninteracting systems only. We must therefore be prepared to consider other alternatives.

A satisfactory alternative is to consider a time evolution parameter defined dynamically rather than kinematically. Dynamical evolution is with respect to a temporal parameter that has different significance in different states of motion. The dynamical evolution is self-referring and "the time" is independent of the external reference frames.

It turns out that the temporal parameter so defined, being Lorentz-invariant, must have a generator of dynamical evolution which is also Lorentz-invariant, and is different from any of the ten generators of the Poincaré group. In this 11 parameter generator formalism it has been found possible to construct interacting relativistic systems with invariant world lines.

The natural mechanism for bringing about such a description is to make use of the Dirac constraint formalism starting with a system with excess degrees of freedom and systematically reducing them by imposing constraints. Among those constraints we include one which explicitly depends on a parameter τ, which then gets identified with being the evolution parameter. We have thus the curious situation in which motion is generated by constraints.

In the recent literature there have been a number of such models constructed; they are of three kinds depending upon how the initial configuration and phase spaces are chosen. Each such group made use of a primary set of dynamical variables and a set of constraints. In the first kind of models each individual particle is described by four pairs of canonical variables. A system of $2N$ constraints are then imposed to produce $3N$ pairs of canonical variables and an evolution parameter to describe N particles in motion. In the second kind of model a pair of 4-vectors represent spacetime specification of a uniformly moving "center" of the system and the total 4-momentum of the system, respectively. The constraints then relate these quantities to the particle configurations. In the third kind of model the new collective variables introduced are a Lorentz matrix and its canonical conjugate carrying the burden of the inertial frame. Constraints can then be used to obtain interacting relativistic particles describing world lines.

Each of these kinds of models has its own number of starting variables and judiciously chosen constraints. It would be desirable to have a systematic method of dealing with all three models and to see if there are other possibilities of a similar kind.

The present paper is devoted to this task. We start with grand abstract configuration space $\tilde{\Sigma}$ consisting of the semidirect product of the Lorentz group with the product of N 4-vectors. This configuration space thus has $4N + 10$ dimensions. The phase space has twice this dimension. We then take an invariant subgroup G of the Poincaré group P and take the equivalence classes.

$$\Sigma = \tilde{\Sigma}/G$$

as the configuration space of a model. It turns out that by

[a] Supported by the U.S. Department of Energy under Contract DE-AC02-76ERO 3533. Permanent address: Physics Department, Syracuse University, Syracuse, NY 13210.

[b] Istituto di Fisica Teorica, Università di Napoli and Istituto Nazionale Fisica Nucleare, Sezione di Napoli Mostra d'Oltremare Pad. 19, Napoli, Italy.

[c] Permanent address: Indian Institute of Science, Bangalore 560012, India.

[d] Permanent address: Institute of Theoretical Physics, S-41296, Göteborg, Sweden.

[e] Supported by the U.S. Department of Energy under Contract DE-AS05-76ERO 3992.

 0022-2488/84/010167-10$02.50

choosing G to be P itself, the Lorentz subgroup α, and the translation subgroup T^4, respectively, we get the three kinds of models mentioned above. By choosing the identity subgroup of P we are able to generate another kind of model.

Much of our previous work as well as that of other authors are stated in traditional language of canonical mechanics. For making the ideas accessible to a wider group of people to whom modern differential geometry is a standard tool as well as to expose the essential geometric aspects of the developments, we have carried out our formulation in the language of differential geometry.

The plan of the paper is as follows: Sec. II recapitulates the essential background to establish notation and provide the setting. The world line condition is formulated in its general form in Sec. III. The grand configuration space is introduced in Sec. IV along with the equivalence classes which realize the four kinds of formalisms. In Sec. V we construct the phase spaces and the choice of constraints to build up a suitable family of sections of the fiber bundle for each of the models. Some remarks in Sec. VI conclude the paper.

II. A GEOMETRICAL SETTING FOR CONSTRAINT DYNAMICS

In dealing with constraint dynamics, the situation we are presented with is the following.

On a given $2n$-dimensional manifold[1] $\Gamma = T^*\Sigma$ a set of real functions $K_1,...,K_k$ is given. By choosing a value for each one of them a hypersurface M in Γ is determined. We consider the smooth map

$$\kappa\colon \Gamma \to \mathbb{R}^k,$$
$$\gamma \to (K_1(\gamma),...,K_k(\gamma)),$$

and by fixing a value, say $0 \in \mathbb{R}^k$, we get

$$M = \kappa^{-1}(0) = \{\gamma \in \Gamma\colon K_1(\gamma) = \cdots = K_k(\gamma) = 0\}.$$

We assume M to be a submanifold of Γ, of codimension k. If $0 \in \mathbb{R}^k$ is a regular value for κ, then M is a submanifold.

By means of the symplectic structure ω on Γ we can define Poisson brackets and associate vector fields with functions. The vector field X_f associated with the function f is defined by the relation

$$L_{X_f} g = \{f, g\}$$

for any function g. An equivalent definition is given by

$$i_{X_f}\omega = df$$

if ω is the symplectic form of Γ.

A set of vector fields $X_1,...,X_r$ spans a tangent subspace for each point of Γ on considering span $\{X_1(\gamma),...,X_r(\gamma)\}$. Such spaces will constitute the tangent space of a submanifold if and only if the relations.

$$[X_i, X_j] = c_{ij}^m X_m \tag{2.1}$$

are satisfied, with the c_{ij}^m being functions on Γ. This is the Frobenius theorem.

A vector field X can be evaluated at points of M. If it turns out that $X(m)$ is tangent to M for any $m \in M$, we will say that X is tangent to M.

With the above set of functions we will associate the vector fields X_{K_j} and inquire about the relation (2.1). It is simple to prove that they satisfy the condition of the Frobenius theorem if and only if the following relations hold:

$$d\{K_i, K_j\} = c_{ij}^m\, dK_m.$$

The c_{ij}^m will then be functions of the K_i. We say in this case that the K_i form a function group. Such a situation leads to a foliation on Γ and the relevant analysis has been carried out in Ref. 2, to which we will refer extensively in what follows.

Here we do not require the K_i to form a function group; nevertheless, we shall show how, starting with the vector fields X_{K_i} restricted to M, we can generate a set of vector fields tangent to M and satisfying the condition for the Frobenius theorem.

If

$$i\colon M \to \Gamma$$

is the identification map, we can consider the 2-form $i^*\omega$ on M, which is the pullback of ω by i. In general, $i^*\omega$ is degenerate. If its rank is constant the vector fields on M annihilated by it constitute an involutive distribution $\mathscr{D}$, i.e., they obey the Frobenius theorem. We will prove that they are combinations (with coefficients functions on M) of the X_{K_i} evaluated on M. (Notice that in general the X_{K_i} are not tangent to M.) They will be denoted by Y, and the hypothesis is that they satisfy

$$i_Y(i^*\omega) = 0.$$

This implies that

$$(i_Y\omega)_{|M} = 0$$

and therefore one can write

$$i_Y\omega = c_i\, dK_i \quad \text{(summed on } i\text{)}$$

or

$$Y = c_i\, X_{K_i}$$

with the c_i being functions on M. (Here there is an abuse of notation, as Y is actually a vector field on M, but we do consider it as a vector field on Γ.)

Such an expression for Y implies

$$c_i\{K_i, K_j\} = 0 \quad \text{on } M \text{ for any } j = 1,...,k.$$

When a relation involving Poisson brackets is true only when evaluated on M, it is customary to replace the equality sign $=$ with the sign $\approx$ and it is said to be true in a weak sense. Thus our relations can be written as

$$c_i\{K_i, K_j\} \approx 0 \quad \text{for any } j = 1,...,k. \tag{2.2}$$

It is useful to define the antisymmetric matrix A:

$$A_{ij} = \{K_i, K_j\} \tag{2.3}$$

related to $i^*\omega$ by

$$\operatorname{rank} A(m) = \operatorname{rank}(i^*\omega)(m), \quad m \in M.$$

The set of (c_i) can now be considered as nullvectors of $A_{|M}$ and the number of independent nonvanishing vector fields satisfying (2.2) turns out to be

$$d = \operatorname{codim} M - \operatorname{rank} A_{|M}.$$

If $\operatorname{rank} A_{|M}$ is to be a constant on M, the vector fields Y define an involutive distribution $\mathscr{D}$ on M with the above dimension. This allows us to foliate M and to consider

$$\mathscr{N} = M/\mathscr{D}.$$

In physics it is customary to assume $\mathcal{N}$ to be a manifold having the property that

$$\pi: M \to \mathcal{N}$$

is a submersion. It can be proved that $\mathcal{N}$ inherits a symplectic structure ρ, which allows us to call it the "reduced phase space" or "the frozen phase space."[3]

But so far no dynamics has been defined at all. This is done by introducing a one-parameter family of sections

$$\mathcal{N} \times \mathbb{R} \overset{\sigma}{\to} M.$$

From a global point of view this assumes that a section for $M \overset{\pi}{\to} \mathcal{N}$ does exist. (If the vector fields Y integrate to a Lie group $\mathcal{G}$, such that the leaves of the submersion $\pi: M \to \mathcal{N}$ are diffeomorphic to $\mathcal{G}$, the existence of such a section requires the $\mathcal{G}$-bundle to be trivial.) It is on $\sigma(\mathcal{N} \times \mathbb{R}) \subset M$ that dynamics will be defined, not on M itself. The leaves of π are d-dimensional, and it turns out that $k + d$ is an even number. Therefore,

$$\dim \mathcal{N} = 2n - (k + d)$$

is even, and

$$\dim[\sigma(\mathcal{N} \times \mathbb{R}) \subset M] = 2n - (k + d) + 1, \quad d > 0.$$

Of course, if $d = 0$, then $\mathcal{N} = M$, $\dim \sigma(\mathcal{N} \times \mathbb{R}) = 2n - k$, and our procedure generates a dynamics (the trivial one), i.e., a one-parameter group of transformations on M, which is independent of K_i. But in general this is not the case and the set of K_i has a further role. All possible dynamics that can be defined in such a fashion, corresponding to different choices of σ, have the property that the manifolds of states of motion are all diffeomorphic among themselves.

If $Y_1, Y_2, \ldots, Y_d$ are a basis of vector fields which span $\ker i^*$ each dynamical vector field Δ can be expressed as

$$\Delta = \alpha^i Y_i$$

with α^i functions on M. All this is restricted to the submanifold $\sigma(\mathcal{N} \times \mathbb{R}) \subset M$. This vector field Δ is tangent to the submanifold.

But another way to build up dynamics and the appropriate submanifold is commonly used in dealing with constraint dynamics. Besides the K_i functions, another set of d real functions $X_1, \ldots, X_d$ is chosen to constitute the smooth map

$$\mathbb{X}: \Gamma \times \mathbb{R} \to \mathbb{R}^d,$$
$$(\gamma, \tau) \to X^\tau(\gamma).$$

The requirement on the X is that they are functionally independent and together with the K_i define for each value of the parameter τ a $[2n - (k + d)]$-dimensional surface in Γ on which ω turns out to be nondegenerate. To put it differently, the equations

$$c_m \{ \xi_m, \xi_n \} \approx 0$$
$$(m, n = 1, \ldots, k + d) \quad \text{(summed on } m\text{)}$$

(where ξ_m stands for $K_1, \ldots, K_k; X_1, \ldots, X_d$) do not have nontrivial solutions. Then for each $\tau \in \mathbb{R}$ the surface generated by

$$\mathbb{K} \times \mathbb{X}^\tau: \Gamma \to \mathbb{R}^{d+k}$$

by taking the inverse image of $0 \in \mathbb{R}^{d+k}$ is of dimension $2n - (k + d)$. In this way one recovers what was earlier called $\sigma(\mathcal{N} \times \mathbb{R})$, as will be seen in the next section.

From the previous discussion it is clear that different X_i define different dynamical systems even if all of them have diffeomorphic spaces of trajectories. Their carrier spaces may be different.

In many physical situations, the starting space Γ carries a symplectic action $\overline{\mathcal{R}}$ of some Lie group G, i.e., G acts on Γ via canonical transformations. We ask ourselves what happens to such an action with respect to the constraint surface M. It is obvious that only that part of G which maps M onto itself is relevant as far as dynamics is concerned. If all the infinitesimal generators X^G for $\overline{\mathcal{R}}$ happen to satisfy the relations

$$(i_{X^G} dK_i)_{|M} = 0 \quad (i = 1, \ldots, k)$$

then the action carries over to the manifold M. Furthermore, as the action of G on M preserves $i^*\omega$, it happens that $\mathcal{N}$ also will carry a G-action, $\overline{\mathcal{R}}$, which is symplectic with respect to the symplectic structure ρ. This statement follows from the fact that the vector fields $\tilde{Y}$ defined by

$$i_{\tilde{Y}} \omega = d(c_i K_i)$$

when restricted to M coincide with

$$Y = c_i X_{K^i}.$$

Since $(\overline{\mathcal{R}})^* \omega = \omega$ and M is invariant under $\overline{\mathcal{R}}$, we have also

$$(\overline{\mathcal{R}})^* \mathcal{D} = \mathcal{D}.$$

In fact $(\overline{\mathcal{R}})^*(i_X \omega) = i_{\overline{\mathcal{R}}_* \omega}$, if X is a vector field on Γ.[4]

As we have already said, a dynamics is specified only after we have a section

$$\sigma: \mathcal{N} \times \mathbb{R} \to M$$

and it will be a dynamics on $\sigma(\mathcal{N} \times \mathbb{R})$. The submanifold $\sigma(\mathcal{N} \times \{0\}) \subset \sigma(\mathcal{N} \times \mathbb{R})$ can be thought of as the set of all possible Cauchy data for our dynamics. Furthermore, the projected action of G on $\mathcal{N}$ gives an action of G on $\sigma(\mathcal{N} \times \{0\})$ by setting

$$\mathcal{R}^*(g)\sigma(n, 0) = \sigma(\overline{\mathcal{R}}(g)n, 0), \quad n \in \mathcal{N}, \ g \in G.$$

This can be extended to $\sigma(\mathcal{N} \times \mathbb{R})$ by the relation

$$\mathcal{R}^*(g)\sigma(n, \tau) = \sigma(\overline{\mathcal{R}}(g)n, \tau).$$

It is obvious that $\mathcal{R}^*$ is equivariant with respect to the projection $\pi: M \to \mathcal{N}$ restricted to $\sigma(\mathcal{N} \times \mathbb{R}) \to \mathcal{N}$. It is also clear that it depends on the section $\sigma: \mathcal{N} \times \mathbb{R} \to M$. Moreover, it is canonical with respect to the Poisson brackets on $\sigma(\mathcal{N} \times \mathbb{R})$ defined by the symplectic form $\pi_\tau^* \rho$ the pullback of the symplectic form ρ on $\mathcal{N}$ by the map $\pi_\tau: \sigma(\mathcal{N} \cdot \{\tau\}) \to \mathcal{N}$. This coincides with the usual action generated by Dirac brackets defined on all Γ and restricted to $\sigma(\mathcal{N} \times \{\tau\})$.

But, to connect all this with the evolution of physical objects, it will be necessary to properly define the physical variables, namely positions and momenta in spacetime. In the following sections, maps ϕ_a and ψ_a will be introduced, respectively, for the position and momentum 4-vectors of the ath particle. As the group G involved will be the Poincaré group, it will have the usual action on them. We will denote it by $\mathcal{R}_{\text{reg}}$.

We remark that as both dynamics and states of motion

are given by the choice of a section σ, it is the above action $\mathscr{R}^*$ of the Poincaré group that is the physically relevant one.

In the following sections we are going to apply the above procedure to some specific models.

In some of the models the starting functions K satisfy the relations

$$\{K_i, K_j\} = c_{ij}^m K_m \quad (i,j,m = 1,...,k),$$

i.e.,

$$\{K_i, K_j\} \approx 0.$$

They are then said to form a first class set of constraints. The additional functions χ, meeting the previously stated requirements, are said to form, together with the K, a second class set of constraints. We have

$$\text{rank } A_{|M} = 0, \quad d = k,$$

and the determinant of the matrix

$$B_{m,n} = \{\xi_m, \xi_n\}_{|M}$$

reduces to $(\det|\{K_i, \chi_j\}|)^2$. The Poisson brackets are evaluated on $(\mathbb{K}\times\mathbb{X})^{-1}(0)$.

In other models the structure of the matrix B allows us to carry out the reduction procedure through intermediate steps. For them $A_{|M}$ is singular and has nonzero rank r. A nonsingular submatrix A', of even rank r, is then formed by a subset of the K, which are a second class system of constraints to begin with, so that Dirac brackets can be computed relative to them only. To have the final set of second class constraints, one adds to the remaining K an equal number of χ satisfying the requirement

$$\det B \neq 0.$$

III. WORLD LINE CONDITION

With the space $\mathscr{N}$ we can associate dynamics according to Sec. II. There we have seen that this dynamics is defined on $\sigma(\mathscr{N}\times\mathbb{R}) \subset M$, not on M itself. As already stated, in each model a map $\phi_a : \Gamma \to$ spacetime will be introduced to denote the position 4-vector of particle a. By restricting ϕ_a to $\sigma(\mathscr{N}\times\mathbb{R})$, with each trajectory we associate a world line on spacetime. The physical interpretation of such world lines requires that this association has a definite Poincaré-covariant property. It is this requirement that is usually called the world line condition (WLC). The formal statement of this condition is as follows.

The association

$$n \in \mathscr{N} \mapsto \sigma(n,\mathbb{R})$$

defines a line in $\sigma(\mathscr{N}\times\mathbb{R})$ for each n. On such a set of lines we had defined a Poincaré group action $\mathscr{R}^*$ by setting

$$\mathscr{R}^*(g)\circ\sigma(n,\mathbb{R}) = \sigma(\bar{\mathscr{R}}(g)n,\mathbb{R}), \quad g\in G.$$

We can now state the WLC

$$\phi_a \circ \mathscr{R}^*(g)\circ\sigma(n,\mathbb{R}) = \mathscr{R}_{\text{reg}}(g)\circ\phi_a\circ\sigma(n,\mathbb{R}),$$

where $\mathscr{R}_{\text{reg}}$ is the usual action on the four-dimensional vector space of spacetime positions.

For computations it is convenient to express the WLC in a more explicit way in terms of parametrized lines. Recall the one parameter family of section σ^τ, introduced in Sec II. By varying τ, a line on M is described for each n. Such a line is in turn projected for each a onto $\mathbb{R}^4_a$ by ϕ_a, thus yielding the world line of particle a:

$$c_a^n : \mathbb{R} \to \mathbb{R}^4_a,$$
$$c_a^n(\tau) = \phi_a \circ \sigma^\tau(n).$$

The WLC becomes in this context the requirement that the actions $\mathscr{R}_{\text{reg}}$ defined on each $\mathbb{R}^4$ and $\bar{\mathscr{R}}$ on $\mathscr{N}$ are physically consistent, in the sense that if $n' = \bar{\mathscr{R}}(g)n$, then there is a τ' such that

$$c_a^{n'}(\tau') = \mathscr{R}_{\text{reg}}(g)c_a^n(\tau). \tag{3.1}$$

Here τ' can depend on τ, g, and a. This obviously poses conditions on σ^τ.

To satisfy the WLC, we construct a section of $\pi: M \to \mathscr{N}$ in terms of the real functions χ of the previous section, and choose the χ suitably. We consider the subsets $(\mathbb{X}^\tau)^{-1}(0) \equiv N^\tau \subset \Gamma$. A first requirement is that

$$N^\tau \cap M \neq 0.$$

A second is that $N^\tau_{|M}$ be transversal with respect to the fibers of $\pi: M \to \mathscr{N}$. This condition is satisfied if no vector field exists in $\mathscr{D}$ with a flow tangent to $N^\tau_{|M}$.

While the first demand is met in all cases by requiring that the components of $\mathbb{X}^\tau$ constitute additional constraints not identically vanishing on M, the second one needs some elaboration.

Referring to Sec. II, a vector field lying in $\mathscr{D}$ was seen to be $X_{\psi_{|M}}$, with ψ being such that

$$\psi = c_i K_i \tag{3.2}$$

and

$$\{\psi, K_j\}_{|M} = 0, \quad \forall j = 1,...,k. \tag{3.3}$$

Hence

$$L_{X_\psi} K_j = c_i \{K_i, K_j\} = 0. \tag{3.4}$$

We proceed to determine the functions c_i. Equation (3.4) can be written as

$$(A\mathbf{c})_{|M} = 0, \tag{3.5}$$

where $\mathbf{c} = (c_1,...,c_k)$ and A is the matrix (2.3). We recall that in all the models

$$\text{rank } A = r < k.$$

This allows us to choose r components of $\mathbb{K}$ in terms of which the submatrix A' of nonzero determinant can be built. They will be denoted K'_l $(l = 1,...,r)$ and the remaining ones K''_h $(h = 1,...,d)$ so that

$$\psi = c'_l K'_l + c''_h K''_h.$$

There are ∞^d solutions of (3.5): the c'' can be arbitrarily chosen and the c' are then computed as the unique solution of a linear inhomogeneous system of dimension r. A set of independent solutions is obtained by starting with each K''_h in turn. We denote it by, ψ_h:

$$\psi_h = K''_h - (A')^{-1}_{ll'}\{K'_l, K''_h\}K'_{l'}.$$

The ψ_h constitute a basis for first class constraints.

Returning now to the transversality condition, this can be formulated as the requirement that the equations

$$(b_h\{\psi_h, \chi_{h'}\}) = 0$$

with b_h real functions on Γ, have only the trivial solution $b_h = 0$. This is possible iff

$$\det|\{\psi_h,\chi_{h'}\}_{|M}|\neq 0. \tag{3.6}$$

We note at this point that

$$\{\psi_h,\chi_{h'}\}_{|M} = \{K''_h,\chi_{h'}\}^*_{|M}, \tag{3.7}$$

the bracket on the right-hand side being the Dirac bracket, relative to the K' only.

When

$$\operatorname{rank} A_{|M} = 0,$$

there are no second class constraints (i.e., no K'), and Eq. (3.6) reduces to

$$\det|\{K_j,\chi_{j'}\}|\neq 0, \quad j,j' = 1,...,k. \tag{3.8}$$

In all the schemes considered χ^τ_i are chosen so that all but one, say χ^τ_d, are τ-independent and constitute a Poincaré-invariant set. The χ_i $(i = 1,...,d-1)$ define a line on each fiber and $\overline{\mathscr{R}}_{|M}$ simply permutes these lines among themselves. Thus the WLC is satisfied because in this action on lines $\overline{\mathscr{R}}_{|M}$ and $\mathscr{R}^*$ agree.

Further imposing $\chi^\tau_d = 0$ then puts a parameter τ on each line which is not necessarily preserved under the $\overline{\mathscr{R}}_{|M}$ action. However, this leads us to define a value for τ' in terms of τ, g and other variables such that the WLC in the form (3.1) is satisfied.

IV. THE CHOICE OF THE VARIABLES

In this section we will discuss the variables used in each model to describe systems of N interacting particles.

The physical positions and momenta, in spacetime, will be denoted by 4-vectors q^μ_a and p^μ_a for the ath particle $(a = 1,...,N)$. They transform under the action $\mathscr{R}_{\text{reg}}$ of $\mathscr{P}$ defined by

$$\mathscr{R}_{\text{reg}} = (L,b)q_a, = Lq_a + b$$

and

$$\mathscr{R}_{\text{reg}}(L,B)p_a = Lp_a,$$

where L is a 4×4 Lorentz matrix and b a translation 4-vector. Let $\mathscr{L}$ denote the Lorentz group $\{L\}$ and T^4 the translation group $\{b\}$.

We start with an abstract space $\tilde{\Sigma}$, on which proper actions of $\mathscr{P}$ will be defined. We will then show how the various models equipped with such q_a and p_a emerge.

Let us define

$$\tilde{\Sigma} = \mathscr{P}\times\Sigma_0.$$

$\mathscr{P}$ is the Poincaré group and $\Sigma_0 = \otimes_{a=1,...,N}\mathbb{R}^4_a$. Elements of $\tilde{\Sigma}$ will be denoted $[(\Lambda,a),(x)]$, in which $(\Lambda,a)\in\mathscr{P}$ and (x) stands for $x_1,...,x_N$, x_a being a vector in $\mathbb{R}^4_a$. The following action of $\mathscr{P}$ is defined:

$$\mathscr{R}^{(1)}(L,b)[(\Lambda,a),(x)] = [(\Lambda,a)(L,b)^{-1},(L,b)(x)],$$

where on the right-hand side the right action on $\mathscr{P}$ is given by group multiplication and the left, on (x), is the $\mathscr{R}_{\text{reg}}$ on each $\mathbb{R}^4$, i.e.,

$$(L,b)x_a = \mathscr{R}_{\text{reg}}(L,b)x_a = Lx_a + b.$$

Endowed with such an action, $\tilde{\Sigma}$ has the structure of a fiber bundle associated with the trivial principal bundle $\mathscr{P}$. It is therefore possible to consider equivalence classes with respect to $\mathscr{R}^{(1)}$ and obtain distinct spaces

$$\Sigma = \tilde{\Sigma}/\mathscr{R}^{(1)}(g) \tag{4.1}$$

corresponding to distinct subgroups g of $\mathscr{P}$,

$$\Gamma = \mathscr{T}^*\Sigma,$$

such that the basic (abstract) variables are taken and the analysis of the previous section starts.

Another action of $\mathscr{P}$ on $\tilde{\Sigma}$ commuting with $\mathscr{R}^{(1)}$ can be defined to make $\tilde{\Sigma}$ a trivial principal $\mathscr{P}$-bundle. This is

$$\mathscr{R}^{(2)}(L,b)[(\Lambda,a),(x)] = [(L\Lambda,La+b),(x)].$$

Going to the quotient as in (4.1), it gives rise to an action $\mathscr{R}$ on Σ, which in turn can be lifted to Γ. The symplectic manifold Γ therefore carries a symplectic action $\overline{\mathscr{R}}$ of $\mathscr{P}$.[5]

Maps will be seen to exist from Γ to spacetime for the physical positions, i.e.,

$$q^\mu_a = \phi^\mu_a(\gamma), \quad \gamma\in\Gamma,$$

with the property that

$$\phi_a\circ\overline{\mathscr{R}}(L,b) = \mathscr{R}_{\text{reg}}(L,b)\circ\phi_a$$

and, analogously, for the momenta, i.e.,

$$p^\mu_a = \psi^\mu_a(\gamma), \quad \gamma\in\Gamma,$$

$$\psi_a\circ\overline{\mathscr{R}}(L,b) = \mathscr{R}_{\text{reg}}(L,b)\circ\psi_a.$$

The above physical maps need not be defined on the whole of Γ but rather on the part $\sigma(\mathscr{N}\times\mathbb{R})$, where dynamics operates, i.e., where all the constraints are satisfied. Furthermore, it is there that the generalized mass shell relations

$$p^2_a - m^2_a - v_a = 0 \tag{4.2}$$

will hold.

In what follows we will consider four models. Each of them corresponds to an invariant subgroup of $\mathscr{P}$ with respect to which the quotient (4.1) is taken. Four such subgroups are considered, namely $\mathscr{P}$ itself, the Lorentz group $\mathscr{L}$, the translations T^4, and the identity.

A. The model I[6-9]

The equivalence classes are taken with respect to $\mathscr{P}$, i.e.,

$$\Sigma = \tilde{\Sigma}/\mathscr{R}^{(1)}(\mathscr{P})$$

and each of them can be represented by a set of N 4-vectors (z), so that

$$\Sigma\simeq(\mathbb{R}^4)^{\otimes N}.$$

In fact, the class to which $[(\Lambda,a),(x)]$ belongs contains also $[(L,0),(\Lambda,a)(x)]$ and if

$$(z) = (\Lambda,a)(x)$$

this can be denoted $\{(z)\}$.

The other variables in $\Gamma = T^*\Sigma$ are (η), the canonical conjugates to (z). So a point in Γ is represented by $\{(z);(\eta)\}$. The action $\overline{\mathscr{R}}$ can be seen to be

$$\overline{\mathscr{R}}(L,b)\{(z);(\eta)\} = \{(Lz+b);(L\eta)\}.$$

This allows us to identify these variables with the physical spacetime positions and momenta. The relations (4.2) will enter in the definition of M.

B. The model II[10,11]

Here the subgroup to be taken in (4.1) is the Lorentz group $\mathscr{L}$ and

$$\Sigma = \tilde{\Sigma}/\mathscr{R}^{(1)}(\mathscr{L}).$$

Since

$$\mathscr{R}^{(1)}(L,0)[(\Lambda,a),(x)] = [(\Lambda L^{-1},a),(Lx)],$$

one sees that $[(\Lambda,a),(x)]$ is equivalent to $[(1,a),(\Lambda x)]$. Thus the elements of Σ can be denoted $\{Q,(z)\}$, the Q and z_a $(a = 1,\ldots,N)$ being 4-vectors, $z = \Lambda x$, so that

$$\Sigma \simeq (\mathbb{R}^4)^{\otimes(N+1)}.$$

The additional variables for $\Gamma = T^*\Sigma$ will be R and (η), the canonical conjugates to Q and (z). A point of Γ may be written $\{Q,(z);R,(\eta)\}$. The action of $\bar{\mathscr{R}}(L,b)$ on it gives $\{LQ + b,(Lz);LR,(L\eta)\}$. The physical variables

$$q_a = Q + z_a, \quad p_a = R + \eta_a$$

transform with $\mathscr{R}_{\text{reg}}$ but are not canonically conjugate. The relations (4.2) are satisfied once all the constraints on Γ have been imposed, i.e., when the sections σ have also been introduced.

C. The model III[12]

The equivalence classes are taken with respect to the translation group T^4, i.e.,

$$\Sigma = \tilde{\Sigma}/\mathscr{R}^{(1)}(T^4).$$

Since

$$\mathscr{R}^{(1)}(1,b)[(\Lambda,a),(x)] = [(\Lambda,a - \Lambda b),(x + b)],$$

we have

$$[(\Lambda,a),(x)] = [(\Lambda,0),(x + A^{-1}a)].$$

This allows us to denote a point of Σ by $\{\Lambda,(z)\}$ where

$$z_a = x_a + \Lambda^{-1}a.$$

This gives

$$\Sigma \simeq \mathscr{L} \times (\mathbb{R}^4)^{\otimes N}.$$

The variables for $\Gamma = T^*\Sigma$ include those for Σ and the "momentum" variables $S_{\mu\nu} = -S_{\nu\mu}$ and (η), which are conjugate to Λ^{μ}_{ν} and (z), respectively. The nonvanishing Poisson brackets are

$$\{z_{a\mu},\eta_{b\nu}\} = \delta_{ab}\delta_{\mu\nu},$$

$$\{\Lambda^{\mu}_{\nu},S_{\alpha\beta}\} = g_{\nu\beta}\Lambda^{\mu}_{\alpha} - g_{\nu\alpha}\Lambda^{\mu}_{\beta},$$

$$\{S_{\mu\nu},S_{\alpha\beta}\} = g_{\mu\alpha}S_{\nu\beta} - g_{\nu\alpha}S_{\mu\beta} + g_{\mu\beta}S_{\alpha\nu} - g_{\nu\beta}S_{\alpha\mu}.$$

As far as $\bar{\mathscr{R}}$ is concerned, we see that

$$\mathscr{R}^{(2)}(L,b)[(\Lambda,0),(x + \Lambda^{-1}a)]$$
$$= [(L\Lambda,b),(x + \Lambda^{-1}a)]$$
$$\simeq [(L\Lambda,0),(x + \Lambda^{-1}a + (L\Lambda)^{-1}b)]$$

so that

$$\bar{\mathscr{R}}(L,b)\{\Lambda,(z);S,(\eta)\} = \{L\Lambda,(z + (L\Lambda)^{-1}b);LSL^{-1},(\eta)\}.$$

The position variables in spacetime are defined as

$$q_a = \Lambda z_a,$$

and these transform by means of the action on Γ as under $\mathscr{R}_{\text{reg}}$.

The physical energy–momenta are

$$p_a^{\mu} = \Lambda^{\mu}_{j}\eta_a^{j} + \Lambda^{\mu}_{0}[m_a^2 + V_a(z) + \boldsymbol{\eta}_a\cdot\boldsymbol{\eta}_a]^{1/2}.$$

This allows us to satisfy the relations (4.2). Such p_a transform properly as

$$p_a \to Lp_a$$

since the $v_a(z)$ will be chosen to be functions of the differences $z_b - z_c$.

D. The model IV

The equivalence classes are taken with respect to the identity subgroup so that

$$\Sigma = \tilde{\Sigma} = \Sigma^0 \times \mathscr{P}.$$

The variables of Σ are then Λ, Q, and (z), where $\Lambda \in \mathscr{L}$ and Q and (z) are vectors in $\mathbb{R}^4$. The variables of $T^*\Sigma$ are those of Σ and the "momentum" variables $S_{\mu\nu} = -S_{\nu\mu}$, R, (η). Here R_{μ} is conjugate to Q_{μ} and $\eta_{a\mu}$ is conjugate to $z_{a\mu}$ in the usual sense while $S_{\mu\nu}$ is the four-dimensional "angular momentum" conjugate to Λ^{μ}_{ν}. The Poisson brackets are the same as for model III with the addition of

$$\{Q_{\mu},R_{\nu}\} = \delta_{\nu\mu}.$$

The physical position and momentum variables are given by

$$q_a = \Lambda z_a + Q, \quad p_a = \Lambda\eta.$$

The action of the physical (geometrical) Poincaré group is given by $\mathscr{R}_{\text{reg}}$. Under this action q_a and p_a transform as they should:

$$\mathscr{R}_{\text{reg}}(L,b)q_a = Lq_a + b,$$

$$\mathscr{R}_{\text{reg}}(L,b)p_a = Lp_a.$$

Note that z_a and η_a are invariant under $\mathscr{R}_{\text{reg}}$. The mass shell relations (4.2) will hold as a consequence of the definition of M.

V. REDUCED PHASE SPACES AND SECTIONS

To see how the four models fit within the geometrical setup of Sec. II, we will construct the reduced phase space $\mathscr{N}$ for each of the four models following the procedure outlined before. The additional step will be to consider the choice of the constraints $\mathbb{X}$ to build up a family of sections of the bundle $\pi: M \to \mathscr{N}$.

The dimension of the $\mathscr{N}$'s turns out to be always $6N$; this is another reason to call them phase spaces. Another common feature is that the map $\mathbb{K}$ is taken to be invariant under the Poincaré group, which therefore renders M invariant.

A. The model I

The phase space Γ is of dimension $8N$. The $\mathscr{P}$-invariant submanifold M is constructed by introducing the set of N

real-valued functions on Γ,

$$\mathbb{K} = \{K_a\},$$

$$K_a = p_a^\mu p_{a\mu} - m_a^2 - v_a, \quad a = 1,...,N,$$

having the following properties:

— the zero value is in the image of each of them;
— $(dK_1 \wedge \cdots \wedge dK_N)(m) \neq 0 \quad \forall\, m \in M \equiv \mathbb{K}^{-1}(0)$
(i.e., zero is a regular value for $\mathbb{K}$);
— each of them is $\mathscr{P}$-invariant.

$M \equiv \mathbb{K}^{-1}(0)$ is then a submanifold of Γ. Since $\dim \Gamma = 8N$, we have $\dim M = 7N$.

The v_a satisfy the requirement[6–9]

$$\{K_a, K_b\} = 0, \quad a,b = 1,...,N.$$

Therefore, the matrix A vanishes; and

$$d = \dim \mathscr{D} = N.$$

The vector fields X_a which generate $\mathscr{D}$ are then defined through the relations

$$i_{X_a}\omega = dK_a.$$

The dimension of each leaf is N; hence

$$\dim \mathscr{N} = \dim M/\mathscr{D} = 6N.$$

A point in each leaf, depending on a parameter τ, is obtained by imposing the constraints

$$\chi_a = \left(\sum_{b=1}^{N} p_b\right)\cdot(q_{a+1} - q), \quad a = 1,...,N-1,$$

$$\chi_N = \left(\sum_{b=1}^{N} p_b\right)\cdot q_1 - \tau.$$

As shown in the references quoted, they form, together with the K_a a second class system of constraints; therefore,

$$\det|\{K_a, \chi_b\}|_M| \neq 0, \quad a,b = 1,...,N.$$

Since $A_{|M} = 0$, our transversality condition (3.8) coincides with the above.

B. The model II

The phase space Γ has dimension $8N + 8$. The construction of the $\mathscr{P}$-invariant submanifold M is made by introducing $2N + 5$ functions:

$$K_a^{(1)} = P\cdot z_a,$$
$$K_a^{(2)} = P\cdot \eta_a, \qquad a = 1,...,N,$$

$$K_i^{(3)} = \sum_{a=1}^{N} \eta_{ai}, \quad i = 1,2,3,$$

$$K^{(4)} = \sqrt{P^2} - \sum_{a=1}^{N} (m_a^2 - \eta_a^2 + v_a)^{1/2},$$

$$K^{(5)} = \sum_{a=1}^{N} \eta_{a0}.$$

$$\mathbb{K} \equiv (K_a^{(1)}, K_a^{(2)}, K_i^{(3)}, K^{(4)}, K^{(5)}),$$

$$M = K^{-1}(0).$$

The "potentials" v_a are taken to be $\mathscr{P}$-invariant functions of $z_b - z_c$ and η_b. Only $2N + 4$ of them are functionally independent as, for instance, $K^{(5)}$ is a combination of the $K_i^{(3)}$ due to the $K_a^{(2)}$ vanishing; however,

$$(dK_1^{(1)} \wedge \cdots \wedge dK_N^{(1)} \wedge dK_1^{(2)} \wedge \cdots \wedge dK_N^{(2)} \wedge dK_1^{(3)} \wedge \cdots \wedge dK^{(4)} \wedge dK^{(5)})(m) \neq 0$$

for all $m \in M$. We have

$$\operatorname{codim} M = 2N + 4.$$

Again the zero value is regular and M is $\mathscr{P}$-invariant since $\bar{\mathscr{R}}$ either leaves the components of $\mathbb{K}$ invariant or permutes them among themselves.

The $(2N + 5)$-dimensional antisymmetric matrix A, the elements of which are the Poisson brackets of components of $\mathbb{K}$, has the form

$A =$

	$K_a^{(1)}$	$K_a^{(2)}$	$K_i^{(3)}$	$K^{(4)}$	$K^{(5)}$
$K_a^{(1)}$	0	$c \cdots 0$ $0 \cdots c$	$P_1\ P_2\ P_3$ $\cdot\ \cdot\ \cdot$ $\cdot\ \cdot\ \cdot$	x_1 $\cdot$ x_N	P_0 $\cdot$ $\cdot$
$K_a^{(2)}$	$-c \cdots 0$ $0 \cdots -c$	0	$0\ 0\ 0$ $\cdot\ \cdot\ \cdot$ $\cdot\ \cdot\ \cdot$	x_{N+1} $\cdot$ x_{2N}	0 $\cdot$ $\cdot$
$K_i^{(3)}$	$-P_1 \cdots$ $-P_2 \cdots$ $-P_3 \cdots$	$0 \cdots$ $0 \cdots$ $0 \cdots$	$\cdot\ \cdot\ \cdot$ $\cdot\ \cdot$ $\cdot$	0 $\cdot$ $\cdot$	$\cdot$ $\cdot$ $\cdot$
$K^{(4)}$	$-x_1 \cdots -x_N$	$-x_{N+1} \cdots -x_{2N}$	$0 \cdots$	$\cdot\cdot$	$\cdot$
$K^{(5)}$	$-P_0 \cdots$	$0 \cdots$	$\cdots$	$\cdots$	$\cdot\cdot$

where

$$c = P^\mu P_\mu,$$

$$x_a = \{P\cdot z_a, K^{(4)}\},$$

$$x_{a+N} = \{P\cdot \eta_a, K^{(4)}\}.$$

To compute the latter matrix elements and then to evaluate them on M, use is made of the $\mathscr{P}$ invariance of the v_a. This means that both x_a and x_{a+N} are combinations of terms each of which has $P\cdot z_2$ or $P\cdot\eta_a$ as factors; then they vanish on M.

To compute the rank of A on M, we note that its minor A', formed by the first $2N$ rows and $2N$ columns, is nonsingular. We then act on the remaining rows and columns, adding to the elements of a line those of other parallel lines multiplied by suitable constants, to transform A on M into a new matrix $\bar{A}$ having the same rank:

$$\bar{A} = \left[\begin{array}{ccccc:ccc} & & & & & P_1 & 0 & \cdots \\ & & & & & 0 & & \cdot \\ & & & A' & & \cdot & & \cdot \\ & & & & & \cdot & & \cdot \\ \hdashline -P_1 & 0 & \cdots & & & & & \\ 0 & & & & & & & \\ \cdot & & & & & & 0 & \\ \cdot & & & & & & & \\ \cdots & & & & & & & \end{array}\right]$$

the rank of which cannot be $2N+1$ it being antisymmetric. Since $\det A' \neq 0$, we conclude that

$$\operatorname{rank} A = 2N;$$

$$\dim \mathscr{D} = \operatorname{codim} M - \operatorname{rank} A = 4.$$

Since

$$\dim M = 8(N+1) - (2N+4) = 6N+4,$$

we have

$$\dim \mathscr{N} = \dim M/\mathscr{D} = 6N.$$

The set of constraints K' leading to the nonsingular matrix A' is made up of the $K^{(1)}_a$ and $K^{(2)}_a$ $(a = 1,\ldots,N)$.

The four remaining functionally independent K form the K'' set. To these are added four constraints χ to form a second class set. (The explicit form for χ is discussed in Ref. 11.) The transversality condition (3.7) involves the Dirac bracket $\{K'',\chi\}^*$ relative to the K' constraints only. This is satisfied as a previous analysis of this model shows.[11]

C. The model III

The phase space Γ has dimension $8N+12$. Here $2N+5$ functions are introduced to construct the $\mathscr{P}$-invariant submanifold M They are

$$K^{(1)}_\alpha = z^0_\alpha - z^0_{\alpha+1},$$
$$K^{(2)}_\alpha = \eta^0_\alpha - \eta^0_{\alpha+1}, \qquad \alpha = 1,\ldots,N-1,$$

$$K^{(3)}_i = \sum_{a=1}^{N} \eta^i_a, \quad i = 1,\ldots,3,$$

$$K^{(4)}_i = \tfrac{1}{2}\epsilon_{ikl}S^{kl} + \sum_{a=1}^{N} (\mathbf{z}_a \wedge \boldsymbol{\eta}_a)_i,$$

$$K^{(5)} = \eta^0_1 + v(\Delta z) + \sum_{a=1}^{N} (m_a^2 + \boldsymbol{\eta}^2 + v_a)^{1/2}.$$

They are either invariant or transformed into each other under $\mathscr{P}$.

Furthermore the wedge product of their differentials is a $(2N+5)$-form which does not vanish on $M \equiv \mathbb{K}^{-1}(0)$. This is therefore a $\mathscr{P}$-invariant submanifold of dimension $6N+7$.

The antisymmetric matrix A of dimension $(2N+5)$ has the following form:

$$A = \left[\begin{array}{c|ccccc} & K^{(1)} & K^{(2)} & K^{(3)} & K^{(4)} & K^{(5)} \\ \hline K^{(1)} & 0 & A^{(1)} & 0 & 0 & \begin{matrix}1\\0\\\vdots\end{matrix} \\ \hline K^{(2)} & -A^{(1)} & 0 & 0 & 0 & \vdots \\ \hline K^{(3)} & 0 & 0 & 0 & A^{(2)} & \vdots \\ \hline K^{(4)} & 0 & 0 & -A^{(2)} & A^{(3)} & \vdots \\ \hline K^{(5)} & -1\ 0\ \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right]$$

Here

$$A^{(1)} = \begin{vmatrix} 2 & 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 2 & 1 & 0 & \cdot & \cdot & \cdot & \cdot \\ 0 & 1 & 2 & 1 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 & 1 & 2 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 1 & 2 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 & 2 \end{vmatrix},$$

$$A^{(2)} = \begin{vmatrix} 0 & K^{(3)}_3 & -K^{(3)}_2 \\ -K^{(3)}_3 & 0 & K^{(3)}_1 \\ K^{(3)}_2 & -K^{(3)}_1 & 0 \end{vmatrix},$$

and

$$A^{(3)} = \begin{vmatrix} 0 & K^{(4)}_3 & -K^{(4)}_2 \\ -K^{(4)}_3 & 0 & K^{(4)}_1 \\ K^{(4)}_2 & -K^{(4)}_1 & 0 \end{vmatrix}.$$

Evaluated on M,

$$\operatorname{rank} A \leqslant 2(N-1) + 1.$$

A further reduction to

$$\operatorname{rank} A \leqslant 2(N-1)$$

is obtained since an antisymmetric matrix of odd order has vanishing determinant. Direct computation shows $\det A^{(1)} \neq 0$. Therefore,

$$\operatorname{rank} A = 2(N-1).$$

In this case $\mathscr{D}$ has dimension $(2N+5) - 2(N-1) = 7$ and again $\dim \mathscr{N} = 6N$.

The set of K' is formed by the $2(N-1)$ $K^{(1)}_\alpha$ and $K^{(2)}_\alpha$. Imposing only these constraints means restricting the analysis to a subset M' of Γ having dimension $6N+14$. If $i_{M'}$ is the

identification map

$$i_{M'}: M' \to \Gamma,$$

then the original symplectic form ω[17],

$$\omega = \sum_{a=1}^{N} \sum_{\mu=0}^{3} dz_a^\mu \wedge d\eta_{a\mu} + \omega',$$

when pulled back to M' gives

$$i_{M'}^* \omega = \sum_{a=1}^{N} \sum_{i=1}^{3} dz_a^i \wedge d\eta_a^i - N\, dz_{10} \wedge d\eta_{10} + \omega',$$

where ω' pertains to the variables Λ_μ^ν and $S_{\mu\nu}$. This 2-form on M' is seen to be nondegenerate as a consequence of the K' being second class. Introducing new variables to replace z_{10} and η_{10},

$$Q = \sqrt{N} z_{10}, \quad R = \sqrt{N} \eta_{10},$$

we can write

$$i_{M'}^* \omega = \sum_{a=1}^{N} \sum_{i=1}^{3} dz_a^i \wedge d\eta_a^i - dQ \wedge dR + \omega',$$

This is actually the starting symplectic form for the model described in Ref. 12 since the relation between a symplectic form

$$\omega = \tfrac{1}{2} \omega_{\mu\nu}(\xi)\, d\xi^\mu \wedge d\xi^\nu$$

and its associated Poisson brackets

$$\{f, g\} = \omega^{\mu\nu}(\xi) \frac{\partial f}{\partial \xi^\mu} \frac{\partial g}{\partial \xi^\nu}$$

is given by

$$\omega^{\mu\nu} \omega_{\nu\lambda} = \delta_\lambda^\mu.$$

To form the section $\sigma(\mathcal{N} \times \mathbb{R})$ we need to make specific choice of χ as described in Ref. 12.

D. The model IV

The dimension of $T^*\Sigma$ is $8N + 20$ so that a second class system of $2N + 20$ constraints is required to obtain $\dim \mathcal{N} = 6N$. We may choose them to be the following:

$$K_\mu^{(1)} = R_\mu - \sum_{a=1}^{N} p_{a\mu}, \quad \mu,\nu = 1,...,4,$$

$$K_{\mu\nu}^{(2)} = (Q \wedge R)_{\mu\nu} + S_{\mu\nu} - \sum_{a=1}^{N} (q_a \wedge p_a)_{\mu\nu},$$

$$K_a^{(3)} = \eta_a^2 - m_a^2 - v_a, \quad a = 1,...,N,$$

$$\chi_\mu^{(1)} = \sum_{a=1}^{N} \epsilon_a z_{a\mu}, \quad \left(\sum_{a=1}^{N} \epsilon_a = 1,\ \epsilon_a > 0 \right),$$

$$\chi_\alpha^{(2)} = z_{1\alpha} - z_{2\alpha}, \quad \alpha \leqslant 2,$$

$$\chi_\alpha^{(3)} = z_{1\alpha} - z_{3\alpha}, \quad \alpha \leqslant 1,$$

$$\chi^{(4)} = z_{10} - z_{40},$$

$$\chi_\alpha^{(5)} = R \cdot (q_\alpha - q_N), \quad \alpha = 1,...,N-1,$$

$$\chi^{(5)} = R \cdot q_N - \tau.$$

Here we choose v_a in $K_a^{(3)}$ to be functions only of the internal variables z_a and η_a. We choose them to be also invariant under the "Poincaré" group with generators $\Sigma\, \eta_a$, $\Sigma\, (z_a \wedge \eta_a)$ and adjust their functional dependence so that the $K_a^{(3)}$ form a first class set. (This is always possible.[9]) With such a choice $K^{(1)}$, $K^{(2)}$, and $K^{(3)}$ together form a first class set of $(N + 10)$ constraints.

The remaining constraints χ turn this first class set into a second class set. Of these, $\chi^{(1)}$ to $\chi^{(4)}$ are generalizations of those in model III. The functions ϵ_a are functions only of the internal variables (z) and (η) and are thus invariant under the physical Poincaré group. In the free particle limit $v_a \to 0$, they become the "renormalized energies" so that the usual free particle trajectories are recovered as in Ref. 12. The conditions $\chi^{(2)}$ to $\chi^{(4)}$ are designed to fix a Lorentz frame, and thus they are conjugate to $K^{(2)}$. For $N \leqslant 3$ they are clearly inadequate: They must then be replaced by some other "frame fixing" condition. Conditions $\chi^{(5)}$ are the familiar constraints conjugate to $K^{(3)}$.

Since $K^{(1)}$ to $K^{(3)}$ form a first class set $\mathbb{K}$ and the $(N + 10) \times (N + 10)$ matrix of their Poisson brackets with the constraints $\mathbb{X}$ is by construction nondegenerate, it is clear that the $(2N + 20) \times (2N + 20)$ matrix of Poisson brackets is nondegenerate. That is, the constraints $\mathbb{K}$ and $\mathbb{X}$ form a second class set. To be precise, there are degeneracies in these matrices whenever $\chi^{(2)}$ to $\chi^{(4)}$ fail to fix a frame, for instance, when z_1, z_2, and z_3 are parallel. Such situations have to be handled as in Ref. 12.

Thus $M = \mathbb{K}^{-1}(0)$ has dimension $7N + 10$ and the distribution $\mathcal{D}$ has dimension $N + 10$ and is formed by the vector fields X_K. The transversality condition for the σ defined in terms of χ reduces to (3.8) and is satisfied as K and χ form a second class set.

We note the following. The constraints $K^{(1)}$ and $K^{(2)}$ ensure that in the reduced phase space the generators of the physical Poincaré group have the desirable expressions $\Sigma\, p_a$ and $\Sigma\, q_a \wedge p_a$. Also, by virtue of the constraints $\chi^{(1)}$, Q becomes the weighted average $\Sigma\, \epsilon_a q_a$ as in other models.[10,12]

VI. DYNAMICS AS A GATHERING OF MANY INTO A SYSTEM

In the present paper we have started with a grand configuration in which we have a private world to each particle with a 4-vector all to itself and a Lorentz matrix describing the inertial frame. At this stage we had no particles and no motion, no interaction, and no dynamics: We need to generate some *togetherness* and some *self-referral* mechanism to introduce evolution. *Interaction comes from togetherness.*

To form a *system*, this "preparticle" collection has to give up part of its free-wheeling style and subject themselves to some constraints. It is from such constraints that the dynamical system specification and even the notion of dynamical evolution and the evolution parameter emerge.

In this paper we show many alternate patterns to the same goal and how the intermediate stage formulations appear drastically different. We also see in the course of time that not all constraints are on the same footing. Some are gauge constraints which change only the language of description; but some are essential constraints. *Changing the latter means changing the physical system.*

It is fairly straightforward to make choice of the constraints so that the world line condition is satisfied thus fulfilling one of the elementary requirements on relativistic interacting systems. But it was essential to go beyond the

ten-parameter descriptions to the generalized *11-parameter form of Dirac's relativistic dynamics.*

In all this discussion the question of separability for systems with more than two particles has not been answered. We have addressed ourselves to this question elsewhere.[13]

In conclusion, we wish to stress the unifying power of geometry allowing us to view different models for relativistic interacting particles from a common perspective. The emphasis on the role of geometry in description of nature goes back to Plato, and this point of view has been enriched over the centuries by many illustrious scholars.[14] We hope that our work is in keeping with this tradition.

[1]We refer to R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Addison-Wesley, Reading, Mass., 1978) for the theory and notation of the calculus on manifolds.

[2]G. Marmo, E. J. Saletan, and A. Simoni, Nuovo Cimento B **50**, 1 (1979).

[3]Peter G. Bergman and Arthur Komar, "The Hamiltonian in Relativistic Systems of Interacting Particles," Syracuse University, 1980; F. Rohrlich, Phys. Rev. D **25**, 2576 (1982).

[4]Ref. 1, p. 116.

[5]Ref. 1, p. 180.

[6]Ph. Droz-Vincent, Lett. Nuovo Cimento **1**, 839 (1969); **1**, 206 (1973); Phys. Scripta **2**, 129 (1970); Rep. Math. Phys. **8**, 79 (1975); Ann. Inst. H. Poincaré **27**, 407 (1977); Phys. Rev. D **19**, 702 (1979).

[7]I. T. Todorov, "Dynamics of Relativistic Point Particles as a Problem with Constraints," Commun. of the JINR, EZ-10125, Dubna, 1976.

[8]A. Komar, Phys. Rev. D **18**, 1881, 1887, 3017 (1978); **19**, 2908 (1979).

[9]E. C. G. Sudarshan, N. Mukunda, and J. N. Goldberg, Phys. Rev. D **23**, 2218 (1981).

[10]F. Rohrlich, Ann. Phys. **117**, 292 (1979); Physica A **96**, 290 (1979); M. J. King and F. Rohrlich, Ann. Phys. **130**, 350 (1980).

[11]N. Mukunda and E. C. G. Sudarshan, Phys. Rev. D **23**, 2210 (1981).

[12]A. P. Balachandran, G. Marmo, N. Mukunda, J. S. Nilsson, A. Simoni, E. C. G. Sudarshan, and F. Zaccaria, Nuovo Cimento A **67**, 121 (1982).

[13]A. P. Balachandran, D. Dominici, G. Marmo, N. Mukunda, J. S. Nilsson, J. Samuel, E. C. G. Sudarshan, and F. Zaccaria, Phys. Rev. D **26**, 3492 (1982).

[14]G. Galilei, Il Saggiatore, **VI**, 232 (1623); S. Lie, "Zur Theorie der Transformationsgruppen," Christ. Forh. Aar. 1888 Nr. 13, Ges. Abh. BNdv, XXIII (especially pp. 554–7); S. Lie and F. Engle, *Theorie der Transformationsgruppen*, (Teubner, Liepzig, 1888–1893), Vols. I–III (in particular Vol. II); E. Cartan, *Leçons sur les invariants integraux* (Hermann, Paris, 1922); C. Caratheodory, *Calculus of Variations and Partial Differential Equation of the First Order (Part I)* (Holden Day, San Francisco, 1965); A. Lichenerowicz, J. Differential Geom. **12**, 253 (1977); R. Herman, *Interdisciplinary Mathematics* (Math. Sci. Press, Brookline, Mass., 1975, 1977), Vol. 14, Chap. 14, and Vol. 15, p. 52.

A smooth transonic flow in the plane

P. D. Smith[a)]
Institute for Advanced Study, Princeton, New Jersey and Johns Hopkins University, Baltimore, Maryland 21218

(Received 18 May 1982; accepted for publication 7 January 1983)

The implicit function theorem is used to study a symmetric exterior problem for the gas dynamics equation—an equation of mixed type. The existence of families of smooth C^1 solutions is demonstrated. These solutions are families of smooth transonic flows in the plane and are of applied interest. Some of these results have appeared in the literature with an incorrect derivation using the Hodograph mapping. This mapping is not invertible in the transonic case. The methods of this paper do not use the Hodograph mapping and extend to general (e.g., plasma) flows.

PACS numbers: 47.40.Hg, 02.30. + g

INTRODUCTION

Recently L. M. and R. J. Sibner have constructed a family of smooth transonic flows on a symmetric torus.[1] Smooth transonic flows are interesting because of the transonic flow controversy (see Bers[2]), but a physicist might object that flows constrained to a torus are not physical. In this paper the method of Ref. 1 is extended to construct families of smooth transonic flows in an exterior plane domain, showing that the above objection is unfounded.

The extension of their method is necessary because of technical difficulties: In a limiting case of our plane flow, certain derivatives, which are always finite in toroidal flow, become infinite. Also our flow domain is not compact. Together, these two facts require the modification of certain Arzela–Ascoli arguments in Ref. 1. The new arguments use Dini's theorem on the convergence of monotone function sequences instead of the Arzela–Ascoli theorem.

In the toroidal flows shock solutions may also occur. In plane flows, when the polytropic constant $\gamma = 3$, we show that shocks do not occur. Our proof uses the Prandle–Rankine–Hugonant relations for shocks in a polytropic gas. The author conjectures that shocks do not exist for any value of $\gamma > 1$.

Our construction of smooth transonic flows is interesting because it never uses the Hodograph mapping, a mapping which may not be invertible in transonic flow. See Bers,[2] for a discussion of the inapplicability of the Hodograph method in transonic flow, and Courant[3] for the Hodograph approach.

1. DESCRIPTION OF THE PROBLEM

We seek an irrotational, stationary polytropic flow in the exterior of the unit circle considered as a domain in the Euclidean plane. This flow is assumed to have a constant angular speed, i.e., to be independent of the polar angle. We show, directly from the defining differential equation, that there are three flows of this type: purely rotational vortex flow, purely radial source flow with constant mass flow through the circle, and spiral flow with constant mass flow through the circle. The most interesting flow is the spiral flow because this case includes a family of smooth transonic flows.

Our results follow from a complete analysis of the mass flow–circulation problem below:

The mass flow–circulation problem

Consider the exterior of the unit circle as a domain in the Euclidean plane: Show that, in this domain, there exists an irrotational, stationary, polytropic flow that is independent of the polar angle, that has prescribed circulation about the circle, and that has prescribed radial mass transport through the circle.

Remark: The data for the mass flow–circulation problem must lie in certain ranges determined later. The reader will find a complete statement of the results in Sec. 4.

2. THE DIFFERENTIAL EQUATION

We now describe the model of polytropic flow used in this discussion. This model was developed by Sibner and Sibner[4,5] to describe stationary irrotational polytropic flow on a Riemannian manifold.

In this model a flow is described by its velocity field given as a differential 1-form ω that satisfies the equations below:

$$d\omega = 0, \tag{2.1a}$$

$$\delta\rho(Q(\omega))\omega = 0, \tag{2.1b}$$

where $Q(\omega) = g^{ij}\omega_i\omega_j$ is the square speed and $\rho = (1 - \frac{1}{2}(\gamma - 1)Q(\omega))^{1/(\gamma-1)}\gamma > 1$ is the polytropic density function (see Bers[2]). We require that ρ be nonnegative which forces $0 \leqslant Q(\omega) \leqslant 2/(\gamma - 1)$.

Remark: Physically, Eq. (2.1a) is the irrotationality of flow, and Eq. (2.1b) is the conservation of mass. These equations are a mixed quasilinear system. When $0 \leqslant Q(\omega) < 2/(\gamma + 1)$, this system is elliptic and the flow is subsonic; hence, $2/(\gamma + 1)$ is the *square sonic speed*; when $Q(\omega) = 2/(\gamma + 1)$, the system is parabolic; when $Q(\omega) > 2/(\gamma + 1)$, the system is hyperbolic, and the flow is said to be supersonic. This system is the prolongation of the gas dynamics equation to the cotangent bundle of a Riemannian manifold. See Ref. 4 for details.

a) Supported in part by NSF Grant MCS 77-18723 A04.

 0022-2488/84/010177-08$02.50

For flows exterior to the unit circle in the Euclidean plane, it is convenient to use polar coordinates R and θ. With these coordinates $g_{11} = 1$, $g_{22} = R^2$, and $g_{12} = g_{21} = 0$.

Equation (2.1a) reduces to

$$\text{(A)}\quad \alpha_\theta = B_R, \quad \text{where } \omega = \alpha\, dR + \beta\, d\theta. \tag{2.2}$$

Equation (2.1b) reduces to

$$\text{(B)}\quad \frac{\partial}{\partial R}[R\rho\alpha] + \frac{\partial}{\partial\theta}\left[\frac{1}{R}\rho\beta\right] = 0.$$

In the next section, we show that solutions of (A) and (B), which are independent of θ, satisfy a nonlinear algebraic equation. Compare Ref. 1.

3. THE MASS-FLOW RELATION

From now on we consider only flows in the exterior of the unit circle, which are independent of the polar angle. These flows have constant angular velocity (hence β is constant) and are radial, rotational, or spiral flows.

We show that such flows satisfy a conservation law given by a nonlinear algebraic equation—the mass flow relation.

Consider Eqs. (A) and (B). In combination they tell us that any solution $\omega = \alpha\, dR + \beta\, d\theta$, which is independent of θ, must satisfy

$$B_R = \alpha_\theta \frac{\partial}{\partial R}(R\rho\alpha) + \frac{\partial}{\partial\theta}\left[\frac{1}{R}\rho\beta\right] = 0. \tag{3.1}$$

Since β is constant, this implies that $\alpha = \alpha(R)$ is a function only of R. Recall that, in this geometry, $g_{11} = 1$, $g_{22} = R^2$, $g_{12} = g_{21} = 0$. Since $Q(\omega) = g^{ij}\omega_i\omega_j = \alpha^2 + \beta^2/R^2$, we see that Q is independent of θ. Moreover, because $\rho^2 = (1 - \frac{1}{2}(\gamma - 1)Q)2/(\gamma - 1)$, we also see that the density ρ is independent of θ.

Thus the partial differential equation above becomes the nonlinear algebraic equation

$$R^2\rho^2\alpha^2 = K, \quad \text{for some nonnegative const } K, \tag{3.2}$$

and, since $Q = \alpha^2 + \beta^2/R^2$ with β constant, we obtain the mass flow relation (MF) $R^2\rho^2(Q - \beta^2/R^2) = K$ with β constant and K a nonnegative constant.

The mass flow relation has physical as well as mathematical importance. Physically, it says that the mass flow through the circle is zero in rotational vortex flow and constant in both radial and spiral flow. Mathematically the mass flow relation is (for fixed values of K and β) a relation for $Q(\omega)$ as a function of the radius R and this relation determines α from $Q = \alpha^2 + \beta^2/R^2$.

In any case, a flow satisfying the mass flow relation is presented by two parameters: β (which determines the circulation $C = 2\pi\beta$) and K (which determines the radial mass flow).

4. A DESCRIPTION OF PLANE FLOWS EXTERIOR TO THE UNIT CIRCLE—A LIST OF THE RESULTS

This section describes all the solutions to the mass flow–circulation problem. These solutions are symmetric, stationary flows, with fixed circulation $2\pi\beta$: Purely radial (source) flow, purely rotational (vortex) flow and spiral flow. Both vortex and spiral flow include solution families that are everywhere smooth and transonic.

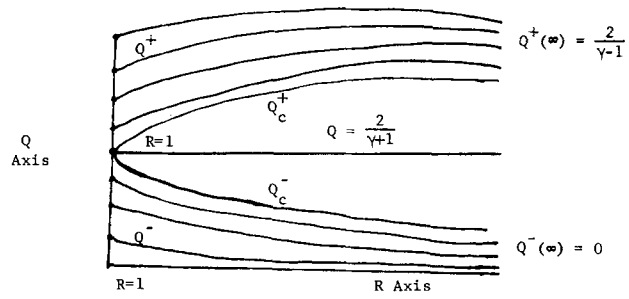

FIG. 1. Radial flow.

Purely radial (source) flow

Here β is zero. The speed is given by $Q = \alpha^2(R)$ and the mass flow relation reduces to $K = R^2\rho(\alpha^2)\alpha^2$.

The constant K parametrizes the solutions. If K is too large, there is no flow at all. Any smaller value K corresponds to two flows. The first flow is everywhere supersonic with a limiting square speed of $2/(\gamma - 1)$ at infinity. We denote this flow by Q^+. The second flow is everywhere subsonic with a limiting speed of zero at infinity. We denote this flow by Q^-. See Fig. 1.

Since $\rho^2(R) = (1 - \frac{1}{2}(\gamma - 1)Q)2(\gamma - 1)$, we see that if $Q = 2/(\gamma - 1)$, the density vanishes. Thus $Q = 2/(\gamma - 1)$ is the square vacuum (or cavitation) speed. Since $Q^+ = 2/(\gamma - 1)$ at infinity and $K = R^2\rho(\alpha^2)\alpha^2$, both Q^+ and Q^- have no mass flow at infinity.

When the initial speed is sonic, i.e., $Q(1) = 2/(\gamma + 1)$, corresponding to the largest value of K for which there is flow, the two flows bifurcate from $R = 1$. See Fig. 1. We call such flow critical flow and $K = K_c$ critical mass flow. Both critical solutions Q_c^+ and Q_c^- are everywhere smooth (C^1), except when $R = 1$, where $d[Q^+]/dR$ is infinite.

Remark: Shock flow—that is, flow that starts on the Q^+ curve and finishes on the Q^- curve labeled by the same K (see Figs. 2 and 3)—might occur. However, we show in Sec. II that, when $\gamma = 3$, shocks never occur.

Purely rotational (vortex) flow

This is a trivial case. This flow is a vortex with constant angular speed parametrized by β. The streamlines are circles, concentric and exterior to the unit circle. There is no radial mass transport ($K = 0$).

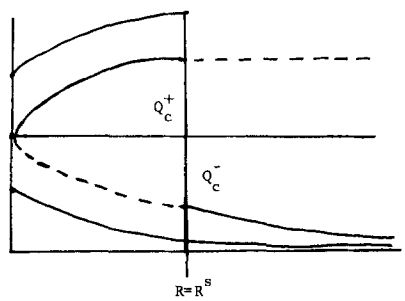

FIG. 2. Shock in critical flow.

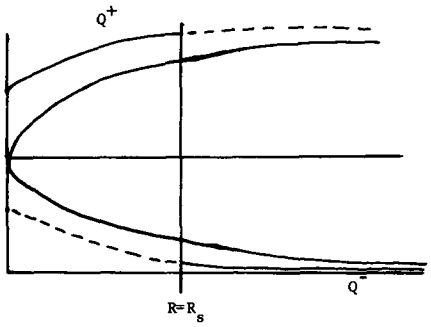

FIG. 3. Shock in noncritical flow.

Since in this case $Q = \beta^2/R^2$, the flow is everywhere subsonic when $\beta > \sqrt{2/(\gamma+1)}$ and otherwise smooth transonic. The limiting speed is zero.

Spiral flow

This is the most interesting flow. It combines the bifurcation feature of radial flow with the smooth transonic flow feature of spiral flow.

In this case the flows are parametrized by K and β. Physically this says the flows are parametrized by mass flow (K) and circulation C ($C = 2\pi\beta$).

Just as in the case of purely radial (source) flow, in spiral flow we have two solutions corresponding to each value of K, up to a critical value K_c of K, and then no solutions if $K > K_c$.

However, in spiral flow the critical mass flow constant K_c depends on β. It is a consequence of this dependence that spiral flows have families of everywhere smooth transonic flows.

To see this, we must think in terms of bifurcation points. In purely radial flow, bifurcation occurred at $R = 1$ when $K = K_c$ and $Q_c^{\pm}(1) = 2/(\gamma+1)$, the sonic speed. But in spiral flow bifurcation occurs at $R = 1$ when $K = K_c$, and K depends on β. Because of the mass flow relation K_c determines $Q_c^{\pm}(1)$ [just let $R = 1$ in (MF)] and this means that bifurcation occurs when $R = 1$ and

$$Q(1) = \hat{Q}(1) = [2/(\gamma+1)](1+\beta^2). \tag{4.1}$$

The new bifurcation point is called the critical speed. The critical speed is generally larger than the sonic speed (if $\beta > 0$) and replaces the sonic speed as a bifurcation point in spiral flow. See Fig. 4.

Since k determines $Q(1)$ from the mass flow relation, we could also parametrize the $Q^{\pm}$ flows by β and $Q^{\pm}(1)$.

In spiral flow shocks might occur (although the author conjectures that they do not). When $\gamma = 3$, we show in Sec. 11 that shocks never occur.

However, we do have families of smooth transonic flows. To see this, consider the Q_- flows, with $\beta^2 < 2/(\gamma-1)$, and with $2/(\gamma+1) < Q^-(1) \leqslant \hat{Q}(1)$. These flows are a family of everywhere smooth transonic spiral flows with zero limiting speed at infinity.

5. TWO EQUIVALENT PROBLEMS

We reformulate the mass flow–circulation problem as an initial value problem.

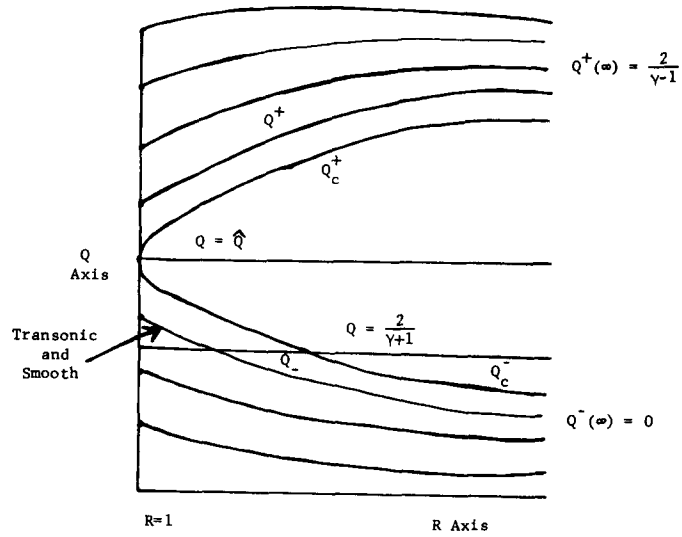

FIG. 4. Spiral flow.

Consider the following two problems:

Problem I: Mass flow–circulation problem: Find C^1 functions $Q^{\pm}(R)$, satisfying $R^2\rho^2(Q - \beta^2/R^2) = K$, where $R \geqslant 1$, $K > 0$, $0 < \beta^2 < 2/(\gamma-1)$, K, and β are prescribed constants, and where $\rho^2 = (1 - \frac{1}{2}(\gamma-1)Q)2/(\gamma-1)$.

Probelm II: Initial value problem: Find C^1 solutions $Q^{\pm}(R)$ with prescribed β, $0 < \beta^2 < 2(\gamma-1)$, with prescribed initial data $Q^+(1) > \hat{Q}(1)$ or $Q^-(1) < \hat{Q}(1)$ satisfying

$$R^2\rho^2(Q - \beta^2/R^2) = K = \rho^2[Q^{\pm}(1) - \beta^2]. \tag{5.1}$$

These problems are equivalent. More precisely, we have:

Proposition 5.1: Solutions of I satisfy II and vice-versa provided that $K = \rho^2(Q^{\pm}(1))(Q^{\pm}(1) - \beta^2)$ in I.

Which follows from:

Proposition 5.2: Solutions of either I or II satisfy the algebraic mass flow relation

$$(MF) \quad R^2\rho^2(Q^{\pm})(Q^{\pm} - \beta^2/R^2) = K,$$

and C^1 solutions to the relation (MF) satisfy problems I and II for noncritical K and noncritical initial value $Q^{\pm}(1)$ such that $\rho^2(Q^{\pm}(1))(Q^{\pm}(1) - \beta^2) = K$.

Proof: The relation $\rho^2(Q^{\pm}(1))(Q^{\pm}(1) - \beta^2) = K$ is simply the mass flow relation (MF) when $R = 1$. The two problems are equivalent since Problem II is a Problem I with this relation used to replace K by $Q^{\pm}(1)$.

6. LOCAL EXISTENCE THEORY

We seek a local solution $Q = Q(R)$ of the initial value problem:

$$R^2(Q - \beta^2/R^2)\rho^2 = K \quad \text{on } [R_1, R_2] \subset [1, \infty),$$

K a positive constant, $Q(R_1) = Q_1$, and $\beta^2/R_1^2 < Q_1 < 2/(\gamma-1)$.

Definition: $\hat{Q}(R) = [2/(\gamma+1)](1 + \beta^2/R^2)$ is called the critical curve. A solution that lies above $\hat{Q}$ is called supercritical and is denoted by Q^+. A solution that lies below $\hat{Q}$ is called subcritical and is denoted by Q^-.

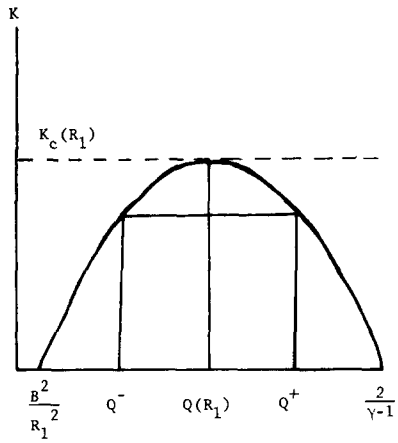

FIG. 5. Dependence of K on Q.

Theorem 6.1: There exists a critical value of K, $K = K_c$, such that for given R_1, Q_1, and K, where $0 \leqslant K < K_c(R_1)$, there is an interval $[R_1, R_2]$ in which there exist two local C^1 solutions Q^+ and Q^- of the initial value problem above.

Moreover, Q^+ and Q^- satisfy:

(i) $K = R^2(Q^\pm - \beta^2/R^2)[1 - \frac{1}{2}(\gamma-1)Q^\pm]2/(\gamma-1)$;

(ii) $\beta^2/R^2 < Q^- < \hat{Q} < Q^+ < 2/(\gamma-1)$;

(iii) (a) Q^+ (resp. Q^-) is a monotone increasing (decreasing) function of K, (b) K increases monotonically as a function of increasing Q^-, (c) K decreases monotonically as a function of increasing Q^+; (6.1)

(iv) $Q^- \nearrow \hat{Q}$ and $Q^+ \searrow \hat{Q}$ as $K \to K_c$, see Fig. 5;

(v) $K_c(r_1) \leqslant K(r_2)$ if $r_1 \leqslant r_2$;

(vi) if $Q^\pm(R)$ satisfying $Q^\pm(R_1) = Q_1$ can be continued to $R = R_2$ then $Q(R_2)$ is an increasing function of $Q(R_1)$.

Proof: These results are proved for toroidal flow in Ref. 1, p. 372. Our theorem follows identically with the substitution $R = f$. The proofs in Ref. 4 follow by application of the implicit function theorem to the function

$$F(R,Q) = R^2(Q - \beta^2/R^2)\rho^2 - K;$$

thus,

$$E_Q(R_1,Q) = R_1^2[1 - \tfrac{1}{2}(\gamma-1)Q_1](3-\gamma)/(\gamma-1) \quad (6.2)$$

and

$$\frac{dQ}{dR} = \frac{-F_R}{F_Q} = \frac{-8}{\gamma+1}\,\frac{1}{R}\left\{\frac{Q[1 - \frac{1}{2}(\gamma-1)Q]}{\hat{Q} - Q}\right\}.$$

The conclusions follow from arithmetic consideration of these expressions. Note that dQ/dR is infinite if $Q = \hat{Q}$. This is the analytic meaning of $\hat{Q}$.

Remark: (vi) tells us that the local solutions are monotone increasing as functions of their initial values. In other words, if $Q_a^+(1) \geqslant Q_b^+(1)$, then $Q_a^+(r) \geqslant Q_b^+(r)$ for any $r \geqslant 1$. When we have found that global solutions of the mass flow–circulation problem (vi) will also tell us that these solutions $Q^\pm$ are also monotone increasing functions of their initial data.

The formula for dQ/dR above tells us that each solution Q^+ is monotone increasing and that each solution Q^- is monotone decreasing as functions of the radius R. The same facts hold for global solutions of the mass flow–circulation problem.

The considerations of this remark are an important tool in the proof of convergence to the critical solutions of Sec. 9.

7. SOME LOCAL LEMMAS

We state some necessary technical lemmas that are almost identical to the corresponding technical lemmas of Ref. 1. Indeed the proofs in Ref. 1 apply to our case with just a change in notation and domain. We need these technical lemmas in the global existence theory.

Recall Problems I and II of Sec. 5. Let $R \geqslant 1$ as usual.

Proposition 7.1: Let $Q^\pm(R)$ be a solution of the initial value problem II on some interval $[R_{\min}, R_{\max}]$ for which $\varphi(R) = [Q(R) - Q^\pm(R)]^2$ is positive. Then $\varphi(R)$ has a unique minimum at $R = R_{\min}$.

The proof of this proposition follows from this lemma:

Lemma 7.1: (a) The function $(\hat{Q} - Q^\pm)^2$ satisfies the differential equation

$$\frac{d}{dR}[\hat{Q} - Q^\pm]^2 = \frac{8}{R} g(R, Q^\pm(R)), \quad (7.1)$$

where

$$g(R,Q^\pm(R)) = [1/(\gamma+1)]\{Q^\pm(R)[1 - \tfrac{1}{2}(\gamma-1)Q^\pm(R)] + (\beta^2/R^2)[Q^\pm(R) - \hat{Q}(R)]\}. \quad (7.2)$$

(b) If $[\hat{Q}(R) - Q(R)] > 0$ on an interval $[R_{\min}, R_{\max}]$, then $g(R,Q(R)) > C(K) > 0$, where

$$C(K) = \frac{2}{(\gamma+1)^2}\left(1 + \frac{\beta^2}{R_{\max}^2}\right) \times \left[\min \frac{K}{R_{\max}^2}, \left(\frac{\gamma-1}{2}\,\frac{K}{R_{\max}^2}\right)^{2/(\gamma-1)}\right] \quad (7.3)$$

is a monotone increasing function of the mass flow constant K.

Proof: With the substitution $R = f$ and by replacing the torus with the interval $[R_{\min}, R_{\max}]$ as domain, the proof is identical to that of Lemma 5.2 of Ref. 4.

Proof of Proposition 7.1: In Ref. 1 the second derivative test was used, but here it does not apply since the minimum now occurs on the left end point.

We see this because the derivative of the function in question is positive and this function is continuous on a compact set. Here, the minimum occurs on the left end point.

Corollary 7.1: If Q^- is a solution of Problem II in some interval $[R_{\min}, R_{\max}]$ in which $Q^+(R) - Q^-(R) > 0$, then

$$Q^+(R) - \hat{Q}(R) > Q^+(R_{\min}) - \hat{Q}(R_{\min}) > 0. \quad (7.4)$$

We also have:

Lemma 7.2: Let $N = [R_{\min}, R_{\max}]$ be an interval contained in $[1, \infty)$. There exists a unique C^1 subcritical (resp. critical) solution Q^- (resp. Q^+) of Problem II (equivalently of Problem I) with noncritical data on N.

Proof: This follows just like Theorem 5.1 in Ref. 1 with the substitution of N for the torus and R for f.

Remark: The above supercritical solutions Q^+ are monotone increasing, and the above subcritical solutions Q^- are monotone decreasing because of the differential equation above: The sign of the derivative of $Q^\pm$ depends on whether $\hat{Q} - Q^\pm$ is positive or negative.

8. GLOBAL EXISTENCE THEORY FOR NONCRITICAL FLOWS

We establish the global existence and uniqueness of noncritical flow solutions of the mass flow–circulation problem. The method of proof is somewhat different from the method of Ref. 1 because of complications due to the noncompact nature of our domain.

We now state and prove the main theorem of this section.

Theorem 8.1: Let $0<\beta^2<2/(\gamma-1)$ and $0\leqslant K\leqslant K_c$, where K_c is the critical mass flow constant, corresponding to β. There exists a unique C supercritical (resp. subcritical) flow Q^+ (resp. Q^-) with mass flow constant K and circulation $2\pi\beta$ about the unit circle.

At no loss of generality, we carry out the proof in the supercritical Q^+ case.

Proof: First, we show uniqueness.

Suppose that we have two solutions. We show that the set S on which they are equal is nonempty and open and closed in the relative topology that S inherits as a subset of $[1,\infty)$. Then, because $[1,\infty)$ is connected, S is $[1,\infty)$.

We start by showing that S is nonempty. Any solution of the mass flow–circulation problem (Problem I) is also a solution of the initial value problem (Problem II) with initial data $Q(1)$ determined by K. Since both solutions have the same mass flow constant K, they have the same initial value $Q(1)$. Thus they agree at $R=1$, and S is nonempty.

We now show that S is relatively closed and open. Any solution of the mass flow–circulation problem must satisfy the mass flow condition (MF) at every point of $[1,\infty)$. This condition is a continuous algebraic functional relation on Q. Thus S is closed. The implicit function theorem shows that S is open.

Thus S is $[1,\infty)$, and the two solutions are equal—uniqueness is proved.

Remark: We used the noncritical nature of K, when we envoked the implicit function theorem. Because K is noncritical dF/dQ is nonzero and noninfinite. See Sec. 6 for the formula giving DF/dQ.

We now show existence by construction. Let $N=[1,R]$ and $M=[1,\tilde{R}]$ be subintervals of $[1,\infty)$ such that $N\subset M$. The local existence theorem (Theorem 6.1) gives us C^1 solutions ${Q^+}_N$ and ${Q^+}_M$ of Problems I and II on N and M. We show that ${Q^+}_M$ is the unique continuous extension of Q^+ to M that is a solution of Problems I and II on M.

Let S be the set of points where ${Q^+}_N$ and ${Q^+}_M$ agree. Clearly $S\subset N$. We show that S is nonempty, and also closed and open in the relative topology that N inherits as a subset of M. The proof is very similar to the uniqueness proof above.

S is nonempty because ${Q^+}_N$ and ${Q^+}_M$ have the same initial value at $R=1$ as solutions of Problem II. S is closed because ${Q^+}_N$ and ${Q^+}_M$ both satisfy the algebraic mass flow relation (MF), which is a continuous algebraic functional relation on Q. Finally, S is open by the implicit function theorem since K is noncritical and $0<dF/dQ<\infty$.

Thus $S=N$ and ${Q^+}_M$ is the unique continuous extension of ${Q^+}_N$ as a solution of Problem I.

Now let $\tilde{R}\to\infty$. The local solution ${Q^+}_M$ tends to a global solution Q^+ of Problem I. QED

The argument for Q^- is identical.

Corollary 8.1: Q^+ is monotone increasing as a function of R with limiting speed $2/(\gamma-1)$ at infinity. Q^- is monotone decreasing as a function of R with limiting speed zero at infinity.

Proof: The monotonicity follows from the sign of $dQ^\pm/dr$ (see Sec. 6), now that we know that Q^+ and Q^- exist. We see the limiting speed behavior by looking at the algebraic mass flow relation (MF). Both Q^+ and Q^- must satisfy (MF). Let $R\to\infty$. Because K/R^2 then goes to zero, either $\rho(Q_\infty)$ or Q_∞ must vanish.

Since Q^+ is monotone increasing, this forces the limiting square speed to be the square cavatation speed $2/(\gamma-1)$, similarly, because Q^- is monotone decreasing its limiting speed at infinity must be zero. QED

We also have two more corollaries.

Corollary 8.2: The global C^1 solutions $Q^\pm$ satisfy the differential equation

$$\frac{d}{dR}[\hat{Q}-Q^\pm]^2=\frac{8}{R}g(R,Q(R)), \tag{8.1}$$

where g is the same as it was in Lemma 7.1.

Proof: $Q^\pm$ exist. The conclusion of this corollary is a local condition on $Q^\pm$ which was proved in Lemma 7.1. QED

Corollary 8.3: The global C^1 solutions $Q^\pm$ satisfy the integral equation:

$$Q^\pm=\hat{Q}\pm\left\{[\hat{Q}(1)-Q(1)]^2+\int_1^R(8/t)g(t,Q(t))\,dt\right\}^{1/2}. \tag{8.2}$$

Proof: Integrate the differential equation of the previous corollary. QED

9. CONVERGENCE TO THE CRITICAL SOLUTIONS

We show that as K approaches its critical value K_c, the solutions Q^+ and Q^- approach limiting functions Q_c^+ and Q_c^- that solve the mass flow–circulation problem when $K=K_c$.

Slightly abusing terminology, we call Q_c^+ and Q_c^- subcritical. The two critical solutions Q_c^+ and Q_c^- satisfy the critical mass flow relation:

$$\text{(MFC)}\quad K_c=R^2(Q_c^\pm-\beta^2/R^2)\rho^2(Q_c^\pm). \tag{9.1}$$

Q_c^+ is monotone increasing as a function of R with a limiting square speed $2(\gamma-1)$ at infinity, Q_c^- is monotone decreasing as a function of R with a limiting square speed of zero at infinity, and Q_c^+ and Q_c^- bifurcate at $R=1$ with vertical slope. See Fig. 4.

Although the critical solutions $Q_c^\pm$ satisfy the algebraic relation (MFC), we cannot prove local existence using the implicit function theorem alone because the required derivative is infinite at $R=1$ (compare Sec. 6). We prove local and global existence and uniqueness using convergence arguments.

These convergence arguments construct $Q_c^\pm$ as the limit of noncritical solutions $Q^\pm$ as K approaches its critical

value $K = K_c$. The convergence arguments are more delicate than one might at first suspect because our domain is *not* compact, and thus we do the convergence arguments in careful detail.

We now state and prove the theorem.

Theorem 9.1: There exist unique solutions Q_c^+ and Q_c^- to the mass flow–circulation problem with critical mass flow $K = K_c$.

These solutions are C^1 when $R > 1$ and satisfy:

(a) Q_c^+ is monotone increasing as a function of R with limiting square speed $2/(\gamma - 1)$ at infinity.

(b) Q_c^- is monotone decreasing as a function of R with zero limiting speed at infinity.

(c) $Q_c^+ > \hat{Q}$ when $R > 1$ (we say then that Q^+ is supercritical) and $Q_c^- < \hat{Q}$ when $R > 1$ (we say then that Q^- is subcritical).

(d) Q_c^+ and Q_c^- bifurcate from $R = 1$ with infinite slope at $R = 1$.

(e) Let Q_K^+ denote solutions of the mass flow–circulation problem with noncritical mass flow K. Then as K approaches the critical value $K = K_c$, $Q_K^\pm$ approach $Q_c^\pm$ pointwise. In fact, $Q_K^+ \searrow Q_c^+$ and $Q_K^- \nearrow Q_c^-$ uniformly on any compact subinterval of $[1, \infty)$.

(f) Q_c^+ and Q_c^- also solve the initial value problem (Problem II) with *critical* initial value $Q_c(1) = \hat{Q}_c(1)$.

(g) Let $Q_{Q_1}^\pm$ denote solutions of the initial value problem with noncritical initial value $Q(1) = Q_1$. Then, as the initial values Q_1 approach the critical initial value $Q_c(1) = \hat{Q}(1)$, $Q_{Q_1}^\pm$ approach $Q_c^\pm$. In fact, $Q_{Q_1}^+ \searrow Q_c^+$ and $Q_{Q_1}^- \nearrow Q_c^-$ and the convergence is uniform on any compact subinterval of $[1, \infty)$.

(h) The critical solutions satisfy the critical mass flow relation

$$K_c = R^2(Q_c^\pm - \beta^2/R^2)\rho^2(Q_c^\pm)). \tag{9.2}$$

(i) They satisfy the integral equation

$$Q_c^\pm(R) = \hat{Q}(R) \pm \left\{ [\hat{Q}_c(1) - Q_c^\pm(1)]^2 + \int_1^R g(t, Q^{\pm c}(t))\, dt \right\}^{1/2}. \tag{9.3}$$

(j) They have derivatives for $R > 1$ given by

$$\frac{dQ_c^\pm}{dR} = \frac{d\hat{Q}}{dR} - \frac{1}{2}\frac{8}{R} g(RQ_c^\pm(R)) \times \left\{ [\hat{Q}_c(1) - Q_c^\pm(1)]^2 + \int_1^R (8/t) g(t, Q^\pm(t))\, dt \right\}^{-1/2}. \tag{9.4}$$

Proof: We prove the theorem for Q^+. The proof for Q^- is identical.

Proof of Uniqueness: Let $Q_{c,1}^+$ and $Q_{c,2}^+$ be C^1 (when $R > 1$) global solutions of the mass flow–circulation problem with critical mass flow constant $K = K_c$. Then, $Q_{c,1}^+$ and $Q_{c,2}^+$ satisfy the critical mass flow relation

$$\text{(MFC)} \quad [Q_c^+(R) - \beta^2/R]\rho^2(Q_c^+)R^2 = K_c.$$

Let S be the set of points where $Q_{c,1}^+ = Q_{c,2}^+$. Then because these solutions have the same initial value, S is nonempty (i.e., $K = 1 \in S$). Because (MFC) is algebraic and thus continuous as a function of Q_c^+, the set S is closed. When $R > 1$, the initial value theorem implies the local solvability of the relation (MFC) and thus S is relatively open in $[1, \infty)$. Therefore, S is all of $[1, \infty)$ and $Q_{c,1}^+ = Q_{c,2}^+$. Now, we prove the existence of a global critical solution that is C^1 when $R > 1$.

We now construct the supercritical solution Q_c^+.

Consider a sequence $(Q_{K_n}^+)$ of solutions to Problem I corresponding to a sequence of noncritical mass flow values K_n (for β fixed!) converging to the critical mass flow value K_c. Since K determines $Q(1)$ from the mass flow relation at $R = 1$, the remark at the end of Sec. 6 (and Theorem 6.1) imply that at each point x of $[1, \infty)$ the sequence $(Q_{K_n}^+)$ is monotone decreasing and bounded from below (by zero); hence it has a limit point $Q_c^+(x)$. Thus the function sequence $(Q_{K_n}^+)$ converges pointwise to a limit function Q_c^+. Moreover, a standard interlacing sequence argument shows that the limit function Q_c^+ is independent of the choice of the sequence (K_n).

The limit function Q_c^+ is continuous, as can be easily proved by a standard "three epsilon" argument. More is true: Because the sequence $(Q_{K_n}^+)$ is monotone, Dini's theorem (Ref. 6, p. 248, Ex. 9.9) assures us that the convergence is *uniform* on any compact subset of $[1, \infty)$.

Each member $(Q_{K_n}^+)$ of the sequence satisfies the mass flow relation

$$K_n = R^2(Q_{K_n}^+ - \beta^2/R^2)\rho^2(Q_K^+). \tag{9.5}$$

Because ρ is continuous as a funciton of Q, the sequential continuity of this relation implies that Q_c^+ satisfies the critical mass flow relation

$$\text{(MFC)} \quad K_c = R^2(Q_c^+ - \beta^2/R^2)\rho^2(Q_c^+), \tag{9.6}$$

which proves (h).

We now show that Q_c^+ is C^1 when $R > 1$. We do this by computing dQ_c^+/dR. Along the way, we establish (i) and (j).

By Corollary 8.3 each element $(Q_{K_n}^+)$ of the sequence globally satisfies the integral equation

$$Q_{K_n}^+(R) = \hat{Q}(R) + \left\{ [\hat{Q}(1) - Q_{K_n}^+(1)]^3 + \int_1^R (8/t) g(t, Q_{K_n}^+)\, dt \right\}^{1/2}, \tag{9.7}$$

where g is given by

$$g(R, Q(R)) = [1/(\gamma + 1)]\{Q(R)[1 - \tfrac{1}{2}(\gamma - 1)Q(R)] + (\beta^2/R^2)[Q(R) - \hat{Q}(R)]\}. \tag{9.8}$$

Because $g(R,Q)$ is continuous as a function of Q, and because the convergence of $Q_{K_n}^+$ to Q_c^+ is uniform on compact sets by Dini's theorem, we have that Q_c^+ satisfies the integral equation

$$Q_c^+ = \hat{Q}(R) + \left\{ [\hat{Q}(1) - Q_c^+(1)] + \int_1^R (8/t) g(t, Q_c^+)\, dt \right\}^{1/2}. \tag{9.9}$$

Now by the fundamental theorem of calculus we can differentiate this relation at any interior point of $[1, \infty)$ to obtain

$$\frac{dQ_c^+}{dR} = \frac{d\hat{Q}}{dR} - \frac{\frac{1}{2}[(8/R)g(R,Q_c^+(R))]}{\left\{[\hat{Q}(1) - Q_c^+(1)]^2 + \int_1^R (8/t)g(t,Q_c^+)\,dt\right\}^{1/2}}. \tag{9.10}$$

Thus Q_c^+ is C^1 if $R > 1$. These last two equations have many consequences.

From the equation for Q_c^+ we see that $\hat{Q}(1) = Q_c^+$ [and similarly $\hat{Q}(1) = Q_c^-$]; thus $Q_c^+(1) = Q_c^-(1) = Q(1)$, showing that Q_c^+ and Q_c^- bifurcate from $R = 1$. Also from this, we see that Q_c^+ solves initial value problem II as well as mass flow–circulation problem I.

From the above expression for dQ_c^+/dR we see that Q_c^+ is monotone increasing as a function of R (similarly Q_c^- is monotone decreasing as a function of R) and also that dQ_c^+/dR is infinite when $R = 1$.

From the integral equation above for Q_c^+ it follows by algebra that:

$$\frac{d}{dR}(\hat{Q} - Q_c^+)^2 = \frac{8}{R}g(R,Q_c^+), \tag{9.11}$$

where

$$g(R,Q_c^+(R)) = [1/(\gamma-1)] \times \{Q_c^+(R)[1 - \tfrac{1}{2}(\gamma-1)Q_c^+(R)] + (\beta^2/R^2)[Q_c^+(R) - \hat{Q}(R)]\}, \tag{9.12}$$

from which, repeating the proof of Corollary 8.1, it follows that Q_c^+ is supercritical. Similarly, Q_c^- is subcritical.

Finally the critical mass flow relation implies that at infinity $Q^+ = 2/(\gamma-1)$ (the square cavitation speed) and $Q^- = 0$. QED

10. THE CASE OF $\gamma = 3$

When $\gamma = 3$, the mass flow relation becomes quadratic and $Q^\pm$ satisfy a quadratic equation. Compare Ref. 1.

In this case we have

$$Q^{\pm 2} - 2\hat{Q}(R)Q^\pm + (\beta^2 + K)/R^2 = 0, \tag{10.1}$$

and $Q^\pm$ satisfy

$$Q^\pm = -\hat{Q}(R) \pm [\hat{Q}^2(R) - (\beta^2 + K)/R^2]^{1/2}, \tag{10.2}$$

where $\hat{Q}(R) = \frac{1}{2}(1 + \beta^2/R^2)$. The discriminate vanishes when $K = K_c$ and $R = 1$, which shows that $K_c = \frac{1}{4}(1+\beta^2)^2 - \beta^2$.

11. SHOCK SOLUTIONS

Consider Figs. 2 and 3. Each figure shows a flow that starts out on the Q^+ curve and drops to the Q^- curve at $R = R_s$. Such a flow is a possible solution of the mass flow–circulation problem because Q^+ and Q^- have the same mass flow constant K.

These solutions ae called *shocks*. In Ref. 1 such solutions also occurred. There because of the periodic nature of the flow, it was shown that shocks were possible only for critical mass flow and indeed must occur there.

In our problem the flow is not periodic, so shocks might also occur if K is not critical. See Fig. 3. However, they might not occur at all. In fact, when $\gamma = 3$, we give a proof that shocks do not exist.

The proof is based on the Prandle–Rankine–Hugonant condition below for shocks in a polytropic gas. See Ref. 3.

Prandle–Rankine–Hugonant condition

Let a shock occur at $R = R_s$. Let V_{n1} be the velocity normal to the shock ahead of the shock and let V_{n2} be the velocity normal to the schock behind the shock. Then:

$$V_{n1}V_{n2} = 2/(\gamma+1) \tag{11.1}$$

We now show:

Theorem 11.1: Let $\gamma = 3$. There are no shock solutions to the mass flow–circulation problem.

Proof: We first show that the shock is oblique and normal to R at R_s.

Since $Q^+(R_s)$ and $Q^-(R_s)$ share the same value of β, the θ-velocity is invariant across the shock. So, the θ direction is tangent to the shock. See Fig. 6.

We now show that $R_s < 1$, which proves shocks are impossible since our flows are exterior to the unit circle.

Recall that when $\gamma = 3$, Q^+ and Q^- satisfy the quadratic equation

$$Q^{\pm 2} - (1 + \beta^2/R^2)Q^\pm + (\beta^2 + K)/R^2 = 0. \tag{11.2}$$

Let

$$Q^+(R_s) = Q_s^+, \quad \alpha^+(R_s) = \alpha_s^+, \tag{11.3}$$

$$Q^-(R_s) = Q_s^-, \quad \alpha^-(R_s) = \alpha_s^-. \tag{11.4}$$

Then,

$$Q_s^{\pm 2} - (1 + \beta^2/R_s^2)Q_s^\pm + (\beta^2 + K)/R_s^2 = 0. \tag{11.5}$$

Thus

$$Q_s^+ + Q_s^- = 1 + \beta^2/R_s^2, \tag{11.6}$$

$$Q_s^+ Q_s^- = (\beta^2 + K)/R_s^2. \tag{11.7}$$

In terms of α_s^+ and α_s^+ we have

$$\alpha_s^{+2} + \alpha_s^{-2} = [Q_s^+ - \beta^2/R_s^2] + (Q_s^- - \beta^2/R_s^2) \tag{11.8}$$

or

$$\alpha_s^{+2} + \alpha_s^{-2} = 1 - \beta^2/R_s^2 \tag{11.9}$$

and also

$$(\alpha^{+2} + \beta^2/R_s^2)(\alpha^{-2} + \beta^2/R_s^2) = (\beta^2/R_s^2) = (\beta^2 + K)/R_s^2 \tag{11.10}$$

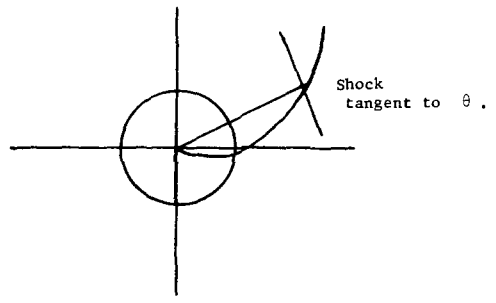

FIG. 6. Oblique shock.

or, equivalently,

$$\alpha_s^{+2}\alpha_s^{-2} + (\alpha_s^{+} + \alpha_s^{-2})\beta^2/R_s^2 + (\beta^2/R_s^2)^2 = (\beta^2 + K)/R_s^2. \quad (11.11)$$

Since $\alpha^{+2} + \alpha^{-2} = 1 - \beta^2/R_s^2$, we have

$$\alpha_s^{+2}\alpha_s^{-2} + (1 - \beta^2/R_s^2)\beta^2 R_s^2 + (\beta^2/R_s^2)^2 = (\beta^2 + K)/R_s^2. \quad (11.12)$$

Consider $\alpha_s^{+2}\alpha_s^{-2}$. Because the shock is normal to R at $R = R_s$, $\alpha_s^{+} = V_{n1}$, and $\alpha_s^{-} = V_{n2}$. Thus by the Prandle–Randkine–Hugonant condition above, $\alpha_s^{+2}\alpha_s^{-2} = 2/(\gamma + 1) = \frac{1}{2}$.

Thus we have

$$\tfrac{1}{2} + (1 - \beta^2/R_s^2)\beta^2/R_s^2 + (\beta^2/R_s^2)^2 = (\beta^2 + K)/R_s^2. \quad (11.13)$$

At present let $K = K_c$. When $\gamma = 3$, we showed in the previous section that

$$\beta^2 + K_c = \tfrac{1}{4}(1 + \beta^2). \quad (11.14)$$

The last two equations give the equation below for R_s.

$$\tfrac{1}{2}R_s^4 + (\tfrac{3}{4}\beta^2 - \tfrac{1}{4})R_s^2 + (\beta^2 - \beta^4) = 0, \quad (11.15)$$

which has roots

$$R_s^2 = (\tfrac{1}{4} - \tfrac{3}{4}\beta^2) \pm \sqrt{(\tfrac{1}{4} - \tfrac{3}{4}\beta^2)^2 - 2(\beta^2 - \beta^4)}. \quad (11.16)$$

Thus,

$$R_s^2 < 1 \quad (\text{when } \beta = 0,\ R_s = 1/\sqrt{2}). \quad (11.17)$$

So, if $K = K_c$, shocks do not occur. Now if $K \neq K_c$, then $K < K_c$ and $K + \beta^2 < \beta^2 + K_c < \frac{1}{4}(1 + \beta^2)$. Replacing the quadratic above by a quadratic inequality proves that again $R_s < 1$. QED

We have thus proved that if $\gamma = 3$, shocks do not occur.

When $\gamma \neq 3$, the author conjectures that shocks also do not occur. A proof proabably would involve Prandle's condition and careful estimates based on the integral equation for $Q^{\pm}$.

12. CONCLUSION

We have demonstrated a family of smooth transonic flows in the plane. The method also can be used to analyze other transonic plane flows (e.g., Ringleb flow[3]), previously incorrectly treated by the Hodograph method. In addition, the author has also treated certain three-dimensional flows by this method (e.g., pipe flow), and here too, smooth transonic families occur.

[1] L. M. Sibner and R. J. Sibner, "Transonic Flow on an Axially Symmetric Torus," J. Math. Anal. (to be published).

[2] L. Bers, *Mathematical Aspects of Subsonic and Transonic Gas Dynamics* (Wiley, New York, 1958).

[3] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves* (Interscience, New York, 1948).

[4] L. M. Sibner, and R. J. Sibner, "Non-Linear Hodge–de-Rham Theorem," Acta Math. **125**, 57–73 (1970).

[5] L. M. Sibner, and R. J. Sibner, "Non-Linear Hodge Theory: Applications," Adv. Math. **31**, 1–15 (1979).

[6] T. Apostol, *Mathematical Analysis* (Addison-Wesley, Reading, Mass., 1974), 2nd ed.

Symmetry of the complete second-order conductivity tensor in a Vlasov plasma

Jonas Larsson
Department of Plasma Physics, Umeå University, S-901 87 Umeå, Sweden

(Received 22 June 1983; accepted for publication 2 September 1983)

This paper has two purposes. The first is to consider the origin of a recently derived symmetry property including the pole contributions of the second-order conductivity. The second is to show how certain general formulas for the conductivities easily lead to much more convenient expressions than those used in the above-mentioned derivation of the symmetry.

PACS numbers: 52.25.Fi, 02.30. + g

The second-order conductivity tensor for a plasma described by the Vlasov–Maxwell equations can be expressed in terms of an integral involving poles due to resonant wave-particle interaction. The nonresonant particles determine the principal part of the integral while the resonant particles give pole contributions. Neglecting the pole contributions, we obtain the very well-known symmetry leading to the Manley–Rowe relations. Recently a symmetry relation was found[1] involving also the pole contributions. The derivation was a straightforward but lengthy calculation resulting in a very extensive formula for the second-order conductivity tensor in an unmagnetized relativistic plasma.

In this paper we observe that previously derived[2–4] general formulas for the conductivities directly lead to symmetries involving both the principal parts and the pole contributions. The symmetries are valid for a relativistic plasma also in the magnetized case. It will be shown below that the symmetry in Ref. 1 is included.

It is convenient to consider the quantities V, related to the second-order conductivity as[4]

$$\mathbf{V}(0,1,2)\equiv V(k_0,\mathbf{E}_0,k_1,\mathbf{E}_1,k_2,\mathbf{E}_2) = (2i/\omega_0)\mathbf{E}_0\cdot\boldsymbol{\sigma}^{(2)}_{k_1,k_2}(\mathbf{E}_1,\mathbf{E}_2), \tag{1a}$$

where $k_j = (\omega_j,\mathbf{k}_j)$, $j = 0,1,2$ and

$$\omega_0 + \omega_1 + \omega_2 = 0, \quad \mathbf{k}_0 + \mathbf{k}_1 + \mathbf{k}_2 = 0 \tag{1b}$$

and $\mathbf{E}_j$ are arbitrary vectors used as arguments in (1a). Now V may be writen as[4]

$$V(0,1,2) = \int f_0(\mathbf{v})A(k_0,\mathbf{E}_0,k_1,\mathbf{E}_1,k_2,\mathbf{E}_2,\mathbf{v})d^3v \equiv \int f_0(\mathbf{v})A(0,1,2\mathbf{v})d^3v, \tag{2}$$

where A is symmetric in the indices (0,1,2), i.e.,

$$A(0,1,2,\mathbf{v}) = A(\alpha,\beta,\gamma,\mathbf{v}) \quad \text{for } \{\alpha,\beta,\gamma\} = \{0,1,2\}. \tag{3}$$

There are, however, poles in the integrand of (2) due to denominators $(\omega_j - \mathbf{k}_j\cdot\mathbf{v})$ for an unmagnetized plasma and $(\omega_j - k_{jz}v_z - n\omega_c)$ for a magnetized plasma. These poles must be treated properly. Let us introduce operators P and R_j, where P stands for the principal part and R_j stands for pole contribution of the denominators containing ω_j mentioned above treated according to the prescription $\omega_j + i\eta$, $\eta \to 0+$. Then we have[4]

$$V(0,1,2) = PV - R_0V + R_1V + R_2V. \tag{4}$$

In (4) we have for brevity not indicated the arguments on the right-hand side since the symmetry (3) implies

$$PV(0,1,2) = PV(\alpha,\beta,\gamma), \quad R_jV(0,1,2) = R_jV(\alpha,\beta,\gamma) \quad \text{for} \quad \{\alpha,\beta,\gamma\} = \{0,1,2\}. \tag{5}$$

It is clear from (4) that $V(0,1,2)$ does not have the corresponding symmetry. A calculation of $V(0,1,2)$ naturally means a calculation of each term in (4). Then we have determined not only $V(0,1,2)$ but also all $V(\alpha,\beta,\gamma)$, where $\{\alpha,\beta,\gamma\} = \{0,1,2\}$. More substantial symmetries may be obtained in situations where some of the pole contributions may be neglected.

Considering resonant wave interaction between two high-frequency waves k_0 and k_1 with the low-frequency wave k_2, we may sometimes take $R_0V = R_1V = 0$. Then $V(0,1,2) = V(1,0,2) = PV + R_2V$, while $V(2,0,1) = PV - R_2V$. The coupled mode equations, in which we now may omit the linear damping of wave 0 and 1, are then simplified. Different particular forms of these equations are considered in Ref. 5.

Let us now compare with the symmetry result (26) in Ref. 1. We may write it in the form

$$V(0,1,2) = (P + R_1 + R_2)S(1,0,2) + (P + R_1 - R_0)S(1,2,0), \tag{6}$$

where S is related to the tensor S_{ijl} in Ref. 1 as

$$S(0,1,2)\equiv S(k_0,\mathbf{E}_0,k_1,\mathbf{E}_1,k_2,\mathbf{E}_2) = -iq(2\pi)^4(\omega_0\omega_1\omega_2)^{-1}S_{ijl}(-k_0,k_1,k_2)E_{0i}E_{1j}E_{2l}, \tag{7}$$

together with (1b). We also have

$$V(0,1,2) = S(0,1,2) + S(0,2,1). \tag{8}$$

It follows directly from (1) in Ref. 1 that

$$R_2S(1,0,2) = R_2V(1,0,2), \quad R_0S(1,2,0) = R_0V(1,2,0), \tag{9}$$

and we may thus rewrite (6) as

$$V(0,1,2) = PV(1,0,2) + R_1V(1,0,2) + R_2V(1,0,2) - R_0V(1,2,0). \tag{10}$$

But (10) follows directly from (4) and the derivation is thus completed.

 0022-2488/84/010185-02$02.50

We finally give a formula for the second-order conductivity tensor in an unmagnetized relativistic plasma. It is a particular case of the general formula[3] rewritten in more familiar notations and is clearly much more convenient to use than the formula which one obtains by straightforward calculations.[1] The result is

$$V(0,1,2) = -\frac{i}{m_0^2}\int F_0(\mathbf{v}) \times \frac{1}{(\omega_0 - \mathbf{k}_0\cdot\mathbf{v} - i\eta)(\omega_1 - \mathbf{k}_1\cdot\mathbf{v} + i\eta)(\omega_2 - \mathbf{k}_2\cdot\mathbf{v} + i\eta)} \times \left(\frac{\mathbf{k}_0\cdot\mathbf{F}_0 - (q\omega_0/c^2)\mathbf{v}\cdot\mathbf{E}_0}{\omega_0 - \mathbf{k}_0\cdot\mathbf{v} - i\eta}(\mathbf{F}_1\cdot\mathbf{F}_2 - q^2c^{-2}\mathbf{v}\cdot\mathbf{E}_1\mathbf{v}\cdot\mathbf{E}_2) + \text{even permutations of } (0,1,2) \right)(1 - \mathbf{v}^2/c^2)d^3v, \tag{11}$$

where

$$F_j = q\left(\mathbf{E}_j + \mathbf{v}\times\frac{\mathbf{k}_j\times\mathbf{E}_j}{\omega_j}\right) \quad \text{and } \eta \to 0+. \tag{12}$$

The property (3) is manifest in (11). The tensor S_{ijl} is explicitly obtained from (7) and (11) by substituting $\mathbf{E}_j = \hat{\mathbf{x}}_j$, where $(\hat{x}_0,\hat{x}_1,\hat{x}_2)$ are our orthonormal unit vectors.

The expression (11) is a good example of the usefulness of the general current response formulas[2,3] and it may simplify future application of them if we consider the notational change needed to obtain (11). From (2.11) and (2.13)–(2.15) in Ref. 3 we obtain in the notation of that paper

$$iu\cdot\kappa_\alpha\,\delta\tilde{x}_\alpha = \delta\tilde{u}_\alpha, \tag{13}$$

$$iu\cdot\kappa_\alpha\delta\tilde{u}_a = i(q/m_0c^2)\,\kappa_\alpha\Lambda\,\tilde{\phi}_\alpha\cdot u, \tag{14}$$

$$\tilde{\phi}_0\cdot\Lambda^{(2)}_{\kappa_1,\kappa_2}:\tilde{\phi}_1\,\tilde{\phi}_2 = \tfrac{1}{2}ic^3m_0\int f_0(u)[\kappa_0\cdot\delta\tilde{x}(0)\delta\tilde{u}(1)\cdot\delta\tilde{u}(2) + \text{even permutations of } (0,1,2)]\,du, \tag{15}$$

where $\kappa_\alpha = (\omega_\alpha/c)e_0 + \mathbf{k}_\alpha$ and $u = u^0(e_0 + \mathbf{v}/c)$ so that

$$u\cdot\kappa_\alpha = -(u^0/c)(\omega_\alpha - \mathbf{k}_\alpha\cdot\mathbf{v}), \tag{16}$$

where $u^0 = (1 - \mathbf{v}^2/c^2)^{-1/2}$.

In (15) the 4-vectors $\tilde{\phi}_\alpha$ are arbitrary. If we take $\tilde{\phi}_\alpha$ related to $\mathbf{E}_\alpha$ as the 4-potential is related to the electric field in Fourier space we obtain

$$\tilde{\phi}_0\cdot\Lambda^{(2)}_{\kappa_1,\kappa_2}:\tilde{\phi}_1\,\tilde{\phi}_2 = -(ic/\omega_0)\mathbf{E}_0\cdot\sigma^{(2)}_{\kappa_1,\kappa_2}[\mathbf{E}_1,\mathbf{E}_2] \tag{17}$$

and

$$\kappa_\alpha\Lambda\,\tilde{\phi}_\alpha\cdot u = -iu^0 \times\left(c^{-1}\mathbf{v}\cdot\mathbf{E}_\alpha\,e_0 + \mathbf{E}_\alpha + \mathbf{v}\times\frac{\mathbf{k}_\alpha\times\mathbf{E}_\alpha}{\omega_\alpha}\right). \tag{18}$$

Finally we need the relation between the distribution functions $f_0(u)$ and $F_0(\mathbf{v})$. The 4-current is

$$qc\int_S f_0(u)u\,du = \mathrm{q}\int(ce_0 + \mathbf{v})F_0(\mathbf{v})d^3v. \tag{19}$$

Taking the e_0-part of (19) we get the correspondence

$$f_0(u)u^0\,du = F_0(\mathbf{v})d^3v. \tag{20}$$

Or in more exact words, when we make the variable change $u \to \mathbf{v}$ defined by the relation $cu = u^0(ce_0 + \mathbf{v})$, then (20) is valid. From (1a), (13)–(18) and (20) we now obtain (11).

ACKNOWLEDGMENT

I thank the referee for checking all the equations above by rederiving them and for pointing out a number of typographical errors.

[1] H. E., Brandt, J. Math. Phys. **24**, 1332 (1983).
[2] J. Larsson, J. Math. Phys. **20**, 1321 (1979).
[3] J. Larsson, J. Math. Phys. **20**, 1331 (1979).
[4] J. Larsson, J. Plasma Phys. **21**, 519 (1979).
[5] J. Weiland and H. Wilhelmsson, *Coherent Non-Linear Interaction of Waves in Plasmas* (Pergamon, New York, 1977).

ERRATA

Erratum: Some remarks on the classical vacuum structure of gauge field theories [J. Math. Phys. 22, 179 (1981)]

M. Asorey
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, Spain

(Received 12 October 1983; accepted for publication 28 October 1983)

PACS numbers: 11.10.Np, 02.40.Vh, 99.10. + g

(1) Page 182 left column: Delete

$$\mathcal{A}_1^N U(1) = \{(\exp i\lambda_1,\ldots,\exp i\lambda_N)\in U(1)^N: \lambda_i\in[0, 2\pi), \lambda_1\leqslant\cdots\leqslant\lambda_{N-1}, (1/2\pi)\sum_{i=1}^{N}\lambda_i\in\mathbb{N}\},$$

and replace it by

$$\mathcal{A}_1^N U(1) = \{(\exp i\lambda_1,\ldots,\exp i\lambda_N)\in U(1)^N: \lambda_i\in[0, 2\pi), \lambda_1\leqslant\cdots\leqslant\lambda_N, (1/2\pi)\sum_{i=1}^{N}\lambda_i\in\mathbb{N}\}.$$

(2) Page 182 left column: Delete

$$\mathcal{V}^{SU(2)}_{\mathbb{R}^3\times S^1}\approx U(1)\approx\mathcal{V}^{SO(3)}_{\mathbb{R}^3\times S^1}\approx SO(2)\approx\mathcal{V}^{U(1)}_{\mathbb{R}^3\times S^1}\approx U(1),$$

and replace it by

$$\mathcal{V}^{SU(2)}_{\mathbb{R}^3\times S^1}\approx[0,\pi];\ \mathcal{V}^{SO(3)}_{\mathbb{R}^3\times S^1}\approx SO(2)\approx\mathcal{V}^{U(1)}_{\mathbb{R}^3\times S^1}\approx U(1).$$

Erratum: Splines and the projection collocation method for solving integral equations in scattering theory [J. Math. Phys. 24, 177 (1983)]

M. Brannigan
Department of Statistics and Computer Science, University of Georgia, Athens, Georgia 30602

D. Eyre
National Research Institute for Mathematical Sciences of the CSIR, P. O. Box 395, Pretoria 0001, Republic of South Africa

(Received 6 October 1983; accepted for publication 19 October 1983)

PACS numbers: 24.10. − i, 02.30.Rz, 25.10. + s, 02.60.Nm, 99.10. + g

1. The line after Eq. (2.1) should read "... space of continuous functions ... ."

2. Since the integral operator $\mathcal{K}$, containing the principal value integral, is not bounded on a space of continuous functions then our proof of convergence is not valid. However, convergence for this method is shown in our subsequent paper [J. Math. Phys. **24**, 1548 (1983)].

We are indebted to Ian H. Sloan for calling our attention to these points.

Erratum: Splines and the Galerkin method for solving the integral equations of scattering theory [J. Math. Phys. 24, 1548 (1983)]

M. Brannigan
Department of Statistics and Computer Sciences, University of Georgia, Athens, Georgia 30602

D. Eyre
National Research Institute for Mathematical Sciences of the CSIR, P. O. Box 395, Pretoria 0001, Republic of South Africa

(Received 6 October 1983; accepted for publication 19 October 1983)

PACS numbers: 03.80. + r, 02.30.Rz, 05.30.Jp, 99.10. + g

1. On page 1553 the scattering energy should read $(k/k_B)^2 = 0.64$.

2. Table II shows the square of the L 2-norm.