Tensor operator realizations of E_6 and structural zeros of the 6*j*-symbol

G. Vanden Berghe, H. De Meyer,[a] and J. Van der Jeugt[b]
Seminarie voor Wiskundige Natuurkunde, R. U. G., Krijgslaan 281-S9, B-9000 Gent, Belgium

(Received 19 July 1983; accepted for publication 27 January 1984)

For the exceptional Lie algebra E_6 minimal SO(3) tensor operator realizations are constructed which correspond to the maximal decomposition of E_6 in the respective chains $E_6 \supset F_4 \supset SO(3)$, $E_6 \supset SU(3) \otimes G_2 \supset SO(3) \otimes SO(3)$, $E_6 \supset F_4 \supset SO(3) \otimes G_2 \supset SO(3) \otimes SO(3)$, $E_6 \supset Sp(8) \supset SO(3)$, and $E_6 \supset G_2 \supset SO(3)$. Two particular realizations are shown to provide a basis in which certain structural zeros of Racah's 6*j*-symbol can be explained.

PACS numbers: 02.20. + b, 03.65.Fd

I. INTRODUCTION

From the observation that the standard SO(3) tensor formulation of the algebra G_2 exhibits the nontrivial vanishing of a 6*j*-coefficient, it has been suggested by Biedenharn and Louck[1] that realizations of the other exceptional Lie algebras might as well provide bases for explaining structural zeros. In two previous papers[2,3] the present authors have established tensor operator realizations of the algebra F_4 from which, taking into consideration Regge symmetries, 11 such zeros followed. In view of enlarging considerably the list of structural zeros that can be explained similarly, the in rank next higher exceptional Lie algebra E_6 is made the object of an analogous investigation in the present paper.

The method which we adopt here for the construction of tensor operator realizations has been originally described by Wadzinski.[4] In short, it consists in selecting for the algebra G under consideration a particular chain of maximal semisimple subalgebras ending at an SO(3) algebra or an outer product of SO(3) algebras. In that chain the consecutive decomposition of the adjoint irrep of G produces SO(3) labels which correspond to the rank labels k of the SO(3) tensors that can realize the algebra. Similarly the decomposition of the lowest-dimensional representation of G provides the numbers l which label the representation spaces on which the SO(3) tensors act. Hence, in the notation of Judd,[5] $v^k(l'l)$ denotes an SO(3) tensor operator of rank k which maps a $(2l+1)$-dimensional representation space into a $(2l'+1)$-dimensional one. Clearly, l, l', and k can have nonnegative integer or half-odd integer values with the restriction, however, that $l + l' + k$ is an integer. A realization $\{G^k\}$ of the algebra G in terms of the SO(3) tensor operators is then straightforwardly obtained from the fact that G closes under commutation and on using the standard commutation properties of the tensors.[2-5] Also for each of the subalgebras in the proposed chain a realization is found as a subset of $\{G^k\}$. It is then a matter of systematic investigation to verify whether structural zeros of the 6*j*-symbol can be explained within the realization. From branching rule tables[6] one readily learns that the exceptional Lie algebras are the best candidates for such explanations.

Throughout the paper we adopt the notations and conventions of Ref. 3. In particular unlike Wadzinski[4] and Judd[5] we do not use spectroscopical notation.

[a] Research Associate N. F. W. O., Belgium.
[b] Research Assistant N. F. W. O., Belgium.

II. TENSOR OPERATOR REALIZATION FROM THE CHAIN $E_6 \supset F_4 \supset SO(3)$

From the branching rule tables such as the ones established by McKay and Patera,[6] it is readily verified that along the chain $E_6 \supset F_4 \supset SO(3)$ of maximal subalgebras the 78-dimensional adjoint irrep of E_6 decomposes into the SO(3) irreps (1), (4), (5), (7), (8), and (11). It should be made clear that we label the SO(3) irreps by half the number McKay and Patera use. Similarly, the 27-dimensional irrep of E_6, the lowest-dimensional one, reduces into the SO(3) irreps (0), (4), and (8). Applying the construction algorithm described earlier,[2-4] we finally arrive at the following tensor operator realization of E_6:

$$
\begin{aligned}
G^1_q &= v^1_q(4,4) + (\tfrac{34}{5})^{1/2} v^1_q(8,8), \\
G^5_q &= v^5_q(4,4) + (\tfrac{38}{51})^{1/2} v^5_q(8,8) \\
&\quad + (-1)^{\alpha} 2(\tfrac{10}{51})^{1/2}[v^5_q(4,8) + v^5_q(8,4)], \\
G^7_q &= v^7_q(4,4) - (\tfrac{391}{209})^{1/2} v^7_q(8,8) \\
&\quad + (-1)^{\alpha}(\tfrac{714}{209})^{1/2}[v^7_q(4,8) + v^7_q(8,4)], \\
G^{11}_q &= v^{11}_q(8,8) - (-1)^{\alpha}(\tfrac{11}{30})^{1/2}[v^{11}_q(4,8) + v^{11}_q(8,4)], \\
G^4_q &= v^4_q(4,4) - (\tfrac{323}{154})^{1/2} v^4_q(8,8) \\
&\quad + (-1)^{\alpha}(\tfrac{85}{84})^{1/2}[v^4_q(4,8) + v^4_q(8,4)] \\
&\quad + (-1)^{\beta}(13/\sqrt{231})[v^4_q(4,0) + v^4_q(0,4)], \\
G^8_q &= v^8_q(4,4) - (\tfrac{437}{99})^{1/2} v^8_q(8,8) \\
&\quad - (-1)^{\alpha}(\tfrac{114}{55})^{1/2}[v^8_q(4,8) + v^8_q(8,4)] \\
&\quad + (-1)^{\alpha-\beta}(26/3\sqrt{55})[v^8_q(8,0) + v^8_q(0,8)]. \qquad (2.1)
\end{aligned}
$$

Herein α is a free parameter. Clearly $\{G^1\}$ forms the SO(3) subalgebra, whereas the subset $\{G^1, G^5, G^7, G^{11}\}$ realizes the maximal F_4 subalgebra. Apart from irrelevant scale factors the latter subset is also in conformity with previous results on F_4 (Ref. 2). Hence, when using (2.1) in order to try to explain structural zeros of the 6*j*-symbol we shall not insist again on those emerging from the realization of F_4 contained in (2.1) on its own. Ignoring momentarily Regge symmetry operations there is, however, one more structural zero that is related to the realization (2.1) of E_6. Indeed, let us consider a commutator of the type $[G^{11}, G^8]$, which at first sight should give origin to a term proportional to G^6. Since this type of generator is obviously missing in the E_6 algebra (2.1) all terms in the commutator which could contribute to G^6 either

 0022-2488/84/092585-04$02.50

vanish or cancel each other out. Effectively, a term $v^6_{q'+q}(4,4)$ could only be generated from the commutators $[v^{11}_{q'}(4,8), v^8_q(8,4)]$ and $[v^{11}_{q'}(8,4), v^8_q(4,8)]$, but since the contributions from both are equal and proportional to

$$\begin{Bmatrix} 11 & 8 & 6 \\ 4 & 4 & 8 \end{Bmatrix},$$

the latter 6j-coefficient has to vanish. On the other hand, terms such as $v^6_{q'+q}(8,8)$ are not generated proportional with a single 6j-coefficient and therefore their absence gives rise to a relation between different 6j-coefficients. In fact, a detailed investigation of (2.1) leads to the conclusion that the above-mentioned zero of the 6j-symbol is the only one that can be explained in the present context. However, due to Regge symmetry this structural zero generates five more different ones, which are collected all together in Table I.

III. THE CHAINS $E_6 \supset SU(3) \otimes G_2 \supset SO(3) \otimes SO(3)$ AND $E_6 \supset F_4 \supset SO(3) \otimes G_2 \supset SO(3) \otimes SO(3)$

From branching rule tables[6] we learn that the E_6 adjoint irrep (respectively, the lowest-dimensional irrep) decomposes in both chains mentioned in the title into the same SO(3) ⊗ SO(3) irreps, namely (0;1), (1;0), (2;3), (0;5), (2;0), and (1;3) [respectively, (1;3), (2;0), and (0;0)]. Application of the basic formulas established in Ref. 3 then leads in a straightforward way to the following SO(3) ⊗ SO(3) tensor operator realization of E_6:

$$\begin{aligned}
G^{0;1}_{q';q} &= v^0_{q'}(1,1)v^1_q(3,3), \\
G^{1;0}_{q';q} &= v^1_{q'}(1,1)v^0_q(3,3) \\
&\quad + (\tfrac{5}{7})^{1/2} v^1_{q'}(2,2)v^0_q(0,0), \\
G^{2;3}_{q';q} &= v^2_{q'}(1,1)v^3_q(3,3) \\
&\quad + [(-1)^\alpha/\sqrt{2}]\,[v^2_{q'}(1,2)v^3_q(3,0) \\
&\quad + v^2_{q'}(2,1)v^3_q(0,3)], \\
G^{0;5}_{q';q} &= v^0_{q'}(1,1)v^5_q(3,3), \\
G^{2;0}_{q';q} &= v^2_{q'}(1,1)v^0_q(3,3) \\
&\quad - (1/\sqrt{3})v^2_{q'}(2,2)v^0_q(0,0) \\
&\quad + (-1)^\beta (2/\sqrt{21})[v^2_{q'}(2,0)v^0_q(0,0) \\
&\quad + v^2_{q'}(0,2)v^0_q(0,0)], \\
G^{1;3}_{q';q} &= v^1_{q'}(1,1)v^3_q(3,3) \\
&\quad - (-1)^\alpha (\tfrac{5}{18})^{1/2}[v^1_{q'}(1,2)v^3_q(3,0)
\end{aligned}$$

TABLE I. Structural zeros of the 6j-symbol $\{\begin{smallmatrix} j_1 & j_2 & j_3 \\ l_1 & l_2 & l_3 \end{smallmatrix}\}$ which can be explained from realizations of the exceptional Lie algebras G_2, F_4, and E_6. The zeros marked with an ∘ are directly explained in the chains whereas the others follow from Regge symmetry.

	j_1	j_2	j_3	l_1	l_2	l_3	Algebra–subalgebra chain
∘	5	5	3	3	3	3	
	5	4	4	3	4	2	$G_2 \supset SO(3)$
∘	11	11	3	4	4	8	
	11	10	2	4	5	9	
∘	11	11	9	8	4	8	
	11	10	10	4	9	7	$F_4 \supset SO(3)$
	12	11	8	5	8	7	
	12	10	9	5	9	6	
	13	10	8	6	7	7	
	13	9	9	6	8	6	
∘	3	2	2	1	2	2	$F_4 \supset SO(3) \otimes G_2 \supset SO(3) \otimes SO(3)$
∘	$\frac{9}{2}$	$\frac{9}{2}$	7	4	4	$\frac{5}{2}$	$F_4 \supset SO(3) \otimes Sp(6) \supset SO(3) \otimes SO(3)$
	$\frac{13}{2}$	5	$\frac{9}{2}$	2	$\frac{9}{2}$	4	
∘	11	8	6	4	4	8	
	11	9	5	4	5	7	
	$\frac{21}{2}$	$\frac{19}{2}$	5	$\frac{7}{2}$	$\frac{11}{2}$	7	$E_6 \supset F_4 \supset SO(3)$
	$\frac{21}{2}$	$\frac{17}{2}$	4	$\frac{7}{2}$	$\frac{13}{2}$	8	
	$\frac{19}{2}$	$\frac{19}{2}$	6	$\frac{11}{2}$	$\frac{5}{2}$	8	
	$\frac{19}{2}$	9	$\frac{13}{2}$	$\frac{5}{2}$	5	$\frac{17}{2}$	
∘	7	6	5	4	6	4	
	7	$\frac{13}{2}$	$\frac{9}{2}$	4	$\frac{11}{2}$	$\frac{9}{2}$	
	$\frac{15}{2}$	$\frac{11}{2}$	5	$\frac{9}{2}$	$\frac{11}{2}$	4	
	$\frac{15}{2}$	6	$\frac{9}{2}$	$\frac{9}{2}$	5	$\frac{9}{2}$	
	$\frac{13}{2}$	6	$\frac{11}{2}$	$\frac{11}{2}$	3	$\frac{11}{2}$	
	6	6	6	6	5	3	
∘	6	6	6	5	4	3	$E_6 \supset Sp(8) \supset SO(3)$
	7	6	5	4	4	4	
	7	$\frac{11}{2}$	$\frac{11}{2}$	4	$\frac{9}{2}$	$\frac{7}{2}$	
	$\frac{13}{2}$	$\frac{13}{2}$	5	$\frac{9}{2}$	$\frac{7}{2}$	4	
	$\frac{13}{2}$	6	$\frac{11}{2}$	$\frac{7}{2}$	5	$\frac{7}{2}$	
	$\frac{13}{2}$	6	$\frac{11}{2}$	$\frac{9}{2}$	3	$\frac{9}{2}$	
∘	9	6	4	2	5	5	
	8	6	5	1	5	6	

$$+ v^1_{q'}(2,1)v^3_q(0,3)]$$
$$-(-1)^{\alpha+\beta}(\sqrt{2}/3)[v^1_{q'}(1,0)v^3_q(3,0)$$
$$+ v^1_{q'}(0,1)v^3_q(0,3)]. \tag{3.1}$$

Again α and β are free parameters. It is not very difficult to verify that $\{G^{0;1}\}\cup\{G^{1;0}\}$ realizes the algebra of SO(3) ⊗ SO(3), $\{G^{1;0},G^{2;0}\}\cup\{G^{0;1},G^{0;5}\}$ the algebra of SU(3) ⊗ G_2, $\{G^{1;0}\}\cup\{G^{0;1},G^{0;5}\}$ the algebra of SO(3) ⊗ G_2, and $\{G^{1;0},G^{0;1},G^{0;5},G^{2;3}\}$ the algebra F_4. A systematic investigation of the realization (3.1) shows that unfortunately the only structural zeros which can be explained from it are those already found from the partial chains $G_2 \supset$ SO(3) and $F_4 \supset$ SO(3) ⊗ $G_2 \supset$ SO(3) ⊗ SO(3).

IV. THE CHAINS $E_6 \supset$ Sp(8) $\supset$ SO(3) AND $E_6 \supset G_2 \supset$ SO(3)

Branching rule tables[6] show that the adjoint and lowest dimensional irrep of E_6 in both chains decompose into the same SO(3) irreps, namely into (1), (2), (3), (4), (5), (5), (7), and (8), and into (2), (4), and (6), respectively. Clearly the E_6 realization will contain two independent types of G^5 generators which we shall distinguish by a supplementary label. Straightforward calculations end up with the following SO(3) tensor operator realization of E_6:

$$G^1_q = v^1_q(2,2) + \sqrt{6}v^1_q(4,4) + (\tfrac{91}{5})^{1/2}v^1_q(6,6),$$
$$G^2_q = v^2_q(2,2) - \tfrac{5}{11}(\tfrac{2}{11})^{1/2}v^2_q(4,4)$$
$$+ \tfrac{7}{11}(\tfrac{13}{11})^{1/2}v^2_q(6,6)$$
$$-(-1)^{\alpha}(4/\sqrt{33})[v^2_q(2,4) + v^2_q(4,2)]$$
$$+(-1)^{\beta}\tfrac{14}{11}(\tfrac{26}{33})^{1/2}[v^2_q(4,6) + v^2_q(6,4)],$$
$$G^3_q = v^3_q(2,2) + (57/11\sqrt{11})v^3_q(4,4)$$
$$- \tfrac{7}{11}(\tfrac{26}{55})^{1/2}v^3_q(6,6)$$
$$+(-1)^{\alpha}2(\tfrac{6}{11})^{1/2}[v^3_q(2,4) + v^3_q(4,2)]$$
$$+(-1)^{\beta}\tfrac{7}{11}(\tfrac{546}{55})^{1/2}[v^3_q(4,6) + v^3_q(6,4)],$$
$$G^4_q = v^4_q(2,2) - 3(\tfrac{5}{143})^{1/2}v^4_q(4,4) - 7(\tfrac{34}{715})^{1/2}v^4_q(6,6)$$
$$-(-1)^{\alpha}\tfrac{2}{5}\sqrt{6}\,[v^4_q(2,4) + v^4_q(4,2)]$$
$$+(-1)^{\beta}\tfrac{7}{5}(\tfrac{6}{143})^{1/2}[v^4_q(4,6) + v^4_q(6,4)]$$
$$-(-1)^{\alpha+\beta}\tfrac{14}{5}(\tfrac{2}{13})^{1/2}[v^4_q(2,6) + v^4_q(6,2)],$$
$$G^{5a}_q = v^5_q(4,4) - (\sqrt{119}/3)v^5_q(6,6)$$
$$+(-1)^{\alpha}(\tfrac{715}{96})^{1/2}[v^5_q(2,4) + v^5_q(4,2)]$$
$$-(-1)^{\beta}\tfrac{7}{4}(\tfrac{5}{6})^{1/2}[v^5_q(4,6) + v^5_q(6,4)]$$
$$+(-1)^{\alpha+\beta}\tfrac{7}{12}(\tfrac{5}{2})^{1/2}[v^5_q(2,6) + v^5_q(6,2)],$$
$$G^{5b}_q = (-1)^{\alpha}[v^5_q(2,4) - v^5_q(4,2)]$$
$$-(-1)^{\beta}(\sqrt{143}/7)[v^5_q(4,6) - v^5_q(6,4)]$$
$$-(-1)^{\alpha+\beta}(\sqrt{39}/7)[v^5_q(2,6) - v^5_q(6,2)],$$
$$G^7_q = v^7_q(4,4) + (\tfrac{323}{147})^{1/2}v^7_q(6,6)$$
$$-(-1)^{\beta}\tfrac{2}{7}\sqrt{51}[v^7_q(4,6) + v^7_q(6,4)]$$
$$+(-1)^{\alpha+\beta}\tfrac{11}{7}(\tfrac{2}{3})^{1/2}[v^7_q(2,6) + v^7_q(6,2)],$$
$$G^8_q = v^8_q(4,4) - (\tfrac{133}{45})^{1/2}v^8_q(6,6)$$
$$-(-1)^{\beta}\tfrac{2}{7}(\tfrac{57}{5})^{1/2}[v^8_q(4,6) + v^8_q(6,4)]$$
$$-(-1)^{\alpha+\beta}\tfrac{11}{21}\sqrt{2}[v^8_q(2,6) + v^8_q(6,2)]. \tag{4.1}$$

Obviously we could replace G^{5a} and G^{5b} by any two independent linear combinations of them. However, the choice made in (4.1) is such that G^{5a} is completely symmetric and G^{5b} completely antisymmetric with respect to the mixed type tensor operators. Moreover, this particular choice implies that the subset $\{G^1,G^3,G^{5a},G^7\}$ is a realization of the maximal Sp(8) subalgebra. On the other hand, it turns out that there exists no real combination of G^{5a} and G^{5b} which, when added to G^1, spans the maximal G_2 subalgebra. This has to do with the fact already demonstrated by Dynkin[7] that there is no real (compact) form of G_2 comprised in a real form of E_6. Hence the algebra-maximal subalgebra inclusion $E_6 \supset G_2$ should be understood in terms of the classical (complex) Lie algebras alone[7,8] and becomes invalid when making the restriction to the real domain. Since all the tensor operator realizations which we have obtained so far are real, we have used for the algebras the notation of the real matrix groups by which they are generated. Nonetheless as a consequence of the previous remarks we should be able to realize from (4.1) the maximal G_2 subalgebra contained in E_6 by taking a complex combination of G^{5a} and G^{5b}. Indeed, actual calculations demonstrate that G_2 is generated from the set $\{G^1,G^{5a} + \lambda G^{5b}\}$, whereby $\lambda = i\epsilon 7\sqrt{385}/4\sqrt{78}$ and ϵ can be freely assigned the value $+1$ or -1. There exists, however, also a real realization of G_2 on the same representation space, of which it can be shown that it is nothing but a subset of a nonminimal realization of SO(7) which has the form

$$G^1_q = v^1_q(2,2) + \sqrt{6}\,v^1_q(4,4) + (\tfrac{91}{5})^{1/2}\,v^1_q(6,6),$$
$$G^3_q = v^3_q(2,2) + (15/4\sqrt{11})v^3_q(4,4)$$
$$- 7(\tfrac{13}{110})^{1/2}v^3_q(6,6)$$
$$+(-1)^{\alpha}(\sqrt{33}/4)[v^3_q(2,4) + v^3_q(4,2)]$$
$$+(-1)^{\beta}\tfrac{7}{4}(\tfrac{39}{55})^{1/2}[v^3_q(4,6) + v^3_q(6,4)],$$
$$G^5_q = v^5_q(4,4) - (\sqrt{119}/3)v^5_q(6,6)$$
$$-(-1)^{\alpha}4(\tfrac{55}{39})^{1/2}[v^5_q(2,4) + v^5_q(4,2)]$$
$$+(-1)^{\beta}2(\tfrac{35}{3})^{1/2}[v^5_q(4,6) + v^5_q(6,4)]$$
$$-(-1)^{\alpha+\beta}(\tfrac{385}{18})^{1/2}[v^5_q(2,6) + v^5_q(6,2)]. \tag{4.2}$$

It is readily verified that the nonminimal realization (4.2) is not suited for explaining structural zeros. On the other hand the realization (4.1) of E_6 offers a rich framework for such explanations. More precisely, the complete antisymmetric generators G^{5b} create the best opportunities to be exploited in connection with the missing of rank 6 and rank 9 tensor operators. Let us first analyze the commutator $[G^{5b},G^7]$, which could give rise to a term proportional to $v^6(4,4)$. But this term is absent and could only have been generated from $[v^5(4,6),v^7(6,4)]$ and $[v^5(6,4),v^7(4,6)]$, which produce exactly the same $v^6(4,4)$ coefficient, since $v^5(4,6)$ and $v^5(6,4)$ have op-

posite sign factors in G^{5b} and since $5+7+6$ is even. Hence, the common coefficient which is proportional to

$$\begin{Bmatrix} 5 & 7 & 6 \\ 4 & 4 & 6 \end{Bmatrix}$$

should vanish, which explains the zero value of the latter $6j$-coefficient. On account of Regge symmetries this structural zero gives rise to five more zeros of the $6j$-symbol which are collected in Table I. By a very similar argument one next succeeds in explaining the structural zero

$$\begin{Bmatrix} 5 & 3 & 6 \\ 6 & 6 & 4 \end{Bmatrix} = 0$$

from the commutator $[G^{5b},G^{3}]$.

Finally, we consider the commutator $[G^{5b},G^{5b}]$ in order to observe that a term of the type $v^{9}(4,6)$ could only originate from the commutators $[v^{5}(4,2),v^{5}(2,6)]$ and $[v^{5}(2,6),v^{5}(4,2)]$ contained in it. The sum $5+5+9$ being odd the only way to come out with a vanishing $v^{9}(4,6)$ term is the zero value of

$$\begin{Bmatrix} 5 & 5 & 9 \\ 4 & 6 & 2 \end{Bmatrix}.$$

As can be verified in Table I this zero is generic for one more structural zero.

Making the final counts we conclude that the chain $E_6 \supset \mathrm{Sp}(8) \supset \mathrm{SO}(3)$ on its own allows a minimal realization of E_6, which indirectly provides an explanation of 14 structural zeros of the $6j$-symbol.

V. CONCLUSIONS

Only a restricted number of minimal realizations of E_6 have been established in the present paper. Indeed, there exist many other chains starting at E_6 and ending with one or more SO(3) algebras, but the corresponding reduction patterns show levels of degeneracy which become too high for explaining structural zeros. Nonetheless we succeeded here to extend largely the list of zeros arising from G_2 and F_4 realizations. This fact suggests that many more zeros will become explained in E_7 and E_8, and we hope to report on them soon.

[1] L. C. Biedenharn and J. D. Louck, "The Racah–Wigner Algebra in Quantum Theory," in *Encyclopedia of Mathematics and Its Applications* (Addison–Wesley, London, 1981), Vol. 9, p. 415.
[2] J. Van der Jeugt, G. Vanden Berghe, and H. De Meyer, J. Phys. A: Math. Gen. **16**, 1377 (1983).
[3] H. De Meyer, G. Vanden Berghe, and J. Van der Jeugt, J. Math. Phys. **25**, 751 (1984).
[4] H. T. Wadzinski, Nuovo Cimento B **62**, 247 (1969).
[5] B. R. Judd, *Operator techniques in Atomic Spectroscopy* (McGraw–Hill, New York, 1963), p. 151.
[6] M. G. McKay and J. Patera, *Tables of Dimensions, Indices and Branching Rules for Representations of Simple Lie Algebras; Lecture Notes in Pure and Applied Mathematics* (Marcel Dekker, New York, 1981), Vol. 69.
[7] E. B. Dynkin, Am. Math. Soc. Transl. **6**, 111 (1957).
[8] R. Gilmore, *Lie Groups, Lie Algebras and Some of Their Applications* (Wiley, New York, 1974), p. 587.

On some general properties of the point spectrum of three particles moving in one dimension

J. Fernando Perez, C. P. Malta, and F. A. B. Coutinho
Departmento de Física Matemática, Instituto de Física da USP, São Paulo, Brazil

(Received 6 October 1983; accepted for publication 13 April 1984)

The eigenstates of three particles moving in one dimension are classified according to the S_3 plus parity group. The ordering of the ground state S_3 band is given for a fairly general class of potentials. Sufficient conditions are given both for existence and nonexistence of bound states of a given symmetry.

PACS numbers: 02.20. + b, 03.65.Ge, 21.40. + d

I. INTRODUCTION

In this paper we investigate some general properties of the bound-state spectrum of three identical particles interacting via two-body potential and moving in one dimension.

In Sec. II the eigenfunctions are classified according to their transformation properties under parity and the group S_3. Using these criteria the states can be grouped in two bands (positive and negative partity) of four states. As it should, each set of four eigenfunctions of a band provides a basis for the irreducible representations of the group S_3: two one dimensional (one compeltely symmetric, one completely antisymmetric) and one two dimensional (of mixed symmetry).

Sufficient conditions for the existence of the lowest energy state of each symmetry type are given in Sec. III. It results that the totally symmetric state of positive parity is always bound if the two-body potential is attractive. There are sufficient conditions for the existence of the lowest energy state of the other symmetry types and it is possible to give the ordering of the ground state band (of positive parity) for a large class of potentials.

Finally in Sec. IV we give sufficient conditions for the unboundness of the lowest-energy, negative-parity, totally symmetric, and totally antisymmetric states. Sufficient conditions for the unboundness of the first excited totally symmetric state of positive parity are also given.

The tools used as the k-harmonics method[1] (also called hyperspherical harmonics), the Hall and Post[2] and Hall[3] theorems, and the comparison theorem.[4]

II. THE S_3 BANDS

We consider the time-independent Schrödinger equation for three identical particles moving in one dimension and interacting via a two-body attractive potential $V(|x_i - x_j|)$. As it is well known, a three-body (identical particles) problem in one dimension can be reduced to a one-body problem in two dimensions. Using the "hyperspherical coordinates," [1] the hyperradius ρ and the hyperangle θ, defined by

$$\eta = \rho \cos\theta, \quad \xi = \rho \sin\theta, \quad 0 \leqslant \theta < 2\pi, \tag{1}$$

where η and ξ are the Jacobi coordinates

$$\eta = (1/\sqrt{2})(x_1 - x_2) \tag{2a}$$

$$\xi = \sqrt{2/3}\,[(x_1 + x_2)/2 - x_3] \tag{2b}$$

$$R = (x_1 + x_2 + x_3)/\sqrt{3} = 0\,, \tag{2c}$$

the Schrödinger equation becomes (using energy in units of $\hbar^2/2m$)

$$-\left[\frac{1}{\rho}\frac{\partial}{\partial\rho}\rho\frac{\partial}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial\theta^2}\right]\psi(\rho,\theta) + V(\rho,\theta)\,\psi(\rho,\theta) = E\psi(\rho,\theta)\,, \tag{3}$$

where

$$V(\rho,\theta) = V(\sqrt{2}\,\rho|\cos\theta|) + V(\sqrt{2}\,\rho|\cos(\theta + \pi/3)|) + V(\sqrt{2}\,\rho|\cos(\theta + 2\pi/3)|)\,. \tag{4}$$

If we now use the (k-harmonics) expansion

$$\psi(\rho,\theta) = \sum_{k=-\infty}^{\infty} \frac{R_k(\rho)}{\rho^{1/2}}\frac{e^{ik\theta}}{(2\pi)^{1/2}}\,, \tag{5}$$

we get the following infinite set of coupled ordinary differential equations:

$$-\left[\frac{d^2}{d\rho^2} - \left(k^2 - \frac{1}{4}\right)\frac{1}{\rho^2}\right]R_k(\rho) + \sum_{k'} V_{k'-k}(\rho)R_{k'}(\rho) = ER_k(\rho)\,, \tag{6}$$

where

$$V_{k'-k}(\rho) = \frac{1}{2\pi}\int_0^{2\pi} e^{i(k'-k)\theta}\,V(\rho,\theta)\,d\theta\,. \tag{7}$$

The parity operator Π and the permutation operators P_{12}, P_{13}, and P_{23} leave ρ invariant and transform θ onto $\pi + \theta$, $\pi - \theta$, $5\pi/3 - \theta$, and $\pi/3 - \theta$, respectively.

From parity invariance of $V(\rho,\theta)$ (4) it immediately follows that

$$V_{k'-k}(\rho) = 0, \quad \text{for } k' - k \text{ odd}, \tag{8a}$$

$$V_{k'-k}(\rho) = V_{k-k'}(\rho) = V_{-(k'-k)}(\rho)\,. \tag{8b}$$

Using the invariance of $V(\rho,\theta)$ under P_{12}, P_{13}, and P_{23} we get

$$V_{k'-k}(\rho) = \begin{cases} \dfrac{6}{2\pi}\displaystyle\int_0^{\pi/3} e^{i(k'-k)\theta}\,V(\rho,\theta)\,d\theta, \\ \quad \text{for } k' - k = 6n,\ n \text{ integer} \\ \quad\quad \text{from } -\infty \text{ to } +\infty, \\ 0, \quad \text{otherwise.} \end{cases} \tag{9}$$

 0022-2488/84/092589-04$02.50

Due to properties (8a) and (9) the system (6) splits into six infinite sets of k-values

$$k \text{ even: } 6n,\ 2+6n, \text{ and } -2+6n, \qquad (10)$$

$$k \text{ odd: } 3+6n,\ 1+6n, \text{ and } -1+6n.$$

Using property (8b) it can be easily verified that the solutions of the system of equations (6) can be chosen such that

$$R_{-k}(\rho) = \pm R_k(\rho). \qquad (11)$$

We shall denote by $R^E_k(\rho)$ the solutions such that $R_{-k}(\rho) = R_k(\rho)$ and by $R^O_k(\rho)$ the solutions such that $R_{-k} = -R_k(\rho)$.

The eigenfunctions of (3) given by expansion (5) can now be classified. Positive parity corresponds to k even and negative parity to k odd. The totally symmetric eigenfunction of positive parity is given by the set $6n$, the totally antisymmetric eigenfunction is given by the set $6n$ ($n \neq 0$), and the mixed symmetry eigenfunctions are given by the set $2+6n$ (or $-2+6n$). Thus, the positive parity ground-state band is given by

$$\psi_S^{(+)}(\rho,\theta) = \frac{1}{(2\pi)^{1/2}} \sum_{n=-\infty}^{\infty} \frac{R^E_{6n}(\rho)}{\rho^{1/2}} \cos 6n\theta, \qquad (12a)$$

$$\psi_A^{(+)}(\rho,\theta) = \frac{i}{(2\pi)^{1/2}} \sum_{n=-\infty}^{\infty} \frac{R^O_{6n}(\rho)}{\rho^{1/2}} \sin 6n\theta, \qquad (12b)$$

$$\psi_M^{(+)}(\rho,\theta) = \frac{1}{(2\pi)^{1/2}} \sum_{n=-\infty}^{\infty} \frac{R_{2+6n}(\rho)}{\rho^{1/2}} \times \exp[i(2+6n)\theta] \text{ and c.c.} \qquad (12c)$$

As the set $-2+6n$ equals minus the set $2+6n$, due to property (11) the eigenfunctions given by the set $-2+6n$ differ from (12c) only by a phase factor.

The negative parity totally symmetric and totally antisymmetric eigenfunctions are given by set $3+6n$ and the mixed symmetry eigenfunctions are given by the set $1+6n$ (or $-1+6n$). The negative parity ground-state band is then given by

$$\psi_S^{(-)}(\rho,\theta) = \frac{i}{(2\pi)^{1/2}} \sum_{n=-\infty}^{\infty} \frac{R^O_{3+6n}(\rho)}{\rho^{1/2}} \sin(3+6n)\theta, \qquad (13a)$$

$$\psi_A^{(-)}(\rho,\theta) = \frac{1}{(2\pi)^{1/2}} \sum_{n=-\infty}^{\infty} \frac{R^E_{3+6n}(\rho)}{\rho^{1/2}} \cos(3+6n)\theta, \qquad (13b)$$

$$\psi_M^{(-)}(\rho,\theta) = \frac{1}{(2\pi)^{1/2}} \sum_{n=-\infty}^{\infty} \frac{R_{1+6n}(\rho)}{\rho^{1/2}} \exp[i(1+6n)\theta] \text{ and c.c.} \qquad (13c)$$

(As in the case of the set $-2+6n$, the eigenfunctions given by the set $-1+6n$ differ from those given by $1+6n$ only by a phase factor.)

The transformation properties of the mixed symmetry eigenfunctions $\psi_M^{(\pm)}$ and $(\psi_M^{(\pm)})^*$ can be easily verified using the complex two-dimension irreducible representation of the group S_3 introduced by Simonov.[1]

Concluding this section we examine the order[4] of the states of the ground-state S_3 band. The lowest state is $\psi_S^{(+)}(\rho,\theta)$ since this state has no centrifugal barrier ($k=0$). Due to the centrifugal barrier the other symmetry states should occur in the following order [see Eq. (6)]: $\psi_M^{(-)}(\rho,\theta)$, $\psi_M^{(+)}(\rho,\theta)$, $\psi_{S,A}^{(-)}(\rho,\theta)$ (apparently degenerate), and finally $\psi_A^{(+)}(\rho,\theta)$. The degeneracy of $\psi_{S,A}^{(-)}(\rho,\theta)$ is easily shown to be only apparent (see the Appendix). Other results concerning the ordering of the states will be presented in Sec. IV.

III. SUFFICIENT CONDITIONS FOR THE EXISTENCE OF THE LOWEST ENERGY STATE OF EACH SYMMETRY TYPE

Truncation of expansions (12) and (13) provides variational functions and consequently an upper bound for the lowest energy state of each symmetry type. Keeping a single term in the expansion (12) and (13) the corresponding system of differential equations (6) for each symmetry type will be reduced to a single equation. The truncated equation is

$$H_k R_k(\rho) = E R_k(\rho), \qquad (14)$$

where

$$H_k = -\frac{d^2}{d\rho^2} + \left(k^2 - \frac{1}{4}\right)\frac{1}{\rho^2} + W_k(\rho),$$

with $W_k(\rho) = V_0$ for $k = 1,2$, and 6, $W_3(\rho) = V_0 - V_6$ for the $k=3$ odd solution (associated with $\psi_S^{(-)}$), and $W_3(\rho) = V_0 + V_6$ for the $k=3$ even solution (associated with $\psi_A^{(-)}$). Since the truncated equation (14) provides variational upper bounds for the system (6) if it has a bound state so will the system (6).

In principle it is not difficult to find sufficient conditions for the existence of a bound state of a given symmetry from Eq. (14). In fact, Hunziker's theorem[5,6] states that for a locally square integrable two-body potential vanishing at infinity, the continuum of a three-particle system starts at the ground state energy E^0_{2B} of the two-particle Hamiltonian

$$\left[-\frac{d^2}{d\rho^2} + V(\sqrt{2}\rho)\right]\psi(\rho) = E^0_{2B}\,\psi(\rho), \qquad (15)$$

where $\rho = (1/\sqrt{2})(x_1 - x_2)$ and where we have measured energy in units of $\hbar^2/2m$.

In fact, according to Hunziker's theorem with statistics,[6] the continuum threshold for the totally symmetric states and the mixed symmetry states is given by the ground state energy E^0_{2B} of (15) and for the completely antisymmetric states the continuum threshold is given by the first excited state energy E^1_{2B} of (15).

Therefore to find a sufficient condition for the existence of a bound state of a given symmetry we have to find a sufficient condition for Eq. (14) to have a bound state with energy less than the appropriate continuum threshold. This in turn can be reduced, by the Rayleigh–Ritz principle, to finding a trial function $\phi(\rho)$ such that $(\phi, H_k\phi) < E_{2B}(\phi,\phi)$, where E_{2B} is the appropriate continuum threshold.

Taking Simon's choice,[6] $\phi(\rho) = \rho^\alpha e^{-\rho/2}$, we obtain

$$(\alpha^2 + k^2 - \tfrac{1}{4})\Gamma(2\alpha-1) - \alpha\Gamma(2\alpha) + \tfrac{1}{4}\Gamma(2\alpha+1) + \int_0^\infty d\rho\,\rho^{2\alpha} e^{-\rho} W_k(\rho) < E_{2B}\Gamma(2\alpha+1), \qquad (16)$$

with $\alpha > \frac{1}{2}$ if $k \neq 0$.

For $k=0$ the limit $\alpha \to \frac{1}{2}$ exists giving the condition

$$\int_0^\infty e^{-\rho}\rho V_0(\rho)\,d\rho < E_{2B} - \frac{1}{4}. \tag{17}$$

This result is a weaker version of a previous result obtained by the authors[7] that guarantees the existence of at least one bound state for N-particle systems in one and two dimensions when the two-body interaction is globally attractive. For $k \neq 0$ a simple condition is obtained by taking $\alpha = 1$.

$$\int_0^\infty d\rho\, e^{-\rho}\rho^2 W_k(\rho) < 2E_{2B} - \left(\frac{1}{4} + k^2\right). \tag{18}$$

Alternatively, taking $\phi(\rho) = \rho^{k+\frac{1}{2}} \times \exp(-\sqrt{-E_{2B}}\,\rho)$ as the trial wave function, the sufficient condition for the existence of at least one bound state of symmetry k is given by

$$\int_0^\infty \rho^{2k+1}\exp(-2\sqrt{-E_{2B}}\,\rho)\, W_k(\rho)\,d\rho \leqslant -\frac{(k+\frac{1}{2})}{(-4E_{2B})^k}(2k)!\,. \tag{19}$$

In Ref. 8 we show that simple sufficient conditions are obtained by using as the trial function the regular and irregular solutions of the modified Helmholtz equation matched at an arbitrary point R (Calogero's sufficient conditions[9] are obtained in this way). In the present case this type of trial wave function gives the following awkward condition for the existence of a bound state with energy $\leqslant -\alpha^2$:

$$\begin{aligned} R\int_0^R \left(\frac{\rho}{R}\right)^{2k+1} e^{-2\alpha\rho} W_k(\rho)\,d\rho &+ R\int_R^\infty \left(\frac{R}{\rho}\right)^{2k-1} e^{-2\alpha\rho} W_k(\rho)\,d\rho \\ &\leqslant -2k\,e^{-2\alpha R} - 2\alpha\int_0^R \frac{(k+\frac{1}{2})}{R^{2k}} e^{-2\alpha\rho}\rho^{2k}\,d\rho \\ &\quad + 2\alpha\int_R^\infty \left(k-\frac{1}{2}\right) R^{2k} e^{-2\alpha\rho}\rho^{-2k}\,d\rho. \end{aligned} \tag{20}$$

IV. SUFFICIENT CONDITIONS FOR THE NONEXISTENCE OF BOUND STATES OF A GIVEN SYMMETRY

We shall use in this section two theorems due to Hall and Post[2] and Hall,[3] respectively.

The two theorems refer to a system of N particles interacting via a two-body potential $V(\mathbf{r}_i - \mathbf{r}_j)$. The translation-invariant Hamiltonian for such a system is

$$H = \sum_{i<j=1}^{N}\left\{-\frac{\hbar^2}{2mN}(\nabla_{r_i} - \nabla_{r_j})^2 + V(\mathbf{r}_i - \mathbf{r}_j)\right\}. \tag{21}$$

Theorem of Hall–Post: A lower bound for the energy of the completely symmetric (antisymmetric) states of system (21) is given by the ground state (first excited state) of the two-body Hamiltonian

$$\mathscr{H}_{\rm HP} = -(N-1)\left[(\hbar^2/2m)\nabla_\rho^2 - (N/2)V(\sqrt{2}\rho)\right], \tag{22}$$

where $\boldsymbol{\rho} = (1/\sqrt{2})(\mathbf{r}_1 - \mathbf{r}_2)$.

The Hall–Post theorem is valid in any dimension. So to apply it to the case of three particles moving in one dimension, taking $N=3$ and using energy in units of $\hbar^2/2m$ we have to consider the two-body problem

$$\mathscr{H}_{\rm HP} = -2\frac{d^2}{d\rho^2} + 3V(\sqrt{2}\rho). \tag{23}$$

Thus a direct application of Hall–Post theorem obtains the following result.

R1: If the Hamiltonian given by Eq. (23) has a single bound state then the state $\psi_A^{(-)}$ is unbound and so is $\psi_A^{(+)}$.

Theorem of Hall: Given a system of N particles described by the Hamiltonian (21) consider the Hamiltonian describing the independent motion of $N-1$ particles

$$\begin{aligned} \mathscr{H}_{\rm H} &= \sum_{i=2}^{N}\mathscr{H}_i \\ &= \sum_{i=2}^{N}\left(-\frac{\hbar^2}{2m}\frac{N}{2(N-1)}\nabla_{\rho_i}^2 + \frac{1}{2}V(\mathbf{r}_1 - \mathbf{r}_i)\right), \end{aligned} \tag{24}$$

where $\boldsymbol{\rho}_i = (1/\sqrt{2})(\mathbf{r}_i - \mathbf{r}_1)$.

Then, the exact energies of the completely symmetric (antisymmetric) states of the N-body problem (21) are bounded one by one by the energies of the completely symmetric (antisymmetric) states of the Hamiltonian $\mathscr{H}_{\rm H}$ (24).

The Hall[3] theorem is also valid in any dimension. So, in the particular case of three particles moving in one dimension lower bounds for the energy spectrum are obtained by solving the two-body problem

$$\mathscr{H}_{\rm H} = \frac{3}{2}\left[-\frac{1}{2}\frac{d^2}{d\rho^2} + V(\sqrt{2}\rho)\right], \tag{25}$$

which is obtained from (24) by setting $N=3$ and $\hbar^2/2m = 1$.

Suppose now that $\mathscr{H}_{\rm H}$ has three bound states with energies ϵ_0, ϵ_1, and ϵ_2 such that $\epsilon_0 < \epsilon_1 < \epsilon_2$ and $(\epsilon_0 + \epsilon_1) < 2\epsilon_2$. Then the Hall theorem tells us that a lower bound for the state $\psi_S^{(+)}$ is given by $2\epsilon_0$ and a lower bound for the states $\psi_S^{(-)}$ and $\psi_A^{(-)}$ is given by $(\epsilon_0 + \epsilon_1)$. Now according to Hunziker's theorem,[5] the continuum threshold of the symmetric states is E_{2B}^0 and the continuum threshold of the antisymmetric states is E_{2B}^1. So we have the following result.

R2: If $\epsilon_0 + \epsilon_1 > E_{2B}^0$, $\psi_S^{(-)}$ is unbound and if $\epsilon_0 + \epsilon_1 > E_{2B}^1$ then $\psi_A^{(-)}$ also is unbound. Since $\psi_A^{(+)}$ is less bound than $\psi_A^{(-)}$ then in this last case $\psi_A^{(+)}$ is also unbound.

The Hall theorem also tells us that a lower bound for $\psi_A^{(+)}$ is given by $(\epsilon_0 + \epsilon_2)$ and a lower bound for the first excited state of the type $\psi_S^{(+)}$ is given by $2\epsilon_2$. Therefore we have the following result.

R3: If $\epsilon_0 + \epsilon_2 - E_{2B}^1 < 0$ and $2\epsilon_2 - E_{2B}^0 > 0$ the states $\psi_S^{(-)}$, $\psi_A^{(-)}$, and $\psi_A^{(+)}$ might be bound but the first excited state of the type $\psi_S^{(+)}$ will be unbound and the ground state band will not intercept the excited state bands.

Finally we would like to comment on the finiteness or infinitude of the bound state spectrum of our system. First we would like to recall a theorem by Sigal[10] which states that for short-range potentials the number bound states of the N-particle Schrödinger operator in the center-of-mass frame is at most finite.

In our case this result follows from the Hall[3] theorem: if the two-body potential has only a finite number of bound

states then so does the Hamiltonian (25). Therefore we can construct only a finite number of completely symmetric (antisymmetric) lower bounds below the respective continuum thresholds and due to the centrifugal barrier the number of bound states of mixed symmetry type is less than the number of symmetric states. Therefore as a particular case of the Sigal[10] theorem we have that for short-range potentials our three-particle system has at most a finite number of bound states.

ACKNOWLEDGMENT

Two of us (J.F.P. and F.A.B.C.) were partially supported by CNPq.

APPENDIX: NONDEGENERACY OF $\psi_S^{(-)}(\rho,\theta)$ AND $\psi_A^{(-)}(\rho,\theta)$

The degeneracy of $\psi_S^{(-)}(\rho,\theta)$ and $\psi_A^{(-)}(\rho,\theta)$ is shown to be false by examining the system of differential equations (6) for the radial vectors $R^{O}_{3+6n}(\rho)$ and $R^{E}_{3+6n}(\rho)$ which enter in $\psi_S^{(-)}(\rho,\theta)$ (13a) and $\psi_A^{(-)}(\rho,\theta)$ (13b), respectively. For $R^{E}_{3+6n}(\rho)$ the system (6) can be cast in the form

$$-\left[\frac{1}{\rho}\frac{d}{d\rho}\rho\frac{d}{d\rho}-\frac{(3+6n)^2}{\rho^2}\right]R^{E}_{3+6n}(\rho)+\sum_{n'=0}^{\infty}\left[V_{6(n'-n)}+V_{6(n'+n+1)}\right]R^{E}_{3+6n'}(\rho)=E\,R^{E}_{3+6n'}(\rho),$$

and for $R^{O}_{3+6n}(\rho)$ it becomes

$$-\left[\frac{1}{\rho}\frac{d}{d\rho}\rho\frac{d}{d\rho}-\frac{(3+6n)^2}{\rho^2}\right]R^{O}_{3+6n}(\rho)+\sum_{n'=0}^{\infty}\left[V_{6(n'-n)}-V_{6(n+n'+1)}\right]R^{O}_{3+6n'}(\rho)=ER^{O}_{3+6n}(\rho).$$

As the potential energy for R^{E}_{k} and R^{O}_{k} are different the states $\psi_S^{(-)}$ and $\psi_A^{(-)}$ will not be degenerate but for an accident. It is not possible to say the order in which they will occur because the potential energy terms cannot be easily compared. The only thing we can say is that for a purely attractive interparticle potential monotonically increasing in the interparticle distance we have the matrix elements $V_{12n}<0$ and $V_{(2n+1)6}>0$ and truncation of the above equations at $n=n'=0$ provides an upper bound for the energy of the state $\psi_S^{(-)}$ which is lower than the upper bound for the energy of the state $\psi_A^{(-)}$.

[1] Yu. A. Simonov, Sov. J. Nucl. Phys. **3**, 461 (1966).
[2] R. L. Hall and H. R. Post, Proc. Phys. Soc. **90**, 381 (1967).
[3] R. L. Hall, Phys. Rev. C **20**, 1155 (1979).
[4] A. Weinstein and W. Stenger, *Methods of Intermediate Problems for Eigenvalues* (Academic, New York, 1972).
[5] W. Hunziker, Helv. Pys. Acta **39**, 451 (1966).
[6] B. Simon, Helv. Phys. Acta **43**, 607 (1970).
[7] F. A. B. Coutinho, C. P. Malta, and J. Fernando Perez, Phys. Lett. A **97**, 242 (1983).
[8] C. P. Malta, J. Fernando Perez, and F. A. B. Coutinho, "Bound states of N-particles: a variational approach," submitted to Nuovo Cimento; "Sufficient conditions for the existence of bound states of N-particles with attractive potentials," Phys. Lett. A **100**, 460 (1984).
[9] F. Calogero, J. Math. Phys. **6**, 161 (1965).
[10] I. Sigal, Commun. Math. Phys. **48**, 137 (1976).

On the classical part of the mean field dynamics for quantum lattice systems in grand canonical representations

Alfred Rieckers
Institut für Theoretische Physik, Universität Tübingen, D-7400 Tübingen, West Germany

(Received 9 November 1983; accepted for publication 3 February 1984)

For a class of discrete mean field models the limiting dynamics is investigated in the representations of generalized grand canonical states. It is demonstrated that for a certain form of spontaneous symmetry breakdown the W^*-automorphism dynamics exhibits a uniquely determined nontrivial classical part, which is essential for the explanation of macroscopic quantum phenomena.

PACS numbers: 02.20. + b, 05.30.Ch, 05.70.Fh

I. INTRODUCTION

In many body physics the dynamics is sometimes formulated by means of the physical Hamiltonian containing only the microscopic energies, and sometimes by means of the reduced Hamiltonian in which also macroscopic contact variables are incorporated. The difference seems to be considered only a question of energy renormalization. This point of view explains perhaps why in the C^*-algebraic formulation[1] of the BCS-model only the KMS-dynamics has been worked out hitherto, which corresponds in virtue of the grand canonical equilibrium state to the reduced Hamiltonian.[2,3] As has been demonstrated in Ref. 4, the physical dynamics has in this case, however, an additional part—in comparison to the KMS-dynamics—which acts nontrivially on the center of the relevant algebra of observables. Thus, only the physical dynamics are capable of explaining all of the macroscopic quantum phenomena of superconductivity.[4]

In the present investigation we treat this dynamical phenomenon for a general class of discrete molecular field models and show that the mathematical construction of the closure of the limiting Heisenberg generators, which originally are defined only on local observables, leads automatically to the classical parts of the physical dynamics, if certain symmetries are spontaneously broken.

In Sec. II we start from a converging subnet of local grand canonical equilibrium states. The family of local subtraction terms in the reduced Hamiltonians with fixed contact variables gives rise to an internal symmetry group acting as automorphisms of the quasilocal algebra. To have a name we call this the generalized gauge group. The additional classical structure terms arise if these symmetry transformations are spontaneously broken in the limiting equilibrium state, which would correspond to the spontaneous breaking of gauge invariance for the BCS-model. The interesting point is that these additional classical features can be determined in a model independent manner. The first step to do this is the extension of the gauge transformations to W^*-automorphisms of the temperature-dependent GNS–von Neumann algebra. The generator of this W^*-automorphism group is shown to contain a differential operator with respect to a continuous variable, which labels part of the overcountably many pure-phase-states.

In Sec. III, the limiting Heisenberg generator is investigated starting from the immediate and weak result that the local Heisenberg generators converge in the strong operator topology of the temperature representation if applied to fixed local observables. This limit defines an antisymmetric $*$-derivation L_0^β with the local observable algebra as domain. The σ-weak closure L^β of L_0^β exhibits the aforementioned differential operator, if the symmetry under gauge transformations is spontaneously broken. It is investigated in which way the image values of L^β can be approached by the image values of L_0^β. The qualitatively new features of L^β in comparison of L_0^β seem to prevent in general a local approximation in any topology stronger than the σ-weak one. The still available approximation property is obtained by means of a certain estimation, which is proved in the Appendix. Avoiding the difficult use of general criteria (cf. Ref. 1, Chap. 3) we show L^β to be the generator of a σ-weakly continuous W^*-automorphism group in the representation–von Neumann algebra by explicitly constructing the dynamical transformations. A remarkable feature is the appearance of the subtraction terms in the thermodynamic limit of the (unsubtracted) physical dynamics in an intertwined manner, dropping out only if the gauge symmetry is not spontaneously broken [cf. formula (3.15)].

In Sec. IV we treat the convergence of the local finite-time dynamics in the infinite volume limit. The crucial point is that in spite of the weak results on the generator convergence, which prevents the application of general theorems of semigroup theory, the convergence of the powers of the local Heisenberg generators is under control, if these are applied to local observables. In virtue of this result is seems still more advantageous to use a generator method for the considered class of long-range interacting models than to prove the convergence of the finite-time translations by means of quite general but rather indirect arguments.[5] For this latter problem we use here a norm-bound [namely (4.5)] for the powers of the loal Heisenberg generators which is rather popular[6] but completely derived only in Ref. 7, at least for our general class of models. The arguments for extending convergence for small times to convergence on the whole time axis are also worked out in Ref. 7. In this way we arrive here at the result that the limits of the local time translations are just the W^*-automorphisms generated by L^β. No dilatation prob-

 0022-2488/84/092593-09$02.50

lem[8] arises in spite of the extreme long-range potentials and in spite of the nontrivial classical features of the limiting dynamics.

Altogether our construction illustrates in which way the properties of the atomic constituents of a many-particle system determine unambiguously also the classical features of the dynamics, a problem which has been put forward in its fundamental importance and discussed from several points of view in Ref. 9.

II. EQUILIBRIUM REPRESENTATION AND GAUGE GROUPS

We consider a discrete quantum system, where on every site $i\in\mathbb{Z}^d$ of a d-dimensional lattice some microscopic constituents occupy states of a finite-dimensional Hilbert space $\mathscr{H}_i$ each of which being an exemplar of the Hilbert space $\mathscr{H}$ with $\dim \mathscr{H} = k\in\mathbb{N}$. Denoting by $|\Lambda|$ the cardinality of a set Λ we introduce $\mathscr{L} := \{\Lambda\subset\mathbb{Z}^d; |\Lambda| < \infty\}$, which is a directed set by means of the inclusion relation. With $\Lambda\in\mathscr{L}$ are associated the Hilbert space $\mathscr{H}_\Lambda := \otimes_{i\in\Lambda}\mathscr{H}_i$ and the algebra of observables $\mathscr{A}_\Lambda := \mathscr{B}(\mathscr{H}_\Lambda)$. In $\mathscr{A}_i$ (where we have identified $\{i\}$ with i) we select a basis $\{s_i^a; a = 1\cdots k^2\}$ in identical manner for all i such that $s_i^{a*} = s_i^a$ and $\|s_i^a\| = 1$ for all a. Writing $s_\Lambda^a := \Sigma_{i\in\Lambda} s_i^a$ we consider the following class of local Hamiltonians:

$$H_\Lambda := \sum_a f_a s_\Lambda^a + \sum_a \frac{g_a}{|\Lambda|} s_\Lambda^{a2}, \quad \Lambda\in\mathscr{L}, \tag{2.1}$$

of mean field type, where the f_a and g_a are real constants. Observe that more general interactions which are Hermitian in the a-indices can be transformed into the diagonal form (2.1). If, beside the temperature, other thermodynamic contact variables as, e.g., some chemical potentials or the global angular velocity, are fixed (by contacting appropriate reservoirs) one introduces the reduced Hamiltonians

$$H_\Lambda^r := H_\Lambda - D_\Lambda, \quad \Lambda\in\mathscr{L}, \tag{2.2}$$

where $D_\Lambda = D_\Lambda^*\in\mathscr{A}_\Lambda$ is the sum of the subtraction terms with fixed contact parameters and with $[H_\Lambda, D_\Lambda] = 0$, for all $\Lambda\in\mathscr{L}$. In the relevant cases the subtraction terms do not destroy the invariance of the Hamiltonian under local lattice permutations and are additive in Λ [cf. (2.20)].

In order to perform the thermodynamic limit we introduce the quasilocal algebra

$$\mathscr{A} := \bigotimes_{i\in\mathbb{Z}^d} \mathscr{A}_i \tag{2.3}$$

by means of the infinite C^*-tensor product,[10] which is a unital, simple, separable, antiliminary algebra with trivial center, every nontrivial representation of which is faithful.[11] After embedding the union of the localized algebras,

$$\mathscr{A}_0 := \bigcup_{\Lambda\in\mathscr{L}} \mathscr{A}_\Lambda \tag{2.4}$$

is a norm-dense $*$-subalgebra of $\mathscr{A}$.

The local grand canonical equilibrium states

$$\langle \psi_\Lambda^\beta; A\rangle := \mathrm{tr}_\Lambda\{\sigma_\Lambda^\beta A\}, \quad A\in\mathscr{A}_\Lambda,$$
$$\sigma_\Lambda^\beta := \exp(-\xi_\Lambda - \beta H_\Lambda^r), \tag{2.5}$$

where tr_Λ denotes the usual unnormalized trace in $\mathscr{H}_\Lambda$, and $\xi_\Lambda\in\mathbb{R}$, constitute, after extension to $\mathscr{A}$, a net in the state space $\mathscr{S}(\mathscr{A})$ of all states on $\mathscr{A}$. Since $\mathscr{S}(\mathscr{A})$ is compact in the weak-$*$-topology this net has limit points from which we select once and for all the state ω^β. The separability of $\mathscr{A}$ implies that the weak-$*$-topology of $\mathscr{S}(\mathscr{A})$ is metrizable and, hence, we may choose a subsequence of local equilibrium states with

$$\text{w-*-}\lim_n \psi_{\Lambda n}^\beta = \omega^\beta. \tag{2.6}$$

From this it follows that ω^β is invariant under all symmetry operations—acting as Jordan automorphisms in $\mathscr{A}$—which leave all H_Λ^r, $\Lambda\in\mathscr{L}$, unchanged, especially under all finite lattice permutations. Because the permutations act in a strongly asymptotic abelian manner on $\mathscr{A}$, the set of all permutation invariant states $\mathscr{S}^P\subset\mathscr{S}(\mathscr{A})$ constitutes a simplex, the extreme points of which are just the product states, uniquely given by a density matrix ρ in the one-lattice-point Hilbert space $\mathscr{H}$.[12] These product states are factorial and are disjoint if the associated ρ's are different. From this one concludes that the unique extremal decomposition of ω^β within $\mathscr{S}^P$ coincides with the central decomposition

$$\omega^\beta = \int_{\mathscr{T}^\beta} \varphi\, d\mu^\beta(\varphi). \tag{2.7}$$

Here the central measure μ^β on $\mathscr{S}(\mathscr{A})$ is supported by the weak-$*$-compact set $\mathscr{T}^\beta$ (and not only pseudosupported) in virtue of the weak-$*$-separability of $\mathscr{S}(\mathscr{A})$.

According to the spatial decomposition theory (Ref. 1, Chap. 4.4) the central decomposition (2.7) implies a direct integral decomposition of the GNS-triple associated with ω^β

$$(\pi_\beta, \mathscr{H}_\beta, \Omega_\beta) = \int_{\mathscr{T}^\beta}^{\oplus} (\pi_\varphi, \mathscr{H}_\varphi, \Omega_\varphi) d\mu^\beta(\varphi) \tag{2.8}$$

into the GNS-triples associated with $\varphi\in\mathscr{T}^\beta$ and of the von Neumann algebra

$$\mathscr{M}^\beta := \pi_\beta(\mathscr{A})'' = \int_{\mathscr{T}^\beta}^{\oplus} \mathscr{M}^\varphi\, d\mu^\beta(\varphi), \tag{2.9}$$

where $\mathscr{M}^\varphi := \pi_\varphi(\mathscr{A})''$ in $\mathscr{H}_\varphi$. The center $\mathscr{Z}^\beta := \mathscr{M}^\beta\cap\mathscr{M}^{\beta\prime}$ is the image of the W^*-isomorphism

$$\kappa^\beta: \mathscr{L}^\infty(\mathscr{T}^\beta, \mu^\beta) \to \mathscr{Z}^\beta \tag{2.10}$$

canonically defined by the property

$$(\Omega_\beta, \kappa^\beta(f)\pi_\beta(A)\Omega_\beta) = \int_{\mathscr{T}^\beta} f(\varphi)\hat{A}(\varphi) d\mu^\beta(\varphi) \tag{2.11}$$

for all $f\in\mathscr{L}^\infty(\mathscr{T}^\beta, \mu^\beta)$ and all $A\in\mathscr{A}$, where

$$\hat{A}(\varphi) := \langle \varphi; A\rangle \tag{2.12}$$

is the weak-$*$-continuous affine function connected with $A\in\mathscr{A}$ (Ref. 1, Chap. 4.1.3). From Ref. 1, Chap. 2.6 it follows, that $\mathscr{Z}^\beta$ consists just of the "observables at infinity".

Let us henceforth identify $\mathscr{A}$ with the faithfully represented algebra $\pi_\beta(\mathscr{A})$ and drop the symbol π_β for this representation morphism. (In some other instances we even drop the symbols for other representations of $\mathscr{A}$ and hope that the context prevents confusion.)

It is useful to introduce also for $\mathscr{M}^\beta$ a quasilocal structure by defining for every $\Lambda\in\mathscr{L}$ the von Neumann algebra

$$\mathscr{M}_\Lambda^\beta := \int_{\mathscr{T}^\beta}^{\oplus} \mathscr{M}_\Lambda^\varphi\, d\mu^\beta(\varphi) \subset \mathscr{M}^\beta \tag{2.13}$$

with

$$\mathscr{M}_\Lambda^\varphi := \pi_\varphi(\mathscr{A}_\Lambda) \subset \mathscr{M}^\varphi.$$

From the definition it is seen that

$$\mathscr{M}_\Lambda^\beta \equiv \mathscr{A}_\Lambda \bar{\otimes} \mathscr{L}^\beta \equiv \mathscr{L}^\infty(\mathscr{T}^\beta, \mu^\beta; \mathscr{A}_\Lambda), \tag{2.14}$$

where $\equiv$ denotes here W^*-isomorphy, $\bar{\otimes}$ the W^*-tensor product,[13] and the last symbol stands for the set of all μ^β-measurable, essentially bounded functions from $\mathscr{T}^\beta$ into $\mathscr{A}_\Lambda$. A typical element $M \in \mathscr{M}_\Lambda^\beta$ will always be written in the form

$$M = \int_{\mathscr{T}^\beta}^{\oplus} \pi_\varphi(A^\varphi) d\mu^\beta(\varphi) \in \mathscr{M}_\Lambda^\beta, \tag{2.15}$$

which is to imply that $\varphi \to A^\varphi$ is in $\mathscr{L}^\infty(\mathscr{T}^\beta, \mu^\beta; \mathscr{A}_\Lambda)$. Since the positive linear functional on $\mathscr{M}_\Lambda^\beta$

$$\mathrm{tr}_\Lambda^\beta\{M\} := \int_{\mathscr{T}^\beta} \mathrm{tr}_\Lambda\{A^\varphi\} d\mu^\beta(\varphi) \tag{2.16}$$

has the trace property, $\mathscr{M}_\Lambda^\beta$ is of type I. If $\Lambda \cap \Lambda' = \phi$, then $\mathscr{M}_\Lambda^\beta$ and $\mathscr{M}_{\Lambda'}^\beta$ commute. The union

$$\mathscr{M}_0^\beta := \bigcup_{\Lambda \in \mathscr{L}} \mathscr{M}_\Lambda^\beta \tag{2.17}$$

is σ-weakly dense in $\mathscr{M}^\beta$.

It should not lead to any confusion if we denote the extension of $\omega^\beta \in \mathscr{S}(\mathscr{A})$ to a normal state on $\mathscr{M}^\beta$, given by the cyclic vector $\Omega_\beta \in \mathscr{H}_\beta$, by the same symbol. Since the localized algebras $\mathscr{A}_\Lambda$ are finite-dimensional every restriction of $\varphi \in \mathscr{S}(\mathscr{A})$ to an $\mathscr{A}_\Lambda$ is normal and may be expressed by means of a density matrix $\rho_\Lambda^\varphi \in \mathscr{A}_\Lambda$, such that the latter depend continuously on φ (in the weak-*-topology). Introducing

$$\rho_\Lambda^\beta := \int_{\mathscr{T}^\beta}^{\oplus} \pi_\varphi(\rho_\Lambda^\varphi) d\mu^\beta(\varphi) \in \mathscr{M}_\Lambda^\beta, \tag{2.18}$$

we have for $M \in \mathscr{M}_\Lambda^\beta$

$$\langle \omega^\beta; M \rangle = \mathrm{tr}_\Lambda^\beta\{\rho_\Lambda^\beta M\}. \tag{2.19}$$

In the usual applications the subtraction terms D_Λ are additive in Λ, thus we postulate

$$D_{\Lambda \cup \Lambda'} = D_\Lambda + D_{\Lambda'}, \quad \text{if } \Lambda \cap \Lambda' = \phi. \tag{2.20}$$

Then, for $A \in \mathscr{A}_0$ and $\theta \in \mathbb{R}$, the definition

$$\alpha_\theta(A) := \exp(i\theta D_\Lambda) A \exp(-i\theta D_\Lambda) \tag{2.21}$$

if $A \in \mathscr{A}_\Lambda$, leads to a well-defined *-automorphism group on $\mathscr{A}_0$ and also gives rise to a C^*-dynamical system $(\mathscr{A}, \alpha, \mathbb{R})$. The restrictions of the dual transformations α_θ^* to $\mathscr{S}(\mathscr{A})$ are denoted by ν_θ, $\theta \in \mathbb{R}$. Since $\alpha_\theta(\sigma_\Lambda^\beta) = \sigma_\Lambda^\beta$, for all $\Lambda \in \mathscr{L}$ we have $\nu_\theta(\omega^\beta) = \omega^\beta$ and the defining relation

$$W_\theta^\beta A \Omega_\beta := \alpha_\theta(A) \Omega_\beta, \quad A \in \mathscr{A}, \ \theta \in \mathbb{R} \tag{2.22}$$

gives rise to a strongly continuous group of unitary operators W_θ^β in $\mathscr{H}_\beta$, which in turn leads to the W^*-dynamical system $(\mathscr{M}^\beta, \alpha^\beta, \mathbb{R})$, where

$$\alpha_\theta^\beta(M) := W_\theta^\beta M W_{-\theta}^\beta, \quad M \in \mathscr{M}^\beta, \ \theta \in \mathbb{R}. \tag{2.23}$$

We even have need for the further extended transformations $\tilde{\alpha}_\theta^\beta \in \mathrm{Aut}(\mathscr{B}(\mathscr{H}_\beta))$ which are also given by (2.23) with $M \in \mathscr{B}(\mathscr{H}_\beta)$.

Proposition 2.1: Denoting for $f \in \mathscr{L}^\infty(\mathscr{S}(\mathscr{A}), \mu^\beta)$ $(f \circ \nu_\theta)(\varphi) := f(\nu_\theta(\varphi))$, $\varphi \in \mathscr{S}(\mathscr{A})$, it holds that

$$\mu^\beta(f \circ \nu_\theta) = \mu^\beta(f) \tag{2.24}$$

$$\theta \in \mathbb{R}, \quad f \in \mathscr{L}^\infty(\mathscr{S}(\mathscr{A}), \mu^\beta)$$

and

$$\alpha_\theta^\beta(\kappa^\beta(f)) = \kappa^\beta(f \circ \nu_\theta) \tag{2.25}$$

entailing

$$\nu_\theta(\mathscr{T}^\beta) = \mathscr{T}^\beta. \tag{2.26}$$

Proof: (i) Let P denote the orthogonal projection onto the closed subspace $\overline{\mathscr{Z}^\beta \Omega_\beta} \subset \mathscr{H}_\beta$. Since $Z \in \mathscr{Z}^\beta$ implies $\alpha_\theta^\beta(Z) \in \mathscr{Z}^\beta$, $\forall \theta \in \mathbb{R}$, we have for every $C \in \mathscr{A}$

$$\begin{aligned}(C\Omega_\beta, \tilde{\alpha}_\theta^\beta(P) Z \Omega_\beta) &= (\alpha_{-\theta}(C)\Omega_\beta, P\alpha_{-\theta}^\beta(Z)\Omega_\beta) \\ &= (C\Omega_\beta, Z\Omega_\beta).\end{aligned}$$

This leads to $\tilde{\alpha}_\theta^\beta(P) Z\Omega_\beta = Z\Omega_\beta$, $\forall \theta \in \mathbb{R}$, and this in turn to $\tilde{\alpha}_\theta^\beta(P) = P$, $\forall \theta \in \mathbb{R}$.

(ii) Equation (2.12) implies

$$\alpha_\theta(A) = \hat{A} \circ \nu_\theta, \quad \forall A \in \mathscr{A}.$$

According to the general theory for orthogonal measures on $\mathscr{S}(\mathscr{A})$ (Ref. 1, Chap. 4.1.3) we obtain the first and last equality in the following calculation, where $A_i \in \mathscr{A}$, for $i = 1, \ldots, n$,

$$\begin{aligned}\mu^\beta(\hat{A}_1 \cdots \hat{A}^n) &= (\Omega_\beta, A_1 P A_2 \cdots P A_n \Omega_\beta) \\ &= (\Omega_\beta, \tilde{\alpha}_\theta^\beta(A_1 P \cdots P A_n)\Omega_\beta) \\ &= (\Omega_\beta, \alpha_\theta(A_1) P \cdots P \alpha_\theta(A_n)\Omega_\beta) \\ &= \mu^\beta(\hat{A}_1 \circ \nu_\theta \cdots \hat{A}_n \circ \nu_\theta). \qquad (*)\end{aligned}$$

Here we have also used the invariance of Ω_β under W_θ^β and (i). Since the $\hat{A}$, $A \in \mathscr{A}$, separate points of $\mathscr{S}(\mathscr{A})$, the polynomials thereof are norm-dense in $\mathscr{C}(\mathscr{S}(\mathscr{A}))$ according to the Stone–Weierstrass theorem and, thus, $\sigma(\mathscr{L}^\infty, \mathscr{L}_1)$-dense in $\mathscr{L}^\infty(\mathscr{S}(\mathscr{A}), \mu^\beta) =: \mathscr{L}^\infty$. The mapping $f \to f \circ \nu_\theta$ defines a *-automorphism of $\mathscr{L}^\infty(\mathscr{S}(\mathscr{A}), \mu^\beta)$ and, therefore, is σ-weakly continuous. $\mu_\theta^\beta(f) := \mu^\beta(f \circ \nu_\theta)$ is then a normal state on $\mathscr{L}^\infty(\mathscr{S}(\mathscr{A}), \mu^\beta)$ and uniquely determined by its values on the polynomials of the $\hat{A}$. Thus our last relation in (*) may be extended to (2.24).

(iii) Let us calculate for arbitrary $C_1, C_2 \in \mathscr{A}$ and $f \in \mathscr{L}^\infty(\mathscr{S}(\mathscr{A}), \mu^\beta)$

$$\begin{aligned}&(C_1\Omega_\beta, \alpha_\theta^\beta(\kappa^\beta(f)) C_2 \Omega_\beta) \\ &\quad = \mu^\beta(\widehat{f C_1^* C_2} \circ \nu_{-\theta}) \\ &\quad \overset{\text{(ii)}}{=} \mu^\beta((f \circ \nu_\theta)\widehat{C_1^* C_2}) \\ &\quad = (C_1\Omega_\beta, \kappa^\beta(f \circ \nu_\theta) C_2 \Omega_\beta),\end{aligned}$$

which gives (2.25).

(iv) Since for the characteristic function $\chi_{\mathscr{T}^\beta}$ of $\mathscr{T}^\beta \subset \mathscr{S}(\mathscr{A})$

$$\mu^\beta(\chi_{\mathscr{T}^\beta}) \overset{\text{(ii)}}{=} \mu^\beta(\chi_{\mathscr{T}^\beta} \circ \nu_\theta) = \mu^\beta(\chi_{\mathscr{T}^\beta_{-\theta}}) = 1$$

and $\mathscr{T}^\beta_{-\theta} := \nu_{-\theta}(\mathscr{T}^\beta)$ is weakly-*-closed, $\nu_{-\theta}$ being weak-*-continuous, we have $\mathscr{T}^\beta_{-\theta} \supset \mathscr{T}^\beta$, for all $\theta \in \mathbb{R}$, and obtain by means of the inverse relation (2.36). □

Since D_Λ is in many cases a linear combination of local particle number operators, let us call α_θ and the modifications thereof "generalized gauge transformations."

Definition 2.2: We say that the symmetry under gauge transformations is spontaneously broken in ω^β if there is at least one pure phase state $\varphi \in \mathcal{T}^\beta$ (the support of the central decomposition of ω^β) with $\nu_\theta(\varphi) \neq \varphi$, for some $\theta \in \mathbb{R}$.

Observation 2.3: If the symmetry under gauge transformations is spontaneously broken, then $\mathcal{T}_\beta$ is an overcountable set.

In fact: In this case there is an $A \in \mathcal{A}$, a $\theta \in \mathbb{R}$, and a $\varphi \in \mathcal{T}^\beta$ with $\langle \nu_\theta(\varphi); A \rangle \neq \langle \varphi; A \rangle$. But, if θ' varies from 0 to θ, the two expectation values are continuously interpolated in virtue of the weak-*-continuity of $\theta' \to \nu_{\theta'}$. Thus, $\{\nu_{\theta'}(\varphi);$ $\theta' \in [0,\theta]\} \subset \mathcal{T}^\beta$ contains overcountably many different states.

The subsequent investigations are only nontrivial if the symmetry under gauge transformations is spontaneously broken in ω^β, but many of them are formally valid also in the other case. If $\varphi \to A^\varphi$ is in $\mathcal{L}^\infty(\mathcal{T}^\beta, \mu^\beta; \mathcal{A}_\Lambda)$, then also $\varphi \to A^{\varphi\theta}$ is so, for all $\theta \in \mathbb{R}$, if we denote

$$\varphi\theta := \nu_\theta(\varphi), \quad \varphi \in \mathcal{S}(\mathcal{A}). \tag{2.27}$$

Thus for $M \in \mathcal{M}_\Lambda^\beta$ the transformed element

$$T_\theta(M) := \int_{\mathcal{T}^\beta} \pi_\varphi(A^{\varphi\theta}) d\mu^\beta(\varphi) \tag{2.28}$$

is also in $\mathcal{M}_\Lambda^\beta$, for $\theta \in \mathbb{R}$. In virtue of the point-wise action of the algebraic operations in an integral von Neumann algebra, T_θ is a *-automorphism of $\mathcal{M}_\Lambda^\beta$, $\forall \Lambda \in \mathcal{L}$, and also of $\mathcal{M}_0^\beta$. $\theta \to T_\theta(M)$ is continuous in the σ-weak topology by means of the dominated convergence theorem. Finally, $\theta \to T_\theta$ is a representation of $\mathbb{R}$. All these properties can be proved also for the following set of transformations in $\mathcal{M}_0^\beta$:

$$\hat{\alpha}_\theta(M) := \int_{\mathcal{T}^\beta}^{\oplus} \pi_\varphi(\alpha_\theta(A^\varphi)) d\mu^\beta(\varphi), \quad \theta \in \mathbb{R}, \tag{2.29}$$

which we introduce mainly for pedagogical reasons.

Proposition 2.4: The group of gauge transformations has on $\mathcal{M}_0^\beta$ the form

$$\alpha_\theta^\beta = \hat{\alpha}_\theta \circ T_\theta = T_\theta \circ \hat{\alpha}_\theta. \tag{2.30}$$

If these symmetry transformations are spontaneously broken in ω^β then $\hat{\alpha}_\theta$ and T_θ for themselves are not σ-weakly continuous mappings in $\mathcal{M}_0^\beta$ (and thus not extendable to W^*-automorphisms of $\mathcal{M}^\beta$) for those $\theta \in \mathbb{R}$, for which $T_\theta \neq 1$.

Proof: (i) The commutativity of $\hat{\alpha}_\theta$ and T_θ can be verified directly.

(ii) For arbitrary $f \in \mathcal{L}^\infty(\mathcal{T}^\beta, \mu^\beta)$ and $A \in \mathcal{A}_\Lambda$ we calculate by means of (2.18), (2.19) and by means of the invariance of ω^β and μ^β under gauge transformations

$$\begin{aligned}
&\int_{\mathcal{T}^\beta} \mathrm{tr}_\Lambda \{\alpha_{-\theta}(\rho_\Lambda^\varphi) A\} f(\varphi) d\mu^\beta(\varphi) \\
&= \int_{\mathcal{T}^\beta} \mathrm{tr}_\Lambda \{\rho_\Lambda^\varphi \alpha_\theta(A)\} f(\varphi) d\mu^\beta(\varphi) \\
&= \langle \omega^\beta; \kappa^\beta(f) \alpha_\theta^\beta(A) \rangle \\
&= \langle \omega^\beta; \alpha_{-\theta}^\beta(\kappa^\beta(f)) A \rangle \\
&= \int_{\mathcal{T}^\beta} \mathrm{tr}_\Lambda \{\rho_\Lambda^\varphi A\} f(\nu_{-\theta}(\varphi)) d\mu^\beta(\varphi) \\
&= \int_{\mathcal{T}^\beta} \mathrm{tr}_\Lambda \{\rho_\Lambda^{\varphi\theta} A\} f(\varphi) d\mu^\beta(\varphi).
\end{aligned}$$

This gives

$$\alpha_{-\theta}(\rho_\Lambda^\varphi) = \rho_\Lambda^{\varphi\theta}, \quad \mu^\beta, \text{ a.e.} \tag{2.31}$$

(iii) For $C_1, C_2 \in \mathcal{A}$ and $M \in \mathcal{M}_\Lambda^\beta$ we obtain

$$\begin{aligned}
&(C_1\Omega_\beta, \alpha_{-\theta}^\beta(M) C_2\Omega_\beta) \\
&= \int_{\mathcal{T}^\beta} \mathrm{tr}_\Lambda \{\rho_\Lambda^\varphi \alpha_\theta(C_1^*) A^\varphi \alpha_\theta(C_2)\} d\mu^\beta(\varphi) \\
&\overset{\text{(ii)}}{=} \int_{\mathcal{T}^\beta} \mathrm{tr}^\Lambda \{\rho_\Lambda^{\varphi\theta} C_1^* \alpha_{-\theta}(A^\varphi) C_2\} d\mu^\beta(\varphi) \\
&\overset{(2.24)}{=} (C_1\Omega_\beta, \hat{\alpha}_{-\theta} \circ T_{-\theta}(M) C_2\Omega_\beta),
\end{aligned}$$

which leads to (2.30).

(iv) Take a $M \in \mathcal{M}_\Lambda^\beta$, not invariant under T_θ, and approximate it by a net $\{A_\alpha; \alpha \in I\} \subset \mathcal{A}_0$ in the strong operator topology of $\mathcal{H}_\beta$. Then $\lim T_\theta(A_\alpha) = M \neq T_\theta(M)$ and $\lim_\alpha \hat{\alpha}_\theta(A_\alpha) = \lim_\alpha \alpha_\theta^\beta(A_\alpha) = \alpha_\theta^\beta(M) \neq \hat{\alpha}_\theta(M)$, as may be seen from (2.30). □

Let us denote the σ-weak generators (Ref. 1, Chap. 3) of α^β and T by δ_α and δ_T, respectively, and introduce the domain

$$\begin{aligned}
\mathcal{M}_\infty^\beta := \{&M \in \mathcal{M}_0^\beta; \text{ for all } n \in \mathbb{N}, d^n A^{\varphi\theta}/d\theta^n \\
&\text{is defined and bounded in } \varphi \text{ for } \mu^\beta, \\
&\text{a.a. } \varphi \in \mathcal{T}^\beta\}
\end{aligned} \tag{2.32}$$

which is a σ-weakly dense sub-*-algebra of $\mathcal{M}^\beta$ containing $\mathcal{A}_0$.

Proposition 2.5: $\mathcal{M}_\infty^\beta$ is in the domains of δ_α^n and δ_T^n for all $n \in \mathbb{N}$. For $M \in \mathcal{M}_\infty^\beta$, $M \in \mathcal{M}_\infty^\beta \cap \mathcal{M}_\Lambda^\beta$ say, it holds

$$\delta_\alpha(M) = [D_\Lambda, M] + \delta_T(M). \tag{2.33}$$

Proof: Let $\varphi \to A^{\varphi\theta}$ be the function associated with $T_\theta(M)$, $M \in \mathcal{M}_\infty^\beta$. Then $dA^{\varphi\theta}/d\theta$ exists for μ^β, a.a. $\varphi \in \mathcal{T}^\beta$ and is μ^β, a.e. bounded according to the assumption. Then for a normal state ψ on $\mathcal{M}^\beta$

$$\begin{aligned}
&\lim_{\theta \to 0} \langle \psi; T_\theta(M) - M \rangle / \theta \\
&= \lim_{\theta \to 0} \int_{\mathcal{T}^\beta} (\langle \psi(\varphi); A^{\varphi\theta} - A^\varphi \rangle / \theta) d\mu^\beta(\theta) \\
&= \int_{\mathcal{T}^\beta} \langle \psi(\varphi); dA^{\varphi\theta}/d\theta \rangle d\mu^\beta(\theta),
\end{aligned} \tag{*}$$

where the last step follows from the dominated convergence theorem, and $\psi(\varphi)$ are the components of the canonical decomposition of ψ into normal states on $\mathcal{M}_\varphi$, $\varphi \in \mathcal{T}^\beta$ (Ref. 13, Prop. 8.34).

Thus the limit in (*) exists and gives $\delta_T(M)$ by definition. This may be iterated and leads to the existence of $\delta_T^n(M)$, for all $n \in \mathbb{N}$. If $M \in \mathcal{M}_\infty^\beta$ then also $[D_\Lambda, M] \in \mathcal{M}_\infty^\beta$. Differentiation of (2.30) leads to (2.33). □

III. THE LIMITING HEISENBERG GENERATOR

Defining $m_\Lambda^a := s_\Lambda^a / |\Lambda|$, it is well known that for every permutation invariant state $\varphi \in \mathcal{S}^P(\mathcal{A})$

$$\text{s-lim}_\Lambda\, m_\Lambda^a =: m_\varphi^a \in \mathcal{Z}^\varphi := \mathcal{M}^\varphi \cap \mathcal{M}^{\varphi'} \tag{3.1}$$

exists in the GNS–Hilbert space $\mathcal{H}_\varphi$ (cf., e.g., Ref. 6).

Proposition 3.1: For every $\varphi \in \mathcal{S}^P(\mathcal{A})$ it holds in $\mathcal{H}_\varphi$ for $A \in \mathcal{A}_\Lambda$ [embedded into $\mathcal{B}(\mathcal{H}_\varphi)$!]

$$\text{s-}\lim_{\Lambda'}[H_{\Lambda'},A] = :\text{s-}\lim_{\Lambda'} L_{\Lambda'}(A) = [H^{\varphi}_{\Lambda},A], \tag{3.2}$$

with

$$H^{\varphi}_{\Lambda} := \sum_a f_a s^a_{\Lambda} + \sum_a 2g_a m^a_{\varphi} s^a_{\Lambda}. \tag{3.3}$$

The expectations of (3.2) in the states φ converge uniformly in $\varphi \in \mathcal{S}^{\beta}(\mathcal{A})$.

Proof: We have for $A \in \mathcal{A}_{\Lambda}$, $\Lambda \subset \Lambda'$,

$$[H_{\Lambda'} - H^{\varphi}_{\Lambda'},A] = \left[\sum_a g_a s^{a2}_{\Lambda},A\right]/|\Lambda'| + 2\left[\sum_a g_a s^a_{\Lambda},A\right]\left(\frac{s^a_K}{|\Lambda'|} - m^a_{\varphi}\right),$$

where $K := \Lambda' \setminus \Lambda$; this tends to zero in the strong operator topology of $\mathcal{H}_{\varphi}$. The expectation in φ is then estimated by

$$\begin{aligned}|\langle\varphi;[H_{\Lambda'} - H^{\varphi}_{\Lambda'},A]\rangle| &\leqslant C_1(A)/|\Lambda'| \\ &\quad + C_2(A)\left(1 - \frac{|K|}{|K| + |\Lambda|}\right)\sum_a |\langle\varphi;s^a_i\rangle|,\end{aligned}$$

where

$$C_1(A) := \left|\left|\left[\sum_a g_a s^{a2}_{\Lambda},A\right]\right|\right|,$$

$$C_2(A) := \left|\left|2\left[\sum_a g_a s^a_{\Lambda},A\right]\right|\right|,$$

and $i \in \mathbb{Z}^d$ is arbitrary. Since $|\langle\varphi_i s^a_i\rangle| \leqslant 1$, the convergence is uniform in $\varphi \in \mathcal{S}^{\beta}(\mathcal{A})$. □

In the case of ω^{β}, (3.2) gives rise to the local form of the limiting Heisenberg generator

$$L^{\beta}_0(A) := [H^{\beta}_{\Lambda},A], \quad A \in \mathcal{A}_{\Lambda}, \tag{3.4}$$

which is a well-defined antisymmetric $*$-derivation of $\mathcal{M}^{\beta}$ with domain $\mathcal{A}_0$. In order to obtain an explicit expression for the closure of (3.4) we make use of the results in Ref. 14. By means of correlation inequalities for the generators of the KMS-dynamics, which here is generated by the reduced Hamiltonians, one can show that for $\varphi \in \mathcal{T}^{\beta}$ the local density matrices of (2.18) have the form

$$\rho^{\varphi}_{\Lambda} = \exp(-\zeta^{\varphi}_{\Lambda} - \beta H^{\varphi r}_{\Lambda}) \tag{3.5}$$

with

$$H^{\varphi r}_{\Lambda} := H^{\varphi}_{\Lambda} - D_{\Lambda}, \tag{3.6}$$

where $\zeta^{\varphi}_{\Lambda} \in \mathbb{R}$ is given from the normalization condition and H^{φ}_{Λ} from (3.3), where for $\varphi \in \mathcal{T}^{\beta}$ the center $\mathcal{Z}^{\varphi}$ is trivial and m^a_{φ} will be identified with a c-number. Thus we may identify H^{φ}_{Λ} with an element of $\mathcal{A}_{\Lambda}$ for all $\Lambda \in \mathcal{L}$. Conforming to this interpretation and in virtue of the dominated convergence theorem we may write (3.1) and (3.3) for the special case $\varphi = \omega^{\beta}$ as

$$m^a_{\beta} = \int^{\oplus}_{\mathcal{T}^{\beta}} m^a_{\varphi} 1_{\varphi}\, d\mu^{\beta}(\varphi), \quad H^{\beta}_{\Lambda} = \int^{\oplus}_{\mathcal{T}^{\beta}} \pi_{\varphi}(H^{\varphi}_{\Lambda}) d\mu^{\beta}(\varphi). \tag{3.7}$$

Application of (2.18) then gives

$$\rho^{\beta}_{\Lambda} = \exp(-\zeta^{\beta}_{\Lambda} - \beta H^{\beta r}_{\Lambda}), \quad \zeta^{\beta}_{\Lambda} \in \mathcal{Z}^{\beta}. \tag{3.8}$$

Using first this explicit expression and then the invariance of ω^{β} under gauge transformations one obtains for $M \in \mathcal{M}^{\beta}_{\infty} \cap \mathcal{M}^{\beta}_{\Lambda}$

$$\langle\omega^{\beta};[H^{\beta}_{\Lambda},M]\rangle = \langle\omega^{\beta};[D_{\Lambda},M]\rangle = -\langle\omega^{\beta};\delta_T(M)\rangle. \tag{3.9}$$

Proposition 3.2: L^{β}_0 is σ-weakly closable with a norm-densely defined pre-adjoint L^{β}_* and it holds

$$L^{\beta} := \overline{L^{\beta}_0}^{\sigma-w} = (L^{\beta}_*)^*. \tag{3.10}$$

The domain $\mathcal{D}(L^{\beta})$ contains $\mathcal{M}^{\beta}_{\infty}$ and there one has

$$L^{\beta}(M) = [H^{\beta}_{\Lambda},M] + \delta_T(M), \quad M \in \mathcal{M}^{\beta}_{\infty} \cap \mathcal{M}^{\beta}_{\Lambda}. \tag{3.11}$$

Proof: (i) Define for $C \in \mathcal{A}_0$ and $M \in \mathcal{M}^{\beta}$

$$\langle\omega^{\beta}_C;M\rangle := \langle\omega^{\beta};C^*MC\rangle/\langle\omega^{\beta};C^*C\rangle, \tag{3.12}$$

whenever the rhs exists. The set of finite linear combinations of these states is norm-dense in $\mathcal{M}^{\beta}_*$. For $A \in \mathcal{A}_{\Lambda}$ one obtains for the adjoint $L^{\beta *}_0$ the relation

$$\begin{aligned}\langle L^{\beta *}_0(\omega^{\beta}_C);A\rangle &:= \langle\omega^{\beta}_C;L^{\beta}_0(A)\rangle \\ &\overset{(3.9)}{=} -\langle\omega^{\beta};[H^{\beta}_{\Lambda},C^*]AC + C^*A[H^{\beta}_{\Lambda},C]\rangle/\langle\omega^{\beta};C^*C\rangle.\end{aligned}$$

This defines uniquely a functional in $\mathcal{M}^{\beta}_*$. Thus, $L^{\beta}_* := L^{\beta *}_0|\mathcal{M}^{\beta}_*$ is a norm-densely and, therefore, $\sigma(\mathcal{M}^{\beta}_*,\mathcal{M}^{\beta})$-densely, defined linear map $\mathcal{M}^{\beta}_* \to \mathcal{M}^{\beta}_*$. Then $(L^{\beta}_*)^*$ exists and is σ-weakly closed (cf. Ref. 1, Lemma 3.1.9). Since $L^{\beta}_0 \subset (L^{\beta}_*)^*$, L^{β}_0 is σ-weakly closable with closure L^{β}. Again using Lemma 3.1.9 of Ref. 1 we know that the densely defined closed operator L^{β} has a densely defined pre-adjoint and coincides with $(L^{\beta}_*)^*$.

(ii) For $M \in \mathcal{M}^{\beta}_{\infty} \cap \mathcal{M}^{\beta}_{\Lambda}$ and $C \in \mathcal{A}_0$ we calculate

$$\begin{aligned}\langle\omega^{\beta}_C;L^{\beta}(M)\rangle &= \langle L^{\beta}_*(\omega^{\beta}_C);M\rangle \\ &\overset{(i)}{=} -\langle\omega^{\beta};[H^{\beta}_{\Lambda},C^*]MC + C^*M[H^{\beta}_{\Lambda},C]\rangle/\langle\omega^{\beta};C^*C\rangle \\ &= \langle\omega^{\beta};C^*[H^{\beta}_{\Lambda},M]C\rangle/\langle\omega^{\beta};C^*C\rangle \\ &\quad - \langle\omega^{\beta};[H^{\beta}_{\Lambda},C^*MC]\rangle/\langle\omega^{\beta};C^*C\rangle \\ &\overset{(3.9)}{=} \langle\omega^{\beta}_C;[H^{\beta}_{\Lambda},M]\rangle + \langle\omega^{\beta}_C;\delta_T(M)\rangle.\end{aligned}$$

Here we have assumed without restriction in generality that $C \in \mathcal{A}_{\Lambda}$. □

Proposition 3.3: Let $M \in \mathcal{D}(L^{\beta})$ and $\{A_{\alpha} \in \mathcal{A}_{\Lambda_{\alpha}}; \alpha \in I\}$ be a net with $\|A_{\alpha}\| \leqslant \|M\|$, $\forall \alpha \in I$, and $A_{\alpha} \to M$ in the strong operator topology of $\mathcal{M}^{\beta}$. (Such a net exists according to Kaplansky's density theorem.) Then

$$\sigma\text{-w}\lim_{\alpha} L^{\beta}(A_{\alpha}) = L^{\beta}(M) = \sigma\text{-w}\lim_{\alpha} L_{\Lambda_{\alpha}}(A_{\alpha}). \tag{3.13}$$

Proof: (i) Relation (3.10) implies that for every $M \in \mathcal{D}(L^{\beta})$ there is a net $\{A'_{\gamma} \in \mathcal{A}_{\Lambda_{\gamma}}; \gamma \in I'\}$ with $A'_{\gamma} \to M$ and $L^{\beta}(A'_{\gamma}) \to L^{\beta}(M)$ in the σ-weak topology of $\mathcal{B}(\mathcal{H}_{\beta})$. Thus the net $\{A_{\alpha} - A'_{\gamma}; (\alpha,\gamma) \in I \times I'\}$ tends σ-weakly to zero, and for $C \in \mathcal{A}_0$ [assume $N_C := \langle\omega^{\beta};C^*C\rangle > 0$]

$$\begin{aligned}&\langle\omega^{\beta}_C;L^{\beta}(A_{\alpha} - A'_{\gamma})\rangle \\ &\quad = -\langle\omega^{\beta};L^{\beta}(C^*)(A_{\alpha} - A'_{\gamma})C + C^*(A_{\alpha} - A'_{\gamma})L^{\beta}(C)\rangle/N_C\end{aligned}$$

tends to zero, which leads to the first equality in (3.13).

(ii) We estimate

$$|\langle\omega_C^\beta;L^\beta(M)-[H_{\Lambda_\alpha},A_\alpha]\rangle|$$
$$\leqslant|\langle\omega_C^\beta;L^\beta(M)-L^\beta(A_\alpha)\rangle|$$
$$+|\langle\omega_C^\beta;L^\beta(A_\alpha)-[H_{\Lambda_\alpha},A_\alpha]\rangle|,$$

where the first summand goes to zero by (i). By the use of (2.19) and (3.12) we write the second summand in the form

$$|\mathrm{tr}_{\Lambda_\alpha}^\beta\{C\rho_{\Lambda_\alpha}^\beta C^*[H_{\Lambda_\alpha}^\beta-H_{\Lambda_\alpha},A_\alpha]\}/N_C|$$
$$\leqslant|\mathrm{tr}_{\Lambda_\alpha}^\beta\{([C,H_{\Lambda_\alpha}^\beta-H_{\Lambda_\alpha}]\rho_{\Lambda_\alpha}^\beta C^*$$
$$+C\rho_{\Lambda_\alpha}^\beta[C^*,H_{\Lambda_\alpha}^\beta-H_{\Lambda_\alpha}])A_\alpha\}/N_C|$$
$$+|\mathrm{tr}_{\Lambda_\alpha}^\beta\{C[\rho_{\Lambda_\alpha}^\beta,H_{\Lambda_\alpha}^\beta-H_{\Lambda_\alpha}]C^*A_\alpha\}/N_C|.$$

Again using the trace property of $\mathrm{tr}_{\Lambda_\alpha}^\beta$ the first term can be written as the sum of two scalar products in $\mathscr{H}_\beta$ and is seen to tend to zero by (3.2). Since $H_{\Lambda_\alpha}^\beta-H_{\Lambda_\alpha}=H_{\Lambda_\alpha}^{\beta r}-H_{\Lambda_\alpha}^r$ and $H_{\Lambda_\alpha}^{\beta r}$ commutes with $\rho_{\Lambda_\alpha}^\beta$ the second term gives

$$\Big|([H_{\Lambda_\alpha}^r,\rho_{\Lambda_\alpha}^\beta]\rho_{\Lambda_\alpha}^{\beta-1}\Omega_\beta,(C^*A_\alpha C-C^*MC)\Omega_\beta)$$
$$+\int_{\mathscr{T}^\beta}\mathrm{tr}_{\Lambda_\alpha}\{\rho_{\Lambda_\alpha}^\varphi[H_{\Lambda_\alpha}^r,\overline{A}^\varphi]\}d\mu^\beta(\varphi)\Big|,$$

where $C^*MC=:\int_{\mathscr{T}^\beta}^\oplus\pi_\varphi(\overline{A}^\varphi)d\mu^\beta(\varphi)$. Here the first term is dominated by

$$\|[H_{\Lambda_\alpha}^r,\rho_{\Lambda_\alpha}^\beta]\rho_{\Lambda_\alpha}^{\beta-1}\Omega_\beta\|\,\|C^*\|\,\|(A_\alpha-M)C\Omega_\beta\|$$

in which the first norm expression is bounded according to the Appendix and the last norm converges to zero by our assumption. Since by Proposition 3.1 for every $\varphi\in\mathscr{T}^\beta$ the integrand tends to zero uniformly in φ the entire integral expression tends to zero, too, and this proves the second equality in (3.13). □

From the proof one has the impression that, in general, the convergence relations in (3.13) can hardly be improved. For special cases we show in Sec. IV that the σ-weak convergence is in fact a strong operator convergence.

There are only few and not easily applicable criteria for proving a closed derivation to generate a W^*-automorphism group. In virtue of the explicit form of L^β we can proceed here directly. The essential idea is to show that $\mathscr{M}_\infty^\beta$ is a core for the generator.

Lemma 3.4: The molecular field values m_φ^a for the pure phase states $\varphi\in\mathscr{T}^\beta$ satisfy the self-consistency relations in $\mathscr{H}_\varphi$

$$m_\varphi^a=\mathrm{tr}_i\{\rho_i^\varphi s_i^a\} \tag{3.14}$$

for every $\{i\}\equiv i\in\mathbb{Z}^d$ and $m_{\varphi\theta}^a$ is analytic in θ.

Proof: We have for $\varphi\in\mathscr{T}^\beta$

$$m_\varphi^a=\langle\varphi;m_\varphi^a 1_\varphi\rangle$$
$$=\lim_\Lambda\langle\varphi,s_\Lambda^a/|\Lambda|\rangle=\langle\varphi;s_i^a\rangle$$

for every $i\in\mathbb{Z}^d$ by permutation invariance and this may be written as in (3.14) by means of the local density matrix ρ_Λ^φ, $\Lambda=i$. Thus

$$m_{\varphi\theta}^a=\mathrm{tr}_i\{\rho_i^{\varphi\theta}s_i^a\}\overset{(2.31)}{=}\mathrm{tr}_i\{\alpha_{-\theta}(\rho_i^\varphi)s_i^a\}$$

which is analytic in θ by (2.21). □

Theorem 3.5: L^β is the generator of a σ-weakly continuous W^*-automorphism group $\{\tau_t^\beta;t\in\mathbb{R}\}\subset\mathrm{Aut}(\mathscr{M}^\beta)$ which leaves $\mathscr{M}_\Lambda^\beta\cap\mathscr{M}_\infty^\beta$ invariant, for all $\Lambda\in\mathscr{L}$. There it holds

$$\tau_t^\beta(M)$$
$$=\int_{\mathscr{T}^\beta}^\oplus\pi_\varphi(\exp(itH_\Lambda^{\varphi r})\exp(itD_\Lambda)A^{\varphi t}$$
$$\times\exp(-itD_\Lambda)\exp(-itH_\Lambda^{\varphi r}))d\mu^\beta(\varphi) \tag{3.15}$$

where, as before, $H_\Lambda^{\varphi r}:=H_\Lambda^\varphi-D_\Lambda$ and $\varphi t:=\nu_t(\varphi)$.

Proof: (i) For $M\in\mathscr{M}_\infty^\beta$ let us *define* the transformations $\tau_t^\beta(M)$ by the rhs of (3.15), which is a well-defined prescription since D_Λ and $H_\Lambda^{\varphi r}$ are additive in Λ and thus (3.15) does not depend on which $\mathscr{M}_\Lambda^\beta\cap\mathscr{M}_\infty^\beta$ is to contain M. From (2.31) and (3.5) we derive

$$\alpha_{-t}(H_\Lambda^\varphi)=H_\Lambda^{\varphi t},\quad\mu^\beta,\ \text{a.e.} \tag{3.16}$$

This gives

$$\exp(it'(H_\Lambda^\varphi-D_\Lambda))\exp(it'D_\Lambda)\exp(it(H_\Lambda^{\varphi t'}-D_\Lambda))\exp(itD_\Lambda)$$
$$=\exp(i(t'+t)(H_\Lambda^\varphi-D_\Lambda))\exp(i(t'+t)D_\Lambda)$$

which leads to the group property

$$\tau_{t'}^\beta\circ\tau_t^\beta=\tau_{t+t'}^\beta,\quad\forall t,t'\in\mathbb{R}, \tag{3.17}$$

if the iterated transformation is defined.

(ii) From (3.15) we also infer that τ_t^β leaves $\mathscr{M}_\Lambda^\beta$ invariant. If we replace in (3.15) in the argument of π_φ the symbol φ by $\varphi\theta$ and assume $M\in\mathscr{M}_\Lambda^\beta\cap\mathscr{M}_\infty^\beta$, and if we use (3.3) and (3.6) to conclude that $H_\Lambda^{\varphi\theta r}$ is analytic in θ with bounded derivatives, we arrive at the invariance of $\mathscr{M}_\infty^\beta$ under τ_t^β, for all $t\in\mathbb{R}$. By inspection we further observe that τ_t^β is a $*$-isomorphism of $\mathscr{M}_\infty^\beta$ for all $t\in\mathbb{R}$. In order to check the σ-weak continuity in t, we investigate $\langle\psi;\tau_t^\beta(M)\rangle$ for a normal state ψ on $\mathscr{M}^\beta$ and for $M\in\mathscr{M}_\Lambda^\beta\cap\mathscr{M}_\infty^\beta$. This expectation is an μ^β-integral over expectations which are continuous in t and uniformly bounded in $\varphi\in\mathscr{T}^\beta$.

(iii) If we calculate $\langle\omega^\beta;\tau_t^\beta(M)\rangle$, $M\in\mathscr{M}_\Lambda^\beta\cap\mathscr{M}_\infty^\beta$, by means of the local density operator ρ_Λ^β from (3.8) we observe the dropping out of the $H_\Lambda^{\varphi r}$-terms, which commute with the ρ_Λ^φ, and are left with $\langle\omega^\beta;\alpha_t^\beta(M)\rangle=\langle\omega^\beta;M\rangle$.

(iv) The combination of (i), (ii), and (iii) shows that the τ_t^β are extendable to elements of a σ-weakly continuous W^*-automorphism group of $\mathscr{M}^\beta$, designated by the same symbols.

(v) From the foregoing considerations it follows that $\mathscr{M}_\infty^\beta$ is left invariant by the τ_t^β, is σ-weakly dense in $\mathscr{M}^\beta$, and is in the domain of the generator of the dynamical group. Thus, $\mathscr{M}_\infty^\beta$ is a core for the σ-weak generator of this W^*-system (Ref. 1, Corollary 3.1.7). Since $\mathscr{M}_\infty^\beta$ contains $\mathscr{A}_0$ it is also a core for L^β [cf. (3.10)], and since the generator of the W^*-system coincides there with L^β it follows that L^β is, in fact, the generator of $\{\tau_t^\beta;t\in\mathbb{R}\}$. □

In the preceding discussion of the finite-time dynamical transformations τ_t^β we have used the disintegrated form of elements $M\in\mathscr{M}^\beta$ into the pure phase components to study the peculiarities of these transformations. In this way it is made evident, that the physical dynamics transforms the pure phase into each other and that its explicit formulation

requires the subtraction terms D_Λ, which locally do not show up before performing the thermodynamic limit, and neither show up in the infinitesimal generator L^β. They drop out only if the symmetry under gauge transformations is not spontaneously broken.

On the other hand, it is quite natural to introduce the KMS-automorphisms generated by the reduced Hamiltonian locally in the form

$$\tau_t^{\beta r}(M) := \exp(itH_\Lambda^{\beta r})M\exp(-itH_\Lambda^{\beta r}), \quad M\in\mathcal{M}_\Lambda^\beta. \tag{3.18}$$

With this (3.15) gives

$$\tau_t^\beta = \tau_t^{\beta r}\circ\alpha_t^\beta \tag{3.19}$$

and the group property of $\{\tau_t^\beta; t\in\mathbb{R}\}$ implies

$$\tau_t^{\beta r}\circ\alpha_{t'}^\beta = \alpha_{t'}^\beta\circ\tau_t^{\beta r}, \quad \forall t,t'\in\mathbb{R}. \tag{3.20}$$

IV. LOCAL APPROXIMATION OF THE LIMITING DYNAMICS

The crucial step for deriving the convergence of the local finite-time-transformations to the limiting dynamics from that of the Heisenberg generators is the following result.

Theorem 4.1: For all $A\in\mathcal{A}_0$ and all $n\in\mathbb{N}$ we have in the GNS-Hilbert space $\mathcal{H}_\beta$

$$\text{s-lim}_\Lambda\, L_\Lambda^n(A) = L^{\beta n}(A) \tag{4.1}$$

[where L_Λ is defined in (3.2) and L^β in (3.10)].

Proof: (i) Since $\{s_i^a; a = 1,2\cdots k^2\}$ is a basis in $\mathcal{A}_i$ there are $\lambda_{abc}\in\mathbb{C}$, independent of i, such that

$$[s_i^a, s_i^b]_- = \sum_c \lambda_{abc} s_i^c.$$

We calculate, with $m_\Lambda^a = s_\Lambda^a/|\Lambda|$,

$$L_\Lambda(s_i^b) = \sum_{a,c}\lambda_{abc}[f_a s_i^c + g_a(m_\Lambda^a s_i^c + s_i^c m_\Lambda^a)] \tag{4.2}$$

which gives by linear superposition

$$L_\Lambda(m_\Lambda^b) = \sum_{a,c}\lambda_{abc}[f_a m_\Lambda^c + g_a(m_\Lambda^a m_\Lambda^c + m_\Lambda^c m_\Lambda^a)]. \tag{4.3}$$

The strong convergence of the norm-bounded m_Λ^a for $\Lambda\to\infty$ entails that of (4.2) and (4.3). To perform a complete induction for the simultaneous iteration of (4.2) and (4.3) let us assume that the existence of s-lim$_\Lambda$ $L_\Lambda^m(s_i^b)$ and s-lim$_\Lambda$ $L_\Lambda^m(m_\Lambda^b)$ together with the norm-boundedness of these nets has been proved for $1\leqslant m\leqslant n-1$. Using (4.3) we obtain

$$L_\Lambda^n(m_\Lambda^b) = \sum_{a,c}\lambda_{abc}[f_a L^{n-1}(m_\Lambda^c) + g_a L_\Lambda^{n-1}(m_\Lambda^a m_\Lambda^c + m_\Lambda^c m_\Lambda^a)].$$

The last term is by means of the Leibniz formula a sum of $L_\Lambda^k(m_\Lambda^a)L_\Lambda^{n-1-k}(m_\Lambda^c)$-expressions, where $1\leqslant k\leqslant n-1$, which are norm-bounded, strongly convergent nets of operators in view of the induction hypothesis. Thus the total expression is convergent and norm-bounded and the analogous reasoning applies to $L_\Lambda^n(s_i^b)$.

(ii) In virtue of Lemma 3.4 $\mathcal{M}_\infty^\beta$ is invariant under $L^{\beta n}$, for all $n\in\mathbb{N}$. Relation (3.2) and Proposition 3.3 imply

$$\sigma\text{-w}\lim_\Lambda L_\Lambda^2(s_i^b) = L^{\beta 2}(s_i^b)$$

and (i) shows that this limit is approached also in the strong operator topology. Thus the reasoning can be repeated for $L_\Lambda^3(s_i^b)$, and so on. In this way we have proved (4.1) for $A = s_i^b$, b and i arbitrary.

(iii) An arbitrary $A\in\mathcal{A}_0$ is a finite sum of products $\otimes_{i\in\Lambda} s_i^a$ for some $\Lambda\in\mathcal{L}$. Applying the Leibniz rule once again, we observe that $L_\Lambda^n(\otimes_{i\in\Lambda} s_\Lambda^{a_i})$ is a finite sum of finite products $\Pi_{i\in\Lambda} L_\Lambda^{n_i}(s_i^{a_i})$, $n_i\in\mathbb{N}\cup\{0\}$, the factors of which are norm-bounded and converge because of (ii). Thus (4.1) is proved for general $A\in\mathcal{A}_0$. □

As far as we know the subsequent reasoning for proving the convergence of the local dynamical power series has been published for the considered class of models only for the case of pure phase representations and only for the KMS-dynamics (cf. Ref. 6 and references therein). (In Ref. 15 the analogous problem is discussed for a rather general class of representations but for z–z-spin interactions only.) Since a detailed elaboration of all necessary steps of our general case is given in Ref. 7, we outline here only the proof of the following result.

Theorem 4.2: Let τ_t^β, $t\in\mathbb{R}$, be the σ-weakly continuous W^*-automorphism group of $\mathcal{M}^\beta$ introduced in Theorem 3.5 and denote by τ_t^Λ the inner W^*-automorphism groups generated by L_Λ, $\Lambda\in\mathcal{L}$. Then it holds for all $t\in\mathbb{R}$

$$\text{s-lim}_\Lambda\, \tau_t^\Lambda(A) = \tau_t^\beta(A), \quad A\in\mathcal{A}_0, \tag{4.4}$$

where the convergence is uniform in all finite t-intervals.

Proof: (i) By complete induction with respect to n and $|\Lambda|$ one finds for every $A\in\mathcal{A}_\Lambda$, $n\in\mathbb{N}$, and $\Lambda\in\mathcal{L}$

$$\|L_{\Lambda'}^n(A)\| \leqslant c(A)n!M_{|\Lambda|}^n, \quad \forall\Lambda'\in\mathcal{L}, \tag{4.5}$$

where the positive constants $c(A)$ and $M_{|\Lambda|}$ depend only on A and $|\Lambda|$, respectively.

(ii) In virtue of (i) and Theorem 4.1 one obtains

$$\text{s-lim}_{\Lambda'}\sum_{n=0}^\infty L_{\Lambda'}^n(A)\frac{t^n}{n!} = \sum_{n=0}^\infty L^{\beta n}(A)\frac{t^n}{n!}, \quad \text{for } |t| < 1/M_{|\Lambda|}. \tag{4.6}$$

Since L^β generates τ_t^β, the rhs of Eq. (4.6) coincides with τ_t^β for the specified, A-dependent t-interval.

(iii) The extension of (4.6) to all of the t-axis is effectuated by theorems for analytic functions with uniform boundedness properties. □

Summing up we have shown that the local interactions between the atomic constituents of a macroscopic system, which is weakly coupled to reservoirs, lead in a unique way to a limiting dynamics with a nontrivial classical part, if the invariance under gauge transformations is spontaneously broken in the thermodynamical equilibrium state. This symmetry breaking is here a property of the local interactions and not brought about—as is usual in elementary particle theories—by an additional external potential of lower symmetry. For superconductors, superfluids and lasers, all of which are mainly treated by means of molecular field mod-

els, the gauge group is, in fact, spontaneously broken below the transition temperature. (The equilibrium phase transition for a laser system seems not to be observable since the required photon density cannot be provided.) Formula (3.15) for the effective, temperature-dependent limiting dynamics exhibits then, beside the usual molecular field part, a rotation of the macroscopic phase angle with a constant velocity proportional to the chemical potential. In spite of the extensive literature on, e.g., superconductivity (cf. Ref. 16 and references therein) the latter feature seems not to have been derived before from a microscopic model theory, such as the BCS theory. Indeed, the two prerequisites for such a derivation, the multiphase representation and the systematic construction of the closure of the dynamical generator, are lacking in the usual treatment of many body physics. Thus, the time dependence of the phase is usually introduced only in an heuristic manner. If now two of those systems are weakly coupled, one obtains, for example, a time-dependent phase difference, if the systems have different chemical potentials, and this gives rise to so-called macroscopic quantum phenomena.[16–18] A model discussion in the spirit of the present investigation[4] incorporates these phenomena into a general quantum dynamics which is uniquely determined by the microscopic interactions. If we apply the concepts of hierarchical systems (cf. Ref. 9, Chap. 6.2) to our models and call the local aspects the "lower level" and the classical aspects the "higher level" of our description, we establish on the one hand a remarkable coincidence of our results with the general scheme, especially we find the two time scales characterized by H_Λ^β and D_Λ, respectively, but we would like to emphasize on the other hand, that the lower level "explains" the higher level, if the contacts of the system to the surrounding (represented by the reservoirs) is specified.

ACKNOWLEDGMENTS

The author benefited from discussions with E. Duffner, W. Fleig, and M. Ullrich.

This work is part of a research project supported by the Deutsche Forschungsgemeinschaft.

APPENDIX: ESTIMATE FOR PROPOSITION 3.3

We derive here the relation

$$\|[H^r_\Lambda,\rho^\beta_\Lambda]\rho^{\beta\,-1}_\Lambda\Omega_\beta\| \leqslant C, \quad \forall \Lambda\in\mathscr{L}, \tag{*}$$

where the positive constant C is independent of Λ, and complete by this the proof of Proposition 3.3.

(i) From (3.8) we have

$$\rho^\beta_\Lambda = \exp(-\zeta^\beta_\Lambda - \beta H^{\beta r}_\Lambda) = \int_{\mathscr{T}^\beta}^{\oplus} \pi_\varphi(\exp(-\zeta^\varphi_\Lambda - \beta H^{\varphi r}_\Lambda))d\mu^\beta(\varphi)$$

which, in fact, is an invertible element of $\mathscr{M}^\beta_\Lambda$. Since

$$\|[H^r_\Lambda,\rho^\beta_\Lambda]\rho^{\beta\,-1}_\Lambda\Omega_\beta\|^2 = \int_{\mathscr{T}^\beta}\|[H^r_\Lambda,\rho^\varphi_\Lambda]\rho^{\varphi\,-1}_\Lambda\Omega_\varphi\|^2\, d\mu^\beta(\varphi),$$

it is sufficient to prove (*) for every pure phase representation with a φ-independent C. In the following all quantities refer to a fixed pure phase representation in spite of the index φ having been dropped. Furthermore, we set

$$H^r_\Lambda \equiv H_\Lambda, \quad H^{\varphi r}_\Lambda \equiv K_\Lambda, \quad \pi_\varphi(\mathscr{A}) \equiv \mathscr{A}.$$

(ii) For every $i\in\mathbb{Z}^d$ we choose a complete set of matrix units of $\mathscr{A}_i$: $\{P^i_{\epsilon i,\epsilon' i};\ \epsilon_i,\epsilon'_i = 1,\ldots,k^2\}$ with

$$P^i_{\epsilon i,\epsilon' i}P^i_{\eta i,\eta' i} = \delta_{\epsilon' i,\eta i}P^i_{\epsilon i,\eta' i}, \quad \sum_{\epsilon i}P^i_{\epsilon i,\epsilon i} = 1,$$

which diagonalize ρ_i and K_i

$$\rho_i = \sum_{\epsilon i}\lambda^i_{\epsilon i}P^i_{\epsilon i,\epsilon i}, \quad K_i = \sum_{\epsilon i}E^i_{\epsilon i}P^i_{\epsilon i,\epsilon i}.$$

For every $\Lambda\in\mathscr{L}$ we denote the vector $(\epsilon i;\ i\in\Lambda)$ by ϵ, the set of all such ϵ's is denoted by I_Λ. For $\epsilon,\epsilon'\in I_\Lambda$ we set

$$\lambda^\Lambda_\epsilon := \prod_{i\in\Lambda}\lambda^i_{\epsilon i}, \quad E^\Lambda_\epsilon := \sum_{i\in\Lambda}E^i_{\epsilon i}, \quad P^\Lambda_{\epsilon,\epsilon'} := \bigotimes_{i\in\Lambda}P^i_{\epsilon i,\epsilon' i}$$

and obtain by the fact that φ is a product state

$$\rho_\Lambda = \sum_{\epsilon\in I_\Lambda}\lambda^\Lambda_\epsilon P^\Lambda_{\epsilon,\epsilon}$$

with positive square root

$$\rho^{1/2}_\Lambda =: \sum_{\epsilon\in I_\Lambda}(\lambda^\Lambda_\epsilon)^{1/2}P^\Lambda_{\epsilon,\epsilon} =: w_\Lambda.$$

In $\mathscr{H}_\Lambda := \mathscr{A}_\Lambda\Omega \subset \mathscr{H}(\equiv\mathscr{H}_\varphi)$ we have the orthonormal basis

$$\{\Omega^\Lambda_{\epsilon,\epsilon'} = P^\Lambda_{\epsilon,\epsilon'}w^{-1}_\Lambda\Omega;\ \epsilon,\epsilon'\in I_\Lambda\}$$

with the property

$$(\Omega,\Omega^\Lambda_{\epsilon,\epsilon'}) = \mathrm{tr}_\Lambda\{\rho_\Lambda P^\Lambda_{\epsilon,\epsilon'}w^{-1}_\Lambda\} = 0, \quad \text{for } \epsilon\neq\epsilon';$$

$$\begin{aligned}([H_\Lambda,\rho_\Lambda]\rho^{-1}_\Lambda\Omega,\Omega_{\epsilon,\epsilon'}) &= (\Omega,(H_\Lambda - \rho_\Lambda H_\Lambda\rho^{-1}_\Lambda)P^\Lambda_{\epsilon,\epsilon'}w^{-1}_\Lambda\Omega)\\ &= (\Omega,H_\Lambda P^\Lambda_{\epsilon,\epsilon'}w^{-1}_\Lambda\Omega) - \mathrm{tr}_\Lambda\{w_\Lambda\rho_\Lambda H_\Lambda P^\Lambda_{\epsilon,\epsilon'}\}/\lambda^\Lambda_\epsilon\\ &= d_{\epsilon,\epsilon'}(\Omega,H_\Lambda P^\Lambda_{\epsilon,\epsilon'}w^{-1}_\Lambda\Omega),\end{aligned}$$

with

$$d_{\epsilon,\epsilon'} := (1 - \lambda_{\epsilon'}/\lambda_\epsilon).$$

(iii) Denoting

$$\mathscr{I}_\Lambda := I_\Lambda\times I_\Lambda$$

we have

$$\|[H_\Lambda,\rho_\Lambda]\rho^{-1}_\Lambda\Omega\|^2 = \sum_{(\epsilon,\epsilon')\in\mathscr{I}_\Lambda}d^2_{\epsilon,\epsilon'}|(\Omega,H_\Lambda P_{\epsilon,\epsilon'}w^{-1}_\Lambda\Omega)|^2.$$

Since K_Λ is diagonal and $d_{\epsilon,\epsilon} = 0$ we may replace H_Λ by ($m^a \equiv m^a_\varphi$!)

$$H_\Lambda - K_\Lambda = \sum_a g_a s^a_\Lambda(m^a_\Lambda - 2m^a).$$

Defining

$$\mathscr{I}_\Lambda(\epsilon_k,\epsilon'_k) := \{(\epsilon,\epsilon')\in\mathscr{I}_\Lambda;\epsilon_k,\epsilon'_k \text{ fixed}\}$$

it is sufficient to sum over $\cup_{k\in\Lambda}\cup_{\epsilon_k\neq\epsilon'_k}\mathscr{I}_\Lambda(\epsilon_k,\epsilon'_k)$, which enlarges the norm expression. For fixed $(\epsilon,\epsilon')\in\mathscr{I}_\Lambda(\epsilon_k,\epsilon'_k)$, $\epsilon_k\neq\epsilon'_k$,

$$\begin{aligned}&\left(\Omega,\sum_a g_a s^a_\Lambda(m^a_\Lambda - 2m^a)\Omega^\Lambda_{\epsilon,\epsilon'}\right)\\ &= \left(\Omega,\sum_a g_a 2s^a_k(m^a_\Lambda - m^a)\Omega^\Lambda_{\epsilon,\epsilon'}\right)\\ &= \left(\Omega,\sum_a 2g_a\left(\frac{s^{a2}_k}{|\Lambda|} + s^a_k\left(\frac{s^a_{\Lambda'}}{|\Lambda|} - m^a\right)\right)\Omega^\Lambda_{\epsilon,\epsilon'}\right),\end{aligned}$$

with $\Lambda' := \Lambda \setminus k$.

We decompose the sum over $\mathscr{I}_\Lambda(\epsilon_k,\epsilon_k')$, $\epsilon_k \neq \epsilon_k'$, into four parts. Stipulating $(\epsilon,\epsilon') \in \mathscr{I}_\Lambda(\epsilon_k,\epsilon_k')$ the first is

$$\sum_{(\epsilon,\epsilon')} \left| \left(\Omega, \sum_a 2g_a s_k^{a2} \Omega^{\Lambda}_{\epsilon,\epsilon'} \right) \Big/ \ |\Lambda| \right|^2$$
$$= \left| \left(\Omega, \sum_a 2g_a s_k^{a2} \Omega^{k}_{\epsilon k,\epsilon' k} \right) \Big/ \ |\Lambda| \right|^2 \leqslant \frac{C_1}{|\Lambda|^2},$$

where the completeness relation in $\mathscr{H}_{\Lambda \setminus k}$ has been employed. The second is

$$\sum_{(\epsilon,\epsilon')} \left(\Omega, \sum_a 2g_a (s_k^{a2}/|\Lambda|) \Omega^{\Lambda}_{\epsilon,\epsilon'} \right)$$
$$\times \left(\Omega^{\Lambda}_{\epsilon,\epsilon'}, \sum_b 2g_b s_k^b \left(\frac{s_{\Lambda'}^b}{|\Lambda|} - m^b \right) \Omega \right)$$
$$= \sum_{a,b} \left(\Omega, 2g_a \frac{s_k^{a2}}{|\Lambda|} \Omega^{k}_{\epsilon k,\epsilon' k} \right)$$
$$\times (\Omega^{k}_{\epsilon k,\epsilon' k}, 2g_b s_k^b \Omega) \left(\Omega, \left(\frac{s_{\Lambda'}^b}{|\Lambda|} - m^b \right) \Omega \right).$$

Observing

$$\left(\Omega, \left(\frac{s_{\Lambda'}^b}{|\Lambda|} - m^b \right) \Omega \right) = -\frac{m^b}{|\Lambda|},$$

the modulus of the considered term is dominated by $C_2/|\Lambda|^2$. The third term is similar to the second one and the fourth is

$$\sum_{(\epsilon,\epsilon')} \left| \left(\Omega, \sum_a 2g_a s_k^a \left(\frac{s_{\Lambda'}^a}{|\Lambda|} - m^a \right) \Omega^{\Lambda}_{\epsilon,\epsilon'} \right) \right|^2$$
$$= \sum_{a,b} (\Omega, 2g_a s_k^a \Omega^{k}_{\epsilon k,\epsilon' k}) (\Omega^{k}_{\epsilon k,\epsilon k'}, 2g_b s_k^b \Omega)$$
$$\times \left(\Omega, \left(\frac{s_{\Lambda'}^a}{|\Lambda|} - m^a \right) \left(\frac{s_{\Lambda'}^b}{|\Lambda|} - m^b \right) \Omega \right),$$

where the last expectation value leads to the upper bound $C_4/|\Lambda|$. The positive constants $C_1 \cdots C_4$ are Λ- and k-independent as well as φ-independent. The k-summation in the augmented norm expression is thus compensated by a factor $1/|\Lambda|$ and the norm itself is dominated by a Λ- and φ-independent constant.

[1] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics I, II* (Springer, New York, 1979, 1981).
[2] W. Thirring and A. Wehrl, Commun. Math. Phys. **4**, 30 (1967); W. Thirring, Commun. Math. Phys. **7**, 181 (1968); F. Jelinek, Commun. Math. Phys. **9**, 169 (1968).
[3] W. Fleig, Masters Thesis, Universitat Tübingen (1981); W. Fleig, "On the Symmetry Breaking Mechanism of the Strong-Coupling BCS-Model," Acta Phys. Aust. **55**, 135 (1983).
[4] M. Ullrich, Masters Thesis, Universitat Tübingen (1983); A. Rieckers and M. Ullrich, "On the Physical Dynamics of the Quansispin BCS-Model," preprint, Tübingen (1983); A. Rieckers and M. Ullrich, "On the Microscopic Derivation of the Josephson-Relation in Operator Form," preprint, Tübingen (1983).
[5] M. B. Ruskai, Commun. Math. Phys. **20**, 193 (1971) and Ref. 1, Chap. 6.
[6] L. van Hemmen, Fortschr. Phys. **26**, 397 (1978).
[7] E. Duffner, "Existence of the Limiting Mean Field Dynamics in General Equilibrium Representations," preprint, Tübingen (1983).
[8] D. A. Dubin and G. L. Sewell, J. Math. Phys. **11**, 2990 (1970); P. E. T. Jörgensen, Tôhoku Math. J. **30**, 277 (1978); Ref. 1, Chap. 6.3.
[9] H. Primas, *Chemistry, Quantum Mechanics and Reductionism, Lecture Notes in Chemistry 24* (Springer, Berlin, 1981).
[10] S. Sakai, *C*-Algebras and W*-Algebras* (Springer, Berlin, 1971).
[11] G. K. Pedersen, *C*-Algebras and their Automorphism Groups* (Academic, London, 1979).
[12] E. Størmer, J. Funct. Anal. **3**, 48 (1969).
[13] M. Takesaki, *Theory of Operator Algebras I* (Springer, New York, 1979).
[14] M. Fannes, H. Spohn, and A. Verbeure, J. Math. Phys. **21**, 355 (1980).
[15] G. G. Emch and H. J. F. Knops, J. Math. Phys. **11**, 3008 (1970).
[16] M. Tinkham, *Introduction to Superconductivity* (McGraw-Hill Kogahuska Ltd., Tokyo, 1975).
[17] B. D. Josephson, Phys. Lett. **1**, 251 (1962); and Rev. Mod. Phys. **46**, 251 (1974).
[18] J. E. Mercereau, in *Superconductivity*, edited by R. D. Parks (Dekker, New York, 1969).

Group theoretic treatment of the Dirac–Coulomb equation and matrix elements of its tensor operators

M. K. F. Wong
Department of Mathematics, Fairfield University, Fairfield, Connecticut 06430
Hsin-Yang Yeh[a)]
Moorhead State University, Moorhead, Minnesota 56560

(Received 2 February 1984; accepted for publication 20 April 1984)

The solution of the Dirac–Coulomb equation obtained by Wong and Yeh is interpreted in terms of the SU(2)×SO(2,1) group. All electromagnetic transition probabilities can be considered as matrix elements of the tensor operators of this group, and evaluated exactly. The cases considered include transitions from bound-to-bound, bound-to-continuum, and continuum-to-continuum states.

PACS numbers: 02.20. + b, 11.10.Qr, 12.20. − m

I. INTRODUCTION

We wish to present in this paper a group theoretic treatment of the Dirac–Coulomb equation in terms of SU(2)×SO(2,1). Though various authors[1,2] have used other groups such as SO(4,1) and SO(4,2) to deal with the relativistic Kepler problem, we find that it is simpler and more direct to use just the product group SU(2)×SO(2,1). Besides, in using the groups SO(4,1) and SO(4,2), there is an implicit involvement of SU(2) and SO(2,1) as subgroups.

We shall show that the solution of the Dirac–Coulomb equation obtained by us[3] can be interpreted under the product group SU(2)×SO(2,1). The group SU(2) is responsible for the angular momentum and spin part of the wave function and SO(2,1) is responsible for the radial wave function. The idea that SO(2,1) is involved with the radial wave function of the Schrödinger equation is well known in the literature; see, e.g., Bacry,[4] Miller,[5] and Armstrong.[6] However, in trying to connect the Schrödinger wave function directly with the basis function of SO(2,1), Miller and Armstrong had to introduce two parameters z and t, where t is not directly related to the radial parameter r. [See, however, Chacón, Levi, and Moshinsky, J. Math. Phys. **17**, 1919 (1976), for the interpretation of t as time in the Schrödinger picture.] We shall restrict our parameter to the one single variable r, or equivalently ρ. In doing so we have to consider a slightly modified wave function G, which is related to the Schrödinger wave function R by the simple relationship

$$G=\rho^{3/4}R. \tag{1.1}$$

What we say about SO(2,1) actually applies to G. However, since G is related to R by the simple relation Eq. (1.1), once we know the group theoretic significance of G we can obtain all results connected with R. The solution to the Dirac–Coulomb equation obtained by us[3] is very similar to the corresponding Schrödinger equation. Thus we can transfer the group theoretic interpretation from the Schrödinger case to the Dirac case.

The electromagnetic transition probabilities of the relativistic electron in a Coulomb field can thus be considered as matrix elements of the tensor operators under this group. It is our intention to show that these matrix elements can all be evaluated exactly. We have thus a twofold result. First, from the group theoretic point of view, it shows that, according to the Wigner–Eckart theorem, both the reduced matrix elements and the Wigner coefficients of SO(2,1) can be explicitly identified, and second, from the physical point of view, the electromagnetic properties of the relativistic electron in a Coulomb field are also exactly known.

It should be mentioned that the exact solution of the relativistic Coulomb problem using SU(2)×SO(2,1) was also given in a paper by Barut and Bornzin.[7] (We wish to thank the referee for bringing this paper to our attention.) However, it seems that the solution we obtained in Ref. 3 is more explicit. Group theoretical calculation of transitions from bound-to-bound, bound-to-continuum, and continuum-to-continuum states have also been discussed by Barut and Wilson[8] (bound–bound, nonrelativistic), Barut, Rasmussen, and Salamó, I[9] (continuum–continuum, elastic scattering), II[10] (bound–continuum). Again we wish to thank the referee for bringing these papers to our attention.

In Sec. II, we start with the Schrödinger equation and show how it can be interpreted under the group SU(2)×SO(2,1). This interpretation is then carried over to the Dirac–Coulomb equation, according to the solution obtained by us.[3] In Sec. III, we show that the electromagnetic matrix elements of the relativistic electron in bound–bound transitions can be exactly evaluated. This corresponds physically to the transition probabilities of the discrete spectrum of hydrogenlike atoms. In Sec. IV, we obtain the exact results for bound–continuum transitions. This corresponds physically to internal conversion and the photoeffect. In Sec. V, we obtain the exact results for continuum–continuum transitions. This corresponds physically to bremsstrahlung and its inverse process, pair creation in the presence of a Coulomb field.

II. SU(2)×SO(2,1) FOR THE DIRAC–COULOMB EQUATION

The Dirac–Coulomb equation can be treated in a similar way as the Schrödinger equation as far as its group properties are concerned. We shall therefore start with the Schrö-

[a)] Current address: Naval Weapons Center, Code 3313, China Lake, CA 93555.

 0022-2488/84/092602-12$02.50

dinger equation. The complete wave function of the Schrödinger equation is

$$\psi(r,\theta,\phi) = R(r)Y_{lm}(\theta,\phi), \tag{2.1}$$

where $Y_{lm}(\theta,\phi)$ is the usual spherical harmonics and

$$R(r) = Ne^{-1/2\rho}\rho^{l}{}_1F_1(-n+l+1,2l+2,\rho), \tag{2.2}$$

$$N = \frac{1}{(2l+1)!}\left[\frac{(n+l)!}{(n-l-1)!2n}\right]^{1/2}\left(\frac{2Z}{n}\right)^{3/2}, \tag{2.3}$$

$$\rho = 2Zr/n. \tag{2.4}$$

The radial wave function of $R(\rho)$ satisfies the equation

$$\frac{1}{\rho^2}\frac{d}{d\rho}\left(\rho^2\frac{dR}{d\rho}\right) + \left[\frac{n}{\rho} - \frac{1}{4} - \frac{l(l+1)}{\rho^2}\right]R = 0. \tag{2.5}$$

Now we define the "modified radial wave function" $G(\rho)$ as

$$G(\rho) = \rho^{3/4}R(\rho). \tag{2.6}$$

Then we find that $G(\rho)$ satisfies the equation

$$-\rho\frac{d^2G}{d\rho^2} - \frac{1}{2}\frac{dG}{d\rho} + \frac{\rho}{4}G + \frac{4l(l+1)+3/4}{4\rho}G = nG. \tag{2.7}$$

Now we assert that G is the basis function for the "positive discrete series representation" of SO(2,1), with generators

$$J_3 = -\rho\frac{d^2}{d\rho^2} - \frac{1}{2}\frac{d}{d\rho} + \frac{\rho}{4} + \frac{l(l+1)}{\rho} + \frac{3}{16\rho}, \tag{2.8}$$

$$J_1 = J_3 - \tfrac{1}{2}\rho, \tag{2.9}$$

$$J_2 = -i\rho\frac{d}{d\rho} - \frac{i}{4}. \tag{2.10}$$

One easily checks that J_1, J_2, and J_3 are the generators of SO(2,1) since

$$[J_1,J_2] = -iJ_3, \quad [J_2,J_3] = iJ_1, \text{ and } [J_3,J_1] = iJ_2. \tag{2.11}$$

Also the Casimir operator I_2 is

$$I_2 = -J_1^2 - J_2^2 + J_3^2 = l(l+1). \tag{2.12}$$

The matrix elements of the generators are

$$J_3G_{nl} = nG_{nl}, \tag{2.13}$$

$$J_+G_{nl} = (J_1 + iJ_2)G_{nl} = [(n-l)(n+l+1)]^{1/2}G_{n+1,l}, \tag{2.14}$$

$$J_-G_{nl} = (J_1 - iJ_2)G_{nl} = [(n+l)(n-l-1)]^{1/2}G_{n-1,l}. \tag{2.15}$$

Thus the "modified Schrödinger wave function" $G_{nl}(\rho)$ $Y_{lm}(\theta,\phi)$ is a basis of the representation SO(2,1)×SO(3) with generators J_1, J_2, J_3 for SO(2,1) and L_1, L_2, L_3 for SO(3), where $\mathbf{L} = \mathbf{r}\times\mathbf{p}$ is the angular momentum operator.

For the Dirac–Coulomb equation, the generators of SU(2) are now the total angular momentum $\mathbf{J}$, where $\mathbf{J} = \mathbf{L} + \mathbf{S}$. Here $\mathbf{S}$ is the spin angular momentum $\mathbf{S} = \frac{1}{2}\boldsymbol{\sigma}$, where σ_1, σ_2, σ_3 are the Pauli matrices. The spinor wave function χ_k^μ are

$$\chi_k^\mu = \sum_\tau \begin{bmatrix} l & \frac{1}{2} & j \\ \mu-\tau & \tau & \mu \end{bmatrix} Y_l^{\mu-\tau}(\theta,\phi)\chi_{1/2}^\tau, \tag{2.16}$$

where

$$\kappa = \tilde{\omega}(j+\tfrac{1}{2}) \tag{2.17}$$

$$l = j + \tfrac{1}{2}\tilde{\omega}, \tag{2.18}$$

and

$$(\boldsymbol{\sigma}\cdot\mathbf{L}+1)\chi_k^\mu = -\kappa\chi_k^\mu. \tag{2.19}$$

Thus we see that three quantum numbers are required to completely specify the state: j, $\tilde{\omega}$, μ, or equivalently j, l, μ, because j, l, and $\tilde{\omega}$ are related to each other according to Eq. (2.18). The complete solution to the Dirac–Coulomb equation is[3]

$$\psi = N\begin{pmatrix} i(E(\kappa/\gamma)-m)^{1/2}\phi_{\lambda(-\gamma)}\chi_{-\kappa}^\mu \\ \tilde{\omega}(E(\kappa/\gamma)+m)^{1/2}\phi_{\lambda(\gamma)}\chi_k^\mu \end{pmatrix}, \tag{2.20}$$

where

$$\lambda(\gamma) = |\gamma| + \tfrac{1}{2}(\tilde{\omega}-1), \tag{2.21}$$

$$\gamma = \tilde{\omega}[(j+\tfrac{1}{2})^2 - (Ze^2)^2]^{1/2}, \tag{2.22}$$

$$\phi_{\lambda(\gamma)} = \frac{[\Gamma(2\lambda+2+n)\Gamma(n+2)]^{1/2}}{\Gamma(2\lambda+2)} \times\rho^\lambda e^{(-1/2)\rho}{}_1F_1(-n_r,2\lambda+2,\rho), \tag{2.23}$$

$$n_r = Ze^2E/\mu_1 - \lambda - 1 = n - \lambda - 1, \tag{2.24}$$

$$\rho = 2\mu_1 r = 2(m^2 - E^2)^{1/2}, \tag{2.25}$$

$$N = 2n^{-1/2}\mu_1^{3/2}[(n-\gamma)!(n-\gamma-1)!]^{-1/2} \times[(n-\gamma)(n-\gamma+1)(E(\kappa/\gamma)-\tilde{\omega}m) + (E(\kappa/\gamma)+\tilde{\omega}m)]^{-1/2}. \tag{2.26}$$

Again we define the "modified radial wave function" G to be

$$G_{\lambda(\gamma)}(\rho) = \rho^{3/4}\phi_{\lambda(\gamma)}(\rho). \tag{2.27}$$

Then $G_{\lambda(\gamma)}(\rho)$ is a basis function of SO(2,1) with generators K_1, K_2, K_3 where

$$K_3 = -\rho\frac{d^2}{d\rho^2} - \frac{1}{2}\frac{d}{d\rho} + \frac{\rho}{4} + \frac{\lambda(\lambda+1)}{\rho} + \frac{3}{16\rho}, \tag{2.28}$$

$$K_1 = K_3 - \tfrac{1}{2}\rho, \tag{2.29}$$

$$K_2 = -i\rho\frac{d}{d\rho} - \frac{i}{4}. \tag{2.30}$$

Note that the solution in (2.20) is obtained after a transformation S, where

$$S = \exp[-\tfrac{1}{2}\rho_2\boldsymbol{\sigma}\cdot\hat{r}\tanh^{-1}(Ze^2/K)], \tag{2.31}$$

$$K = \beta(\boldsymbol{\sigma}\cdot\mathbf{L}+1). \tag{2.32}$$

However, this transformation commutes with the generators K_1, K_2, K_3 and therefore we have $SK_iS^{-1} = K_i$, $i = 1,2.3$.

Comparing the Dirac case with the Schrödinger case, we see that the only difference is the change from l in the Schrödinger case to λ, and the introduction of the additional quantum number $\tilde{\omega}$, which takes on the values of $+1$ or -1. Thus the bound states are completely specified under the group SU(2)×SO(2,1).

The continuum is also very easy to deal with. All one has to do is to reinterpret ρ. By defining $k = (E^2 - m^2)^{1/2}$, we find that the radial wave function for the continuum is a function of $\rho = 2ikr$, with everything else unchanged. The normalization of the continuum wave function will be discussed in Sec. IV.

Thus we have shown that the wave function of the

Dirac–Coulomb equation is related to the basis function of the irreducible representations of $SU(2)\times SO(2,1)$.

III. ELECTROMAGNETIC MATRIX ELEMENTS AS TENSOR OPERATORS IN BOUND-BOUND TRANSITIONS

In a previous paper[11] we have obtained the dipole transition probabilities for hydrogenic atoms using the exact solution of the Dirac–Coulomb equation. This result is equivalent to the neglect of retardation effects. In this paper we shall give the complete result for this problem, including retardation and for all multipoles, both electric and magnetic.

The exact formula in bound–bound transitions has been given by Babushkin,[12] Scofield,[13] Rosner and Bhalla,[14] and Moses.[15] Moreover, the results of Babushkin and Scofield are basically contained in Rose,[16] who considered internal conversion. The result of Rosner and Bhalla differs from Scofield's in appearance. However, it has been pointed out by Grant[17] that the two results are the same under gauge invariance. Moses[15] has also given an alternate expression, at least for the electric part. The simplest form is given by Grant.

According to Grant, the spontaneous emission probability per unit time for the transition $\beta\to\alpha$ is

$$A_{\beta\to\alpha}=\frac{1}{2j_\beta+1}\sum_L\sum_{M_\alpha}\sum_{M_\beta}2\pi|M_{\alpha\beta}|^2$$
$$=2\alpha\omega\sum_L\frac{(2j_\alpha+1)}{(2L+1)}\begin{pmatrix}j_\beta & L & j_\alpha\\ \frac{1}{2} & 0 & -\frac{1}{2}\end{pmatrix}^2|\bar{M}_{\alpha\beta}|^2, \tag{3.1}$$

where

$$\bar{M}_{\alpha\beta}=\bar{M}^m_{\alpha\beta}+\bar{M}^e_{\alpha\beta}, \tag{3.2}$$

$$M^{\bar{m}}_{\alpha\beta}=\frac{i^{L+1}2L+1}{L(L-1)^{1/2}}(\kappa_\alpha+\kappa_\beta)$$
$$\times\int_0^\infty(g_\alpha f_\beta+f_\alpha g_\beta)j_L(kr)r^2\,dr, \tag{3.3}$$

$$\bar{M}^e_{\alpha\beta}=i^L\left\{\left(\frac{L}{L+1}\right)^{1/2}\left[(\kappa_\alpha-\kappa_\beta)\right.\right.$$
$$\times\int_0^\infty(g_\alpha f_\beta+f_\alpha g_\beta)j_{L+1}(kr)r^2\,dr$$
$$\left.+(L+1)\int_0^\infty(g_\alpha f_\beta-f_\alpha g_\beta)j_{L+1}(kr)r^2\,dr\right]$$
$$-\left(\frac{L+1}{L}\right)^{1/2}\left[(\kappa_\alpha-\kappa_\beta)\right.$$
$$\times\int_0^\infty(g_\alpha f_\beta+f_\alpha g_\beta)j_{L-1}(kr)r^2\,dr$$
$$\left.\left.-L\int_0^\infty(g_\alpha f_\beta-f_\alpha g_\beta)j_{L-1}(kr)r^2\,dr\right]\right\}, \tag{3.4}$$

where the solution to the Dirac–Coulomb equation is written as

$$\psi^\mu_\kappa=\begin{pmatrix}g\,\chi^\mu_\kappa\\ if\chi^\mu_{-\kappa}\end{pmatrix}, \tag{3.5}$$

χ^μ_κ being the usual spinors. The f and g are normalized such that

$$\int_0^\infty(g^2+f^2)r^2\,dr=1. \tag{3.6}$$

This result can be easily carried over into our notation. Thus we obtain for the transition probability per unit time, for magnetic multipole radiation:

$$\Gamma_{a'\to a}(m)=2\alpha\omega\sum_L\frac{(2L+1)(2j+1)}{L(L+1)}$$
$$\times(\kappa+\kappa')^2I^{+2}_L\begin{pmatrix}j & j' & L\\ -\frac{1}{2} & \frac{1}{2} & 0\end{pmatrix}^2, \tag{3.7}$$

where

$$I^+_L=\int j_L(kr)[\tilde{\omega}\,g_af_{a'}+\tilde{\omega}'g_{a'}f_a]r^2\,dr, \tag{3.8}$$

$$\psi=\begin{pmatrix}if\chi^\mu_{-\kappa}\\ \tilde{\omega}g\,\chi^\mu_\kappa\end{pmatrix}. \tag{3.9}$$

The selection rule for magnetic multipole radiation is such that the term is zero unless $j-\frac{1}{2}\tilde{\omega}+j'+\frac{1}{2}\tilde{\omega}'+L$ is even. For the electric multipole radiation, the matrix element is zero unless the term above is odd. For the electric multipole radiation, we find

$$\Gamma_{a'\to a}(e)=2\alpha\omega\sum_L\frac{1}{(2L+1)}\left\{\left(\frac{L}{L+1}\right)^{1/2}\right.$$
$$\times[(\kappa-\kappa')I^+_{L+1}+(L+1)I^-_{L+1}]$$
$$\left.-\left(\frac{L+1}{L}\right)^{1/2}[(\kappa-\kappa')I^+_{L-1}-LI^-_{L-1}]\right\}^2$$
$$\times(2j'+1)\begin{pmatrix}j & j' & L\\ -\frac{1}{2} & \frac{1}{2} & 0\end{pmatrix}^2, \tag{3.10}$$

where

$$I^-_{L\pm1}=\int j_{L\pm1}(kr)\,[-\tilde{\omega}'g_{a'}f_a+\tilde{\omega}g_af_{a'}]r^2\,dr. \tag{3.11}$$

Thus the complete result can be obtained if one can evaluate the integrals in (3.8) and (3.11). For bound–bound transitions, these integrals are further equivalent to the following integral $\mathcal{J}$.

$$\mathcal{J}=\int_0^\infty e^{-\mu_1r-\mu_1'(r)}(2\mu_1r)^\lambda(2\mu_1'r)^{\lambda'}$$
$$\times j_L(kr)L^{2\lambda+1}_{n_r}(2\mu_1r)L^{2\lambda'+1}_{n_r'}(2\mu_1'r)r^2\,dr, \tag{3.12}$$

where L^q_p is a generalized Laguerre polynomial. This integral can be evaluated exactly by using the generating function of the generalized Laguerre polynomial, a technique we have used successfully in a previous paper.[18] Here we shall only give the final result.

A convergent sum for the integral $\mathcal{J}$ can be obtained in all cases if we expand in powers of k/μ_1'. We can distinguish between two cases: (a) $k/\mu_1'<1$, and (b) $k/\mu_1'>1$. For case (a) we obtain the result

$$\mathcal{J} = \sum_{\alpha,\beta,\gamma,\delta} \pi^{1/2} 2^{\lambda+\lambda'-L-1} (-)^{\gamma+\delta+n_r'} \mu_1'^{-\lambda-L-3-2\delta-\gamma} k^{L+2\delta} \mu_1^{\lambda+\gamma}$$
$$\times \frac{\Gamma(\lambda+\lambda'+L+3)}{\delta!\Gamma(\frac{1}{2}(\lambda+\lambda'+L+3))} \frac{\Gamma(\frac{1}{2}(\lambda+\lambda'+L+3)+\delta)\Gamma(\frac{1}{2}(\lambda+\lambda'+L+4)+\delta)}{\Gamma(\frac{1}{2}(\lambda+\lambda'+L+4))\Gamma(L+\frac{3}{2}+\delta)}$$
$$\times \binom{\lambda+\lambda'+L+2+2\delta+\gamma}{\gamma} \binom{\lambda+\lambda'+L+2+2\delta+\gamma+\alpha}{\alpha} \binom{\gamma}{\beta}$$
$$\times \binom{\lambda-\lambda'+L+2\delta+1+\gamma}{n_r'-\alpha} \binom{\gamma+2\lambda+1+n_r-\beta}{n_r-\beta}, \tag{3.13}$$

where, without loss of generality, we have taken $\mu_1 < \mu_1'$. The series in (3.13) is convergent. For the parameters α and β, they take values from 0 to n_r and from 0 to n_r' respectively and are therefore finite. The summation over γ is convergent because $\mu_1 < \mu_1'$. The summation over δ is convergent because $k/\mu_1' < 1$. For different multipoles, the transition probability goes down roughly as $(k/2\mu_1')^{2L}$. This is the justification of the "dipole approximation" when the energy of the photon is low, since in that case we have $k/\mu_1' < 1$. Also we can see that this is equivalent to the neglect of retardation, since in that case we also have $k/\mu_1' < 1$.

For case (b), $k/\mu_1' > 1$, we obtain

$$\mathcal{J} = \pi^{1/2} 2^{\lambda+\lambda'-L-1} \mu_1^{\lambda} \mu_1'^{\lambda'} k^{-(\lambda+\lambda'+3)} \Gamma(\lambda+\lambda'+L+3)$$
$$\times \Bigg\{ \sum_{\alpha,\beta,\gamma,\delta} \frac{\sin[(\pi/2)(L-\lambda-\lambda')]\Gamma(\frac{1}{2}(\lambda+\lambda'+L+3)+\delta)\Gamma(\frac{1}{2}(\lambda+\lambda'-L+2)+\delta)}{\Gamma(\frac{1}{2}(\lambda+\lambda'+L+4))\Gamma(\frac{1}{2}(\lambda+\lambda'+L+3))\Gamma(\frac{1}{2}+\delta)\delta!}$$
$$\times \frac{(-1)^{\delta}(2\delta)!\mu_1^{\gamma}\mu_1'^{2\delta-\gamma}}{k^{2\delta}\alpha!(\gamma-\alpha)!\beta!(2\delta-\gamma-\beta)!} \binom{n_r+\lambda+\gamma-\alpha}{n_r-\alpha} \binom{n_r'+\lambda'+2\delta-\gamma-\beta}{n_r'-\beta}$$
$$+ \sum_{\alpha'\beta'\gamma'\delta'} \frac{\cos[(\pi/2)(\lambda+\lambda'-L)]\Gamma(\frac{1}{2}(\lambda+\lambda'+L+4)+\delta')\Gamma(\frac{1}{2}(\lambda+\lambda'+L+3)+\delta')}{\Gamma(\frac{1}{2}(\lambda+\lambda'+L+3))\Gamma(\frac{1}{2}(\lambda+\lambda'+L+4))\Gamma(\frac{3}{2}+\delta')\delta'!\alpha'!(\gamma'-\alpha')!\beta'!}$$
$$\times \frac{k^{-2\delta'-1}\mu_1^{\gamma}\mu_1'^{2\delta-\gamma+1}}{(2\delta'+1-\gamma'-\beta')!} \binom{n_r+\lambda+\gamma'-\alpha'}{n_r'-\alpha'} \binom{n_r'+\lambda'+2\delta'+1-\gamma'-\beta'}{n_r'-\beta'} \Bigg\}, \tag{3.14}$$

where we have used the relation

$$\Gamma(z)\Gamma(1-z) = \pi \csc(\pi z). \tag{3.15}$$

This series is also convergent. The summation over $\alpha(\alpha')$ and $\beta(\beta')$ goes from zero to n_r and zero to n_r', respectively, and is therefore finite. The summation over $\gamma(\gamma')$ is convergent because $\mu_1/\mu_1' < 1$. The summation over $\delta(\delta')$ is convergent because $\mu_1'/k < 1$. For different multipoles, the transition probability goes down as 2^{-2L}. Thus when the photon energy is large, higher order multipoles, at least for the first few, might be important.

Finally we wish to mention a new selection rule for magnetic radiation. This rule is that if $\tilde{\omega}$ and $\tilde{\omega}'$ are of opposite sign, then the magnetic radiation is negligible. Using this rule, we easily conclude that $s_{1/2} \leftrightarrow d_{3/2}$ transitions are forbidden magnetically. This result was deduced by Rose[16] after considering nuclear forces in a lengthy way. However, our results are straightforward and can be applied to other angular momentum states.

IV. BOUND–CONTINUUM TRANSITIONS

In the case of bound–continuum transitions, there are two distinct processes: (1) internal conversion and (2) photoeffect and its inverse process, radiative capture.

For internal conversion, the relevant tensor operator in the radial integral is a spherical Hankel function of the first kind. The internal conversion coefficient, as one can expect, increases as L increases, and therefore from the mathematical point of view, the series does not converge as L increases to infinity. However, in actual practice, one need not calculate beyond, say, $L = 10$, since then the photon number becomes extremely small. For the photoeffect, the relevant tensor operator is the spherical Bessel function, just as the bound–bound case, and the series converges as L increases.

There is also a difference in the continuum electron wave function between internal conversion and photoeffect. For internal conversion, the continuum wave function can be taken directly from the solution of the Dirac–Coulomb equation with a suitable normalization, for example, in the energy scale. For the photoeffect, the outgoing electron wave function must be written in the form of a plane wave plus its scattered part. Therefore it must be written as a superposition of solutions to the Dirac–Coulomb equation. In this section we give a detailed construction of such a wave function. But first let us discuss the internal conversion coefficient.

The main work on internal conversion coefficients was done by Rose[16,19,20] and his co-workers. In this paper we shall restrict ourselves to the "static" part of the nucleus. Moreover, we shall treat the nucleus as a point charge. Our main interest is in the correct expression for the normalized wave function of the electron, and in the analytic evaluation of the radial integral.

For the photon vector potential we use the multipole expansion as given by Rose.[16] Then we obtain basically two internal conversion coefficients β_L and α_L, corresponding to the magnetic and electric multipoles, respectively. In a derivation completely similar to Rose except with regard to factors in the averaging process,

$$\beta_L = \frac{\pi\alpha k}{L(L+1)}(2l'+1)\sum_{\kappa} \times(2j+1)(2j-\tilde{\omega}+1)C^2(j-\tfrac{1}{2}\tilde{\omega},j'+\tfrac{1}{2}\tilde{\omega}',L,0,0) \times W^2(j,j-\tfrac{1}{2}\tilde{\omega},j',j'+\tfrac{1}{2}\tilde{\omega}',\tfrac{1}{2},L)(\kappa+\kappa')^2 I_L^+, \tag{4.1}$$

$$I_L^+ = \int_0^\infty h_L^{(1)}(\tilde{\omega}g_\kappa f_{\kappa'} + \tilde{\omega}' g_{\kappa'} f_\kappa) r^2\, dr, \tag{4.2}$$

$$h_L^{(1)} = (\pi/2kr)^{1/2} H^{(1)}_{L+\frac{1}{2}}(kr), \tag{4.3}$$

$$j - \tfrac{1}{2}\tilde{\omega} + j' + \tfrac{1}{2}\tilde{\omega} + L = \text{even}. \tag{4.4}$$

We also note that

$$C^2(j-\tfrac{1}{2}\tilde{\omega},j'+\tfrac{1}{2}\tilde{\omega}',L,0,0)W^2(j,j-\tfrac{1}{2}\tilde{\omega},j',j'-\tfrac{1}{2}\tilde{\omega}',\tfrac{1}{2},L) = \left(\frac{2L+1}{(2j-\tilde{\omega}+1)(2j'+\tilde{\omega}'+1)}\right)\begin{pmatrix} j & j' & l \\ \frac{1}{2} & -\frac{1}{2} & 0 \end{pmatrix}^2, \tag{4.5}$$

$$\alpha_L = \frac{\pi\alpha k}{L(L+1)}(2j'+\tilde{\omega}'+1)\sum_{\kappa}(2j+\tilde{\omega}+1)(2j+1) \times C^2(j+\tfrac{1}{2}\tilde{\omega},j'+\tfrac{1}{2}\tilde{\omega}',L,0,0) \times W^2(j,j+\tfrac{1}{2}\tilde{\omega},j',j'+\tfrac{1}{2}\tilde{\omega}',\tfrac{1}{2},L) \times \Big|(\kappa'-\kappa)\int_0^\infty h^{(1)}_{L-1}(\tilde{\omega}g_\kappa f_{\kappa'} + \tilde{\omega}' g_{\kappa'} f_\kappa) r^2\, dr + L\int_0^\infty h^{(1)}_{L-1}(\tilde{\omega}g_\kappa f_{\kappa'} - \tilde{\omega}' g_{\kappa'} f_\kappa) r^2\, dr + L\int_0^\infty h_L^{(1)}(\tilde{\omega}g_\kappa f_{\kappa'} + \tilde{\omega}' g_{\kappa'} f_\kappa) r^2\, dr\Big|^2. \tag{4.6}$$

As one can see, our results in Eqs. (4.1) and (4.6) differ from those of Rose by a factor of $(2j'+1)/(2L+1)$. This comes from the difference in calculating the averaging process. The process is, according to Rose, to average over the initial states and sum over final states. We perform this operation by dividing the sum over the final states by $(2j'+1)$, the number of initial states. However, Rose performs this operation by dividing the sum by $(2L+1)$. Hence the difference of the factor $(2j'+1)/(2L+1)$. We think our factor is correct because this result must agree with the corresponding result in bound–bound transitions. But our result agrees with Grant[17] as given in Sec. III. Let us also add that all the other authors such as Scofield,[13] Rosner and Bhalla,[14] do not have the factor $(2L+1)$ in their formula for bound–bound transitions.

For the electron wave function in the bound state, we have the same g and f as in Sec. III. For the electron wave function in the continuum, we normalize the wave function in the energy scale such that[21]

$$\int_0^\infty r^2\, dr\, R_W(r)\int_{W-\Delta W^R}^{W+\Delta W} R_{W'}(r)\, dW' = 1. \tag{4.7}$$

Then it is seen that

$$g(\tilde{\omega}) = (E/p)^{1/2}\Omega(\tilde{\omega})N_p(E\kappa/\gamma)^{-1/2} \tag{4.8}$$

where N_p = normalization factor in the p scale,

$$\Omega(\tilde{\omega}) = (2pr)^\lambda e^{-\pi\eta/2}\frac{|\Gamma(\lambda+1+i\eta)|}{\Gamma(2\lambda+2)}e^{-ipr} \times {}_1F_1(\lambda+1-i\eta,2\lambda+2,2ipr). \tag{4.9}$$

Let us note that

$$e^{-ipr}{}_1F_1(\lambda+1-i\eta,2\lambda+2,2ipr) = e^{ipr}{}_1F_1(\lambda+1+i\eta,2\lambda+2,-2ipr) \tag{4.10}$$

because of Kummer's relation for the confluent hypergeometric function. Thus $\Omega(\tilde{\omega})$ is pure real.

The $f(\tilde{\omega})$ is defined in exactly the same way and is also pure real. The radial wave functions for the continuum therefore satisfy the same reality conditions as those obtained by Rose.[27] However, they are simpler than Rose's in that each f and g contains only one term of a confluent hypergeometric function.

We also find that our normalization differs from Rose[22] in that we have an extra factor containing $E^{1/2}$. This comes from the relativistic equation

$$E^2 = m^2 + p^2 \tag{4.11}$$

so that

$$E\, dE = p\, dp. \tag{4.12}$$

In changing from the p scale to the E scale, one has[21]

$$R_E = \left(\frac{dE}{dp}\right)^{-1/2} R_p = \left(\frac{p}{E}\right)^{-1/2} R_p. \tag{4.13}$$

Therefore, the $E^{1/2}$ term must enter into the normalization. Now Bethe and Salpeter were considering the nonrelativistic case where $E = k^2/2$, and $dE/dk = k$. Therefore in the nonrelativistic case there is no $E^{1/2}$ term in the normalization. Rose left out the $E^{1/2}$ term because he used the nonrelativistic relation, but not the relativistic one. The term $E^{1/2}$, however, is canceled by the term $E^{-1/2}$ at the end of Eq. (4.8), which comes from the normalization of the first-order Dirac equation.

Thus in the case of internal conversion, all we have to do is to evaluate the integral K.

$$K = \int_0^\infty e^{-\mu_1' r + ipr} r^{\lambda+\lambda'+2} h_L^{(1)}(kr) L_{n_r'}^{2\lambda'+1}(2\mu_1' r) \times {}_1F_1(\lambda+1+i\eta,2\lambda+2,-2ipr)dr. \tag{4.14}$$

We have obtained a closed-form expression for K. The method makes use of the generating function for the generalized Laguerre polynomial, and the integral representation for the confluent hypergeometric function. The final result is given as follows.

Again we distinguish between two cases: (a) $\mu_1'/(k+p) < 1$, and (b) $\mu_1'/(k+p) > 1$. Then in case (a) we have

$$K=\sum_{\alpha,\beta,\nu}(-i)^{L+1}\frac{(L+\nu)!}{\nu!(L-\nu)!}\left(\frac{1}{2}i\right)^{\nu}k^{-1-\nu}\Gamma(2+\lambda+\lambda'-\nu)$$
$$-\binom{-(2+\lambda+\lambda'-\nu)}{\beta}(-\mu_1)^{\beta}[i(p+k)]^{-(2+\lambda+\lambda'-\nu+\beta)}(-1)^{n_r'-\alpha}$$
$$\times\binom{\beta}{\alpha}\binom{-2\lambda'-2-\beta}{n_r'-\alpha}{}_2F_1\left(2+\lambda+\lambda'+\beta-\nu,\lambda+1+i\eta,2\lambda+2,\frac{2p}{p+k}\right). \tag{4.15}$$

For case (b), $(k+p)/\mu_1'<1$, we have

$$K=\sum_{\alpha'\beta'\nu}(-i)^{L+1}\frac{(L+\nu)!}{\nu!(L-\nu)!}\left(\frac{1}{2}i\right)^{\nu}k^{-1-\nu}\Gamma(2+\lambda+\lambda'-\nu)\binom{-(2+\lambda+\lambda'-\nu)}{\beta'}$$
$$\times(-\mu_1)^{-(2+\lambda+\lambda'-\nu+\beta')}[i(p+k)]^{\beta'}(-1)^{n_r'-\alpha'}\binom{-(2+\lambda+\lambda'-\nu+\beta')}{\alpha'}$$
$$\times\binom{-\lambda'+\lambda-\nu+\beta'}{n_r'-\alpha'}{}_2F_1\left(2+\lambda+\lambda'-\nu+\beta',\lambda+1+i\eta,2\lambda+2,\frac{2p}{p+k}\right). \tag{4.16}$$

Next we use Kummer's relation[23] to transform the argument of the hypergeometric function in (4.15) and (4.16) from z to $1/z$. Then we have

$${}_2F_1(2+\lambda+\lambda'+\beta-\nu,\lambda+1+i\eta,2\lambda+2,2p/(p+k))=\frac{\Gamma(2\lambda+2)\Gamma(\nu-\beta-\lambda'-1+i\eta)}{\Gamma(\lambda-\lambda'+\nu-\beta)\Gamma(\lambda+1+i\eta)}\left(-\frac{p+k}{2p}\right)^{2+\lambda+\lambda'+\beta-\nu}$$
$$\times{}_2F_1(2+\lambda+\lambda'-\nu+\beta,\lambda'-\lambda-\nu+\beta+1,2+\lambda'+\beta-\nu-i\eta,(p+k)/2p)$$
$$+\frac{\Gamma(2\lambda+2)\Gamma(\beta+\lambda-\nu+1-i\eta)}{\Gamma(\lambda+1-\eta)\Gamma(2+\lambda+\lambda'+\beta-\nu)}\left(-\frac{p+k}{2p}\right)^{\lambda+1+i\eta}$$
$$\times{}_2F_1(-\lambda+i\eta,\lambda+1+i\eta,\nu-\beta-\lambda'+i\eta,(p+k)/2p). \tag{4.17}$$

Note that if one has a term $(-1)^{\lambda}$ where λ is irrational, we write it as $e^{i\pi\lambda}$. The two hypergeometric functions in Eq. (4.17) are now convergent because $(p+k)/2p<1$. Thus we have obtained an exact evaluation of the matrix elements in internal conversion.

For the photoeffect, we have[24]

$$d\sigma=(2\pi)^{-2}|M|^2\delta(\tilde{E})d^3p, \tag{4.18}$$

$$M=-e(2\pi)^{1/2}k^{-1/2}\int d^3r\,\psi_p^{*}\boldsymbol{\alpha}\cdot\boldsymbol{\epsilon}e^{i\mathbf{k}\cdot\mathbf{r}}\psi_B, \tag{4.19}$$

$$\tilde{E}=(p^2+m^2)^{1/2}-k-E_B. \tag{4.20}$$

The bound state wave function ψ_B is defined in exactly the same way as before. The continuum wave function is obtained by a superposition of solutions from the Dirac–Coulomb equation such that it takes the form of a plane wave plus the scattered part asymptotically. We construct such a wave function in the same fashion as indicated by one of us in a previous paper.[25] The four components of ψ_p have the following form, where the small components occupy the first two rows and the large components occupy the third and fourth rows. Thus for $\mu=\frac{1}{2}$, we have in the first row,

$$\sum_l -i^{l+1}\left(\frac{E(\kappa/\gamma)-m}{E(\kappa/\gamma)+m}\right)^{1/2}P_l(\cos\theta)[l(l+1)]^{1/2}$$
$$\times(e^{i\zeta_{l+1}}R_{\rho_{l+1}}+e^{i\zeta_l}R_{\rho_l-1}); \tag{4.21}$$

in the second row,

$$\sum_l i^{l+1}\left(\frac{E(\kappa/\gamma)-m}{E(\kappa/\gamma)+m}\right)^{1/2}P_l^1(\cos\theta)e^{i\phi}[l(l+1)]^{1/2}$$
$$\times(-e^{i\zeta_{l+1}}R_{\rho_{l+1}}+e^{i\zeta_l}R_{\rho_l-1}); \tag{4.22}$$

in the third row,

$$\sum_l i^lP_l(\cos\theta)\left[(l+1)l^{i\zeta_{l+1}}R_{\rho_{l+1}-1}+le^{i\zeta_l}R_{\rho_l}\right]; \tag{4.23}$$

and in the fourth row,

$$\sum_l i^lP_l^1(\cos\theta)e^{i\phi}\left[e^{i\zeta_{l+1}}R_{\rho_{l+1}-1}\right]-e^{i\zeta_l}R_{\rho_l}; \tag{4.24}$$

where

$$R_\lambda=\frac{(2pr)^\lambda|\Gamma(\lambda+1+i\eta)|}{\Gamma(2\lambda+2)\Gamma(1+i\eta)}$$
$$\times e^{-ipr}{}_1F_1(\lambda+1-i\eta,2\lambda+2,2ipr), \tag{4.25}$$

$$\rho_l=(l^2-\alpha^2Z^2)^{1/2}, \tag{4.26}$$

$$\exp(2i\zeta_l)=e^{-i\pi(\rho_l-l)}\frac{\Gamma(\rho_l+1+i\eta)}{\Gamma(\rho_l+1-i\eta)}, \tag{4.27}$$

$$\exp(2i\zeta_{l+1})=e^{-i\pi(\rho_{l+1}-l-1)}\frac{\Gamma(\rho_{l+1}+i\eta)}{\Gamma(\rho_{l+1}-i\eta)}. \tag{4.28}$$

This construction is in complete accordance with the nonrelativistic case where the Schrödinger equation is used. Let us note that in the nonrelativistic limit, the third row reproduces the plane wave plus the scattered part obtained from the Schrödinger equation in parabolic coordinates.[26] The fourth row vanishes, and the small components in the first two rows are negligible because of the factors under the square root.

It has been shown in Ref. 25 that this construction gives the correct result for the Coulomb scattering of fast electrons in agreement with McKinley and Feshbach.[27] In each of

Eqs. (4.21) to (4.24) the first term refers to $\tilde{\omega} = -1$ and the second term refers to $\tilde{\omega} = 1$. In terms of spinors χ_κ^μ and the quantum numbers j and $\tilde{\omega}$, we can rewrite the third and fourth rows as

$$\sqrt{4\pi}\sum_{j,\tilde{\omega}} i^{j-1/2}\tilde{\omega}\left(j+\frac{1}{2}\tilde{\omega}\right)^{1/2} \times e^{i\zeta_{j+\tilde{\omega}/2}} R_{\rho_{j+1/2+(1/2)(\tilde{\omega}-1)}}\chi_\kappa^{1/2}. \tag{4.29}$$

The first and second rows for the small component can then be written accordingly. In a similar manner we can construct a plane wave with $\mu = -\frac{1}{2}$. This is for the large component,

$$\sqrt{4\pi}\sum_{j,\tilde{\omega}} i^{j-1/2}\left(j+\frac{1}{2}\tilde{\omega}\right)^{1/2} \times e^{i\zeta_{j+(1/2)\tilde{\omega}}} R_{\rho_{j+1/2+(1/2)(\tilde{\omega}-1)}}\chi_\kappa^{-1/2}. \tag{4.30}$$

and for the small component,

$$\sqrt{4\pi}\sum_{j,\tilde{\omega}} \tilde{\omega}\left(\frac{E(\kappa/\gamma)-m}{E(\kappa/\gamma)+m}\right)^{1/2} i^{j+1/2}\left(j+\frac{1}{2}\tilde{\omega}\right)^{1/2} \times e^{i\zeta_{j+(1/2)\tilde{\omega}}} R_{\rho_{j+1/2-(1/2)(\tilde{\omega}+1)}}\chi_{-\kappa}^{-1/2}. \tag{4.31}$$

Combining Eqs. (4.21)–(4.24), (4.30), and (4.31), we can write the electron wave function in the z direction as for the small component:

$$\sqrt{4\pi}\sum_{j,\tilde{\omega},\mu} \left(\frac{E(\kappa/\gamma)-m}{E(\kappa/\gamma)+m}\right)^{1/2} \tilde{\omega} i^{j+1/2}\left(j+\frac{1}{2}\tilde{\omega}\right)^{1/2} \times e^{i\zeta_{j+(1/2)\tilde{\omega}}}(\tilde{\omega})^{\mu+1/2} R_{\rho_{j+1/2-(1/2)(\tilde{\omega}+1)}}\chi_\kappa^\mu; \tag{4.32}$$

and for the large component,

$$\sqrt{4\pi}\sum_{j,\tilde{\omega},\mu} i^{j-1/2}\left(j+\frac{1}{2}\tilde{\omega}\right)^{1/2} e^{i\zeta_{j+(1/2)\tilde{\omega}}}(\tilde{\omega})^{\mu+1/2} \times R_{\rho_{j+1/2+(1/2)(\tilde{\omega}-1)}}\chi_\kappa^\mu; \tag{4.33}$$

where $\mu = \pm\frac{1}{2}$. Our wave function is normalized to unit amplitude.

So far the wave function is propagating in the z direction. In the case of the photoeffect the electron goes in the direction $\mathbf{p}'$, different in general from the direction of the photon $\mathbf{k}$. Though it is possible to rotate the electron wave function, it is much easier to rotate the photon wave function. This is what we shall do in this paper.

The photon vector potential $\mathbf{A}$ is thus expanded in multipoles

$$\mathbf{A} = \pi\sum_{L=1}^{\infty}\sum_{M=-L}^{L} i^L \mathscr{D}^L_{MP}(\phi',\theta',0) \times [\mathbf{A}_{LM}(m) + iP\,\mathbf{A}_{LM}(e)], \tag{4.34}$$

where P takes the values of $+1$ or -1, and $\mathbf{A}_{LM}(m)$ and $\mathbf{A}_{LM}(e)$ are given as follows:

$$\mathbf{A}_{LM}(m) = -(2/\pi)^{1/2} j_L(kr)\mathbf{Y}^M_{LL}(\theta,\phi)(2L+1)^{1/2}, \tag{4.35}$$

$$\mathbf{A}_{LM}(e) = (2/\pi)^{1/2}[(L+1)^{1/2} j_{L-1}(kr)\mathbf{Y}_L{}^M{}_{L-1}(\theta,\phi) - L^{1/2} j_{L+1}(kr)\mathbf{Y}_L{}^M{}_{L+1}(\theta,\phi)]. \tag{4.36}$$

The $\mathbf{Y}^M_{LL'}(\theta,\phi)$ are the vector spherical harmonics discussed by Blatt and Weisskopf.[28] They are further expressed by Brink and Satchler[29] as follows:

$$[L(L+1)]^{1/2}\mathbf{Y}^M_{LL} = \mathbf{L}Y_{LM}, \tag{4.37}$$

$$[L(2L+1)]^{1/2} r^{L-1}\mathbf{Y}_L{}^M{}_{L-1} = \nabla(r^L Y_{LM}), \tag{4.38}$$

$$[(L+1)(2L+1)]^{1/2} r^{-L-2}\mathbf{Y}_L{}^M{}_{L+1} = \nabla(r^{-L-1}Y_{LM}). \tag{4.39}$$

In Eq. (4.34), θ' is the angle between $\mathbf{p}'$ and $\mathbf{k}$, where $\mathbf{p}'$ is now the z axis, and ϕ' is the azimuthal angle of $\mathbf{k}$. The angular integration can then be evaluated easily.

The plane wave construction of the electron beam has been done by other authors, for example, Akheizer and Beretetskii[30] and Johnson and Deck.[31] However, their expression seems to be more complicated. These authors rotate their electron wave function to its direction $\mathbf{p}'$ and hence obtain a more complicated expression. We are able to avoid this process by using $\mathbf{p}'$ as the z axis but rotating the photon vector potential.

We now proceed to the exact evaluation of the matrix elements in the photoeffect. Without loss of generality, we take a' to be the continuum and a to be the bound state. The photon vector potential $\mathbf{A}$ has been expanded in terms of electric and magnetic multipoles. The selection rules that $j+\frac{1}{2}\tilde{\omega}+j'+\frac{1}{2}\tilde{\omega}'+L =$ odd for the magnetic part and even for the electric part still apply. After performing the angular integration, we obtain the matrix elements as follows:

$$\begin{aligned}\langle a|\boldsymbol{\alpha}\cdot\mathbf{A}_{LM}(m)|a'|\rangle = {} & (\kappa+\kappa')(-1)^{L-j+j'+\mu+3/2}\sqrt{\frac{(2j+1)(2j'+1)}{2L(L+1)}}\frac{(2L+1)}{\pi} \\ & \times\int dr\, r^2 j_L(kr)\begin{pmatrix} j & j' & L \\ -\frac{1}{2} & \frac{1}{2} & 0\end{pmatrix}\left\{\begin{pmatrix} j & j' & L \\ -\mu & \frac{1}{2} & M\end{pmatrix}\left[\tilde{\omega}\, g^*_\kappa i^{j'+1/2}\left(\frac{E'(\kappa'/\gamma')-m}{E'(\kappa'/\gamma')+m}\right)^{1/2}\right.\right. \\ & \times((j'-\tfrac{1}{2})^{1/2}e^{i\zeta_{j'}-1/2}R_{\rho_{j'-1/2}-1} + (j'+\tfrac{1}{2})^{1/2}e^{i\zeta_{j'}+1/2}R_{\rho_{j'+1/2}}) \\ & \left.+\, i^{j'+1/2}f^*_\kappa(-(j'-\tfrac{1}{2})^{1/2}e^{i\zeta_{j'}-1/2}R_{\rho_{j'-1/2}} + (j'+\tfrac{1}{2})^{1/2}e^{i\zeta_{j'}+1/2}R_{\rho_{j'+1/2}-1})\right] \\ & +\begin{pmatrix} j & j' & L \\ -\mu & -\frac{1}{2} & M\end{pmatrix}\left[-\tilde{\omega}g^*_\kappa i^{j'+1/2}\left(\frac{E'(\kappa'/\gamma')-m}{E'(\kappa'/\gamma')+m}\right)^{1/2}\right. \\ & \times(-(j'-\tfrac{1}{2})^{1/2}e^{i\zeta_{j'}-1/2}R_{\rho_{j'-1/2}-1} + (j'+\tfrac{1}{2})^{1/2}e^{i\zeta_{j'}+1/2}R_{\rho_{j'+1/2}}) \\ & \left.\left.-\, i^{j'+1/2}f^*_\kappa((j'-\tfrac{1}{2})^{1/2}e^{i\zeta_{j'}-1/2}R_{\rho_{j'-1/2}} + (j'+\tfrac{1}{2})^{1/2}e^{i\zeta_{j'}+1/2}R_{\rho_{j'+1/2}-1})\right]\right\},\end{aligned} \tag{4.40}$$

with $j+j'+\frac{1}{2}(\tilde{\omega}+\tilde{\omega}')+L=\text{odd}$;

$$\langle a|A_{LM}(e)|a'\rangle=(-1)^{L+\mu+3/2-j+j'}\sqrt{\frac{(2j+1)(2j'+1)}{2\pi^2}}$$
$$\times\begin{pmatrix} j & j' & L\\ -\frac{1}{2} & \frac{1}{2} & 0\end{pmatrix}\int r^2\,dr\left\{\begin{pmatrix} j & j' & L\\ -\mu & \frac{1}{2} & M\end{pmatrix}\left[\left\{\left(\frac{L+1}{L}\right)^{1/2}j_{L-1}(kr)(-L-\kappa'+\kappa)-\left(\frac{L}{L+1}\right)^{1/2}\right.\right.\right.$$
$$\left.\times j_{L+1}(kr)(L+1+\kappa-\kappa')\right\}\tilde{\omega}g_\kappa^* i^{j'+1/2}\left(\frac{E'(\kappa'/\gamma')-m}{E'(\kappa'/\gamma')+m}\right)^{1/2}$$
$$\times\left\{\sqrt{j'-1/2}\,e^{i\zeta_{j'-1/2}}R_{\rho_{j'-1/2}-1}+\sqrt{j'+1/2}\,e^{i\zeta_{j'-1/2}}R_{\rho_{j'+1/2}}\right\}$$
$$-\left\{\left(\frac{L+1}{L}\right)^{1/2}j_{L-1}(kr)(-L-\kappa+\kappa')-\left(\frac{L}{L+1}\right)^{1/2}j_{L+1}(kr)(L+1)+\kappa'-\kappa)\right\}$$
$$\left.\times if_\kappa^* i^{j'-1/2}\left\{-\sqrt{j'-\tfrac{1}{2}}\,e^{i\zeta_{j'-1/2}}R_{\rho_{j'-1/2}}+\sqrt{j'+\tfrac{1}{2}}\,e^{i\zeta_{j'+1/2}}R_{\rho_{j'+1/2}-1}\right\}\right]$$
$$+\begin{pmatrix} j & j' & L\\ -\mu & -\frac{1}{2} & M\end{pmatrix}\left[\left\{\left(\frac{L+1}{L}\right)^{1/2}j_{L-1}(kr)(-L-\kappa'+\kappa)-\left(\frac{L}{L+1}\right)^{1/2}j_{L+1}(kr)(L+1+\kappa-\kappa')\right\}\right.$$
$$\times\tilde{\omega}g_\kappa^* i^{j'+1/2}\left(\frac{E'(\kappa'/\gamma')-m}{E'(\kappa'/\gamma')+m}\right)^{1/2}\left\{-\sqrt{j'-1/2}\,e^{i\zeta_{j'-1/2}}R_{\rho_{j'-1/2}-1}+\sqrt{j'+1/2}\,e^{i\zeta_{j'+1/2}}R_{\rho_{j'+1/2}}\right\}$$
$$-\left\{\left(\frac{L+1}{L}\right)^{1/2}j_{L-1}(kr)(L-\kappa+\kappa')-\left(\frac{L}{L+1}\right)^{1/2}j_{L+1}(kr)(L+1+\kappa'-\kappa)\right\}if_\kappa^* i^{j'-1/2}$$
$$\left.\left.\times\left\{\sqrt{j'-1/2}\,e^{i\zeta_{j'-1/2}}R_{\rho_{j'-1/2}}+\sqrt{j'+1/2}\,e^{i\zeta_{j'+1/2}}R_{\rho_{j'+1/2}-1}\right\}\right]\right\},\tag{4.41}$$

with $j+j'+\frac{1}{2}(\tilde{\omega}+\tilde{\omega}')+L=\text{even}$.

These formulas are very similar to those in Sec. III for bound–bound transitions, but now they represent the exact matrix elements in the photoeffect. In Eqs. (4.40) and (4.41), given μ and M, only one term in

$$\begin{pmatrix} j & j' & L\\ -\mu & \frac{1}{2} & M\end{pmatrix}\text{ and }\begin{pmatrix} j & j' & L\\ -\mu & -\frac{1}{2} & M\end{pmatrix}$$

will survive. Thus these equations are not as complicated as they look.

It remains for us to obtain the radial integrals in (4.40) and (4.41). This can be summarized in the integral $\mathcal{J}$:

$$\mathcal{J}=\int_0^\infty L_{n_r}^{2\lambda+1}(2\mu_1 r)j_L(kr)\,{}_1F_1(\lambda'+1+i\eta,2\lambda'+2,-2ipr)$$
$$\times e^{-\mu_1 r+ip'r}r^{\lambda+\lambda'+2}dr.\tag{4.42}$$

The final results are as follows. We distinguish between two cases. (a) $\mu_1<k<p'+k$, or $k<\mu_1<p'+k$, and (b) $k<p'+k<\mu_1$.

For case (a), we obtain

$$\mathcal{J}=\sum_{\alpha,\beta}\frac{\sqrt{\pi}K^L\Gamma(a+1)}{2^{L+1}\Gamma(L+\frac{3}{2})}\begin{pmatrix}-a-1\\ \beta\end{pmatrix}\mu_1^\beta(ic)^{-a-1-\beta}(-1)^{\alpha+\beta}\begin{pmatrix}-2\lambda-2-\beta\\ \alpha\end{pmatrix}$$
$$\times\begin{pmatrix}\beta\\ n_r-\alpha\end{pmatrix}F_2\left(a+\beta+1,L+3/2,\lambda'+1+i\eta,2L+3,2\lambda'+2,\frac{-2k}{p'-k},\frac{2p'}{p'-k}\right),\tag{4.43}$$

$$a=L+\lambda+\lambda'+2,\tag{4.44}$$

$$c=p'-k.\tag{4.45}$$

For case (b), we have

$$\mathcal{J}=\sum_{\alpha',\beta'}\frac{\sqrt{\pi}\Gamma(a+1)k^L}{2^{L+1}\Gamma(L+3/2)}\begin{pmatrix}-1-a\\ \beta'\end{pmatrix}(-1)^{-a-1-\beta'}(ic)^{\beta'}\begin{pmatrix}-2\lambda-1+a-\beta'\\ \alpha'\end{pmatrix}$$
$$\times(-1)^{\alpha'+a+1+\beta'}\begin{pmatrix}-a-\beta'-1\\ n_r-\alpha'\end{pmatrix}F_2\left(-\beta',L+3/2,\lambda'+1+i\eta,2L+3,2\lambda'+2,\frac{-2k}{p'-k},\frac{2p'}{p'-k}\right).\tag{4.46}$$

Next we write[32]

$$F_2(\alpha\beta\beta'\gamma\gamma',xy)=(1-y)^{-\alpha}F_2\left(\alpha,\beta,\gamma',\beta',\gamma,\gamma',\frac{x}{1-y},\frac{-y}{1-y}\right).\tag{4.47}$$

Then we expand the F_2 as follows:

$$F_2(\alpha\beta\beta'\gamma\gamma',xy)=\sum_{m=0}^{\infty}\frac{(\alpha)_m(\beta)_m}{(\gamma)_m m!}\,{}_2F_1(\alpha+m,\beta',\gamma',y)x^m. \tag{4.48}$$

Finally we use Kummer's relation for case (a)

$${}_2F_1(abcz)=\frac{\Gamma(c)\Gamma(b-a)}{\Gamma(c-a)\Gamma(b)}(-z)^a{}_2F_1(a,a+1-c,a+1-b,1/z)+\frac{\Gamma(c)\Gamma(a-b)}{\Gamma(c-b)\Gamma(a)}(-z)^b{}_2F_1(b+1-c,b,b+1-a,1/z). \tag{4.49}$$

For case (b) we use

$$\begin{aligned}{}_2F_1(abcz)=&\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}\,{}_2F_1(a,b,a+b+1-c,1-z)\\&+\frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}(1-z)^{c-a-b}\,{}_2F_1(c-a,c-b,c+1-a-b,1-z).\end{aligned} \tag{4.50}$$

Then we obtain for case (a)

$$\begin{aligned}\mathscr{J}=&\sum_{\alpha,\beta,m}\frac{\sqrt{\pi}k^L\Gamma(a+1)}{2^{L+1}\Gamma(L+3/2)}\binom{-a-1}{\beta}\mu_i^{\beta}i^{\alpha+\beta+1}(-1)^{\alpha+\beta}\\&\times\left(\frac{1}{p'+k}\right)^{L+\lambda+\lambda'+\beta+3}\binom{-2\lambda-2-\beta}{\alpha}\binom{\beta}{n_r-\alpha}\left(\frac{2k}{p'+k}\right)^m\\&\times\frac{(a+\beta+1)_m(L+3/2)_m}{(2L+3)_m m!}\left\{\frac{\Gamma(2\lambda'+2)\Gamma(\lambda'-a-\beta-m-i\eta)}{\Gamma(2\lambda'+1-a-\beta-m)\Gamma(\lambda'+1-i\eta)}\right.\\&\times\left(\frac{p'+k}{-2p'}\right)^{a+\beta+m+1}{}_2F_1\left(a+\beta+m+1,a+\beta+m-2\lambda',a+\beta+1+m-\lambda'+i\eta,\frac{p'+k}{2p'}\right)\\&+\frac{\Gamma(2\lambda'+2)\Gamma(a+\beta+m-\lambda'+i\eta)}{\Gamma(\lambda'+1+i\eta)\Gamma(a+1-\beta+m)}\left(\frac{p'+k}{-2p'}\right)^{\lambda'+1-i\eta}\\&\left.\times{}_2F_1\left(-\lambda'-i\eta,\lambda'+1-i\eta,\lambda'+1-a-\beta-m-i\eta,\frac{p'+k}{2p'}\right)\right\}.\end{aligned} \tag{4.51}$$

For case (b) we obtain

$$\begin{aligned}\mathscr{J}=&\sum_{\alpha'\beta'm}\frac{\sqrt{\pi}\Gamma(a+1)}{2^{L+1}\Gamma(L+3/2)}k^L\binom{-a-1}{\beta'}\mu_1^{-L-\lambda-\lambda'-3-\beta'}(-i)^{\beta'}(p'+k)^{\beta'}\\&\times(-1)^{\alpha'+\beta'+a+1}\binom{-2\lambda-1+a+\beta'}{\alpha'}\binom{-a-\beta-1}{n_r-\alpha'}\frac{(-\beta')_m(L+3/2)_m}{(2L+3)_m m!}\left(\frac{2k}{p'+k}\right)^m\\&\times\left\{\frac{\Gamma(2\lambda'+2)\Gamma(\lambda'+1+\beta'-m+i\eta)}{\Gamma(2\lambda'+2+\beta'-m)\Gamma(\lambda'+1+i\eta)}\,{}_2F_1\left(-\beta'+m,\lambda'+1-i\eta,-\beta'+m-\lambda'-i\eta,\frac{-(p'-k)}{p'-k}\right)\right.\\&\times\frac{\Gamma(2\lambda'+2)\Gamma(-\beta'-1-\lambda'+m-i\eta)}{\Gamma(-\beta'+m)\Gamma(\lambda'+1-i\eta)}\left(\frac{-(p'-k)}{p'+k}\right)^{\lambda'+1+\beta'-m+i\eta}\\&\left.\times{}_2F_2\left(2\lambda'+2+\beta'-m,\lambda'+1+i\eta,\lambda'+2+\beta'-m+i\eta,\frac{-(p'-k)}{p'+k}\right)\right\}.\end{aligned} \tag{4.52}$$

Equation (4.51) is convergent for the following reasons. The summation over L is convergent because $k/(p'+k)<1$. The summation over β is convergent because $\mu_1/(p'+k)<1$ by virtue of the case (a) condition. The summation over m is convergent because $2k/(p'+k)<1$, and the summation over λ' is convergent because $p'/(p'+k)<1$.

Equation (4.52) is convergent for the following reasons. The summation over L is convergent because $k/\mu_1<1$ by virtue of the case (b) condition. The summation over β' and m goes as follows:

$$\left(\frac{p'-k}{\mu_1}\right)^{\beta'-m}\left(\frac{p'+k}{\mu_1}\right)^m\left(\frac{2k}{p'+k}\right)^m.$$

All three factors are less than unity. The first one because $p'-k<p'+k<\mu_1$. Thus the summation over m and β' is convergent. The summation over λ' is convergent because $p'<p'+k<\mu_1$.

Finally we wish to remark that the three conditions with two from case (a) and one from case (b) exhaust all possible relations between μ_1, k, and $p'+k$. Thus the matrix elements for the photoeffect have been obtained in closed analytic form.

V. CONTINUUM–CONTINUUM TRANSITIONS

We now discuss the third case, continuum–continuum transitions. Physically, this corresponds to bremsstrahlung

and pair creation in the presence of a Coulomb field.

It is interesting to note that this case has been essentially solved by Rozics and Johnson.[33] However, we have made some improvements on their solutions. First, with regard to the angular part, we have obtained a much simpler expression than theirs. Second, the electron wave function in the continuum used by us is also simpler than theirs. Third, we have put the final expresssion in such a way that it converges in all cases.

The differential cross section for this process is given by

$$d\sigma = \frac{\alpha}{(2\pi)^4}\frac{E_1}{p_1} p_2 E_2 k\, dk\, d\Omega_k \sum_{\text{pol}} \int d\Omega_2\, \epsilon |M|^2, \tag{5.1}$$

where

$$M = \int d^3 r\, \psi^+(a)\boldsymbol{\alpha}\cdot\boldsymbol{\epsilon} e^{i\mathbf{k}\cdot\mathbf{r}}\psi(a'). \tag{5.2}$$

The photon vector potential is expanded in exactly the same way as in Eq. (4.34). The electron wave functions which must behave as a plane wave plus a scattered part have also been constructed in Sec. IV. We now wish to introduce one more symbol $e^{i\delta_l}$. This is the phase shift for the complex conjugate of $e^{i\zeta_l}$, i.e.,

$$e^{i\delta_l} = (e^{i\zeta_l})^*. \tag{5.3}$$

It can be easily seen that $e^{i\delta_l}$ can be obtained from $e^{i\zeta_l}$ by changing i to $-i$. For example, if

$$e^{i2\xi_l} = e^{-i\pi(\rho_l - l)}\frac{\Gamma(\rho_l + 1 + i\eta)}{\Gamma(\rho_l + 1 - i\eta)}, \tag{5.4}$$

then

$$e^{i2\delta_l} = e^{i\pi(\rho_l - l)}\frac{\Gamma(\rho_l + 1 - i\eta)}{\Gamma(\rho_l + 1 + i\eta)}. \tag{5.5}$$

Now we have to rotate the wave functions to their proper directions.

We choose $\mathbf{p}_1$ to be the z axis, then the photon would have angles θ_k, ϕ_k, and p_2 would have angles θ_2, ϕ_2.

The rotated photon vector potential is written in exactly the same way as in Eq. (4.34).

$$\mathbf{A} = \pi \sum_{L=1}^{\infty}\sum_{M=-L}^{L} i^L \mathscr{D}^L_{MP}(\phi_k,\theta_k,0)[\mathbf{A}_{LM}(m) + iP\mathbf{A}_{LM}(e)]. \tag{5.6}$$

The rotation to be performed on the electron wave function is limited to the space coordinates only, the spin coordinates being unchanged. All primed quantities refer to the electron 1, and all unprimed quantities refer to the electron 2. It can be seen that the rotated wave function is the following: For the small component,

$$\sum_{j,\tilde{\omega},\mu,\tau} D^{j+1/2\tilde{\omega}}_{0,\mu-\tau}(\phi_2,\theta_2,0) i^{j+1/2}\left(\frac{E(\kappa/\gamma)-m}{E(\kappa/\gamma)+m}\right)^{1/2} 2\pi^{1/2}\left(j+\frac{1}{2}\tilde{\omega}\right)^{1/2}$$
$$\times \tilde{\omega}^{\mu+3/2} e^{i\zeta_{j+(1/2)\tilde{\omega}}} R_{\rho_{j+1/2}-(1/2)(\tilde{\omega}+1)} C\begin{pmatrix} j-\frac{1}{2}\tilde{\omega} & \frac{1}{2} & j \\ \mu-\tau & \tau & \mu \end{pmatrix} Y^{\mu-\tau}_{j-(1/2)\tilde{\omega}}\chi^{\tau}_{1/2}; \tag{5.7}$$

and for the large component,

$$\sum_{j,\tilde{\omega},\mu,\tau} D^{j+(1/2)\tilde{\omega}}_{0,\mu-\tau}(\phi_2,\theta_2,0) i^{j-1/2} 2\pi^{1/2}\left(j+\frac{1}{2}\tilde{\omega}\right)^{1/2}\tilde{\omega}^{\mu+1/2}$$
$$\times e^{i\zeta_{j+(1/2)\tilde{\omega}}} R_{\rho_{j+1/2}+(1/2)(\tilde{\omega}-1)} C\begin{pmatrix} j+\frac{1}{2}\tilde{\omega} & \frac{1}{2} & j \\ \mu-\tau & \tau & \mu \end{pmatrix} Y^{\mu-\tau}_{j+(1/2)\tilde{\omega}}\chi^{\tau}_{1/2}. \tag{5.8}$$

Our rotated wave function (5.7) and (5.8) is essentially the same as that of Johnson and Deck,[31] if one remembers that[34]

$$D^{l}_{0,m}(\alpha,\beta,\gamma) = (4\pi/(2l+1))^{1/2} Y^m_l(\beta,\gamma). \tag{5.9}$$

We can now evaluate the angular integration in M. We find

$$\langle a|\boldsymbol{\alpha}\cdot\mathbf{A}_{LM}(m)|a'\rangle = i^{j+j'}\tilde{\omega}^{\mu+1/2}\tilde{\omega}^{\mu'+1/2}(j+1/2\tilde{\omega})^{1/2}(j'+1/2\tilde{\omega})^{1/2}(2/\pi)^{1/2}$$
$$\times(2L+1)^{1/2}L^{-1/2}(L+1)^{-1/2}\sum_{\nu,\tau}\int j_L(kr)r^2\,dr(\kappa+\kappa')$$
$$\times(2L+1)^{1/2}(2j-\tilde{\omega}+1)^{1/2}(2j'+\tilde{\omega}+1)^{1/2}(4\pi)^{-1/2}$$
$$\times Y^{\mu'-\tau}_{j'+\frac{1}{2}\tilde{\omega}'}(\theta_2,\phi_2) C\begin{pmatrix} j-\frac{1}{2}\tilde{\omega} & j'+\frac{1}{2}\tilde{\omega}' & L \\ 0 & 0 & 0 \end{pmatrix} S_\nu e^{i\zeta_{j'+(1/2)\tilde{\omega}'}} e^{i\delta_{j+(1/2)\tilde{\omega}}}$$
$$\times\left\{R^*_{\rho_{j+1/2}+(1/2)(\tilde{\omega}-1)} R_{\rho_{j'+1/2}-(1/2)(\tilde{\omega}'+1)}\left(\frac{E'(\kappa'/\gamma')-m}{E'(\kappa/\gamma')+m}\right)^{1/2}\tilde{\omega}' - R^*_{\rho_{j+1/2}-(1/2)(\tilde{\omega}+1)}\right.$$
$$\left.\times R_{\rho_{j'+1/2}+(1/2)(\tilde{\omega}'-1)}\tilde{\omega}\left(\frac{E(\kappa/\gamma)-m}{E(\kappa/\gamma)+m}\right)^{1/2}\right\}, \tag{5.10}$$

with $j + j' + \frac{1}{2}(\tilde{\omega}+\tilde{\omega}') + L = \text{odd}$;

$$\langle a|\boldsymbol{\alpha}\cdot\mathbf{A}_{LM}(e)|a'\rangle = i^{j+j'}\tilde{\omega}^{\mu+1/2}\tilde{\omega}'^{\mu'+1/2}(j+1/2\tilde{\omega})^{1/2}(j'+1/2\tilde{\omega}')^{1/2}$$

$$\times\sum_{\nu,\tau}\int r^2\,dr(2j+\tilde{\omega}+1)^{1/2}(2j'+\tilde{\omega}'+1)^{1/2}2^{-1/2}\pi^{-1}Y^{\mu'-\tau}_{j'+(1/2)\tilde{\omega}'}(\theta_2,\phi_2)$$

$$\times C\begin{pmatrix} j+\frac{1}{2}\tilde{\omega} & j'+\frac{1}{2}\tilde{\omega}' & L \\ 0 & 0 & 0\end{pmatrix}S_\nu e^{i\xi_{j+(1/2)\tilde{\omega}'}}e^{i\delta_{j+(1/2)\tilde{\omega}}}\left\{\tilde{\omega}'\left(\frac{E'(\kappa'/\gamma')-m}{E'(\kappa'/\gamma')+m}\right)^{1/2}\right.$$

$$\times R^*_{\rho_{j+1/2}+(1/2)(\tilde{\omega}-1)}R_{\rho_{j+1/2}-(1/2)(\tilde{\omega}'+1)}\left[(L+1)^{1/2}L^{-1/2}j_{L-1}(kr)(-L-\kappa'+\kappa)\right.$$

$$\left.-L^{1/2}(L+1)^{-1/2}j_{L+1}(kr)(L+1+\kappa-\kappa')\right]+\tilde{\omega}\left(\frac{E(\kappa/\gamma)-m}{E(\kappa/\gamma)+m}\right)^{1/2}R^*_{\rho_{j+1/2}-(1/2)(\tilde{\omega}+1)}$$

$$\times R_{\rho_{j+1/2}+(1/2)(\tilde{\omega}'-1)}\left[(L+1)^{1/2}L^{-1/2}j_{L-1}(kr)(-L+\kappa'-\kappa)-L^{1/2}(L+1)^{-1/2}j_{L+1}(kr)\right.$$

$$\left.\left.\times(L+1-\kappa+\kappa')\right]\right\} \tag{5.11}$$

with $j+j'+\frac{1}{2}(\tilde{\omega}+\tilde{\omega}')+L=$ even; and

$$S_{\bar{\nu}}=(-1)^{\mu-\tau}\begin{pmatrix} j-\frac{1}{2}\tilde{\omega} & j'+\frac{1}{2}\tilde{\omega}' & L \\ \tau-\mu & \mu'-\tau & \mu-\mu'\end{pmatrix}C\begin{pmatrix} j-\frac{1}{2}\tilde{\omega} & \frac{1}{2} & j \\ \mu-\tau & \tau & \mu\end{pmatrix}C\begin{pmatrix} j'+\frac{1}{2}\tilde{\omega}' & \frac{1}{2} & j' \\ \mu'-\tau & \tau & \mu'\end{pmatrix}, \tag{5.12}$$

$$S_{\nu}=(-1)^{\mu-\tau}\begin{pmatrix} j+\frac{1}{2}\tilde{\omega} & j'+\frac{1}{2}\tilde{\omega}' & L \\ \tau-\mu & \mu'-\tau & \mu-\mu'\end{pmatrix}C\begin{pmatrix} j+\frac{1}{2}\tilde{\omega} & \frac{1}{2} & j \\ \mu-\tau & \tau & \mu\end{pmatrix}C\begin{pmatrix} j'+\frac{1}{2}\tilde{\omega}' & \frac{1}{2} & j' \\ \mu'-\tau & \tau & \mu'\end{pmatrix}. \tag{5.13}$$

Let us now perform the integration over θ_2,ϕ_2. This is obtained by taking $|M|^2$ and then integrating over $d\Omega_2$. However, because of the orthonormality property of Y, we have

$$\int_0^{2\pi}\int_0^{\pi}Y^{*m}_l(\theta_2,\phi_2)Y^{m'}_{l'}(\theta_2,\phi_2)\sin\theta_2\,d\theta_2\,d\phi_2=\delta_{ll'}\delta_{mm'}. \tag{5.14}$$

Hence the term $Y^{\mu'-\tau}_{j'+(1/2)\tilde{\omega}'}(\theta_2,\phi_2)$ becomes either 0 or 1 after the integration. The summation over τ in $S_{\bar{\nu}}$ and S_ν can then be performed, leading to a product of a $3-j$ symbol and a $6-j$ symbol as before.

It remains for us to evaluate the radial integral, which is of the form I:

$$I=\int_0^{\infty}j_{L+1}(kr)R^*_{\lambda}(2ipr)R_{\lambda'}(2ip'r)r^2\,dr, \tag{5.15}$$

where R has been defined in (4.25).

The basic structure of the radial integral is the product of three nonterminating hypergeometric functions. This integral can be immediately obtained in terms of a generalized hypergeometric function of three variables. In fact, mathematically this result extends to n variables, i.e., the integral of n nonterminating confluent hypergeometric functions is equal to a generalized hypergeometric function of n variables. This formula can be found in Erdelyi *et al.*[35]:

$$\int_0^{\infty}t^{\nu-1}M_{k,\mu_1-1/2}(\alpha_1 t)\cdots M_{\kappa_n,\mu_n-1/2}e^{(\alpha_n t)}e^{-p_1 t}dt$$

$$=\alpha_1^{\mu_1}\cdots\alpha_n^{\mu_n}(p_1+A)^{-\nu-M}\Gamma(\mu+M)$$

$$\times F_A\left((\nu+M;\mu_1-\kappa_1\cdots\mu_n-\kappa_n,\right.$$

$$\left.2\mu_1\cdots2\mu_n;\frac{\alpha_1}{p_1+A}\cdots\frac{\alpha_n}{p_1+A}\right) \tag{5.16}$$

$$A=\tfrac{1}{2}(\alpha,+\cdots+\alpha_n), \tag{5.17}$$

$$M=\mu_1+\cdots+\mu_n, \tag{5.18}$$

$$\mathrm{Re}(\nu+M)>0,\quad \mathrm{Re}(p_1\pm\tfrac{1}{2}\alpha_q\pm\cdots\pm\tfrac{1}{2}\alpha_n)>0, \tag{5.19}$$

$$M_{\kappa\mu}(z)=z^{1/2+\mu}e^{-1/2z}{}_1F_1(\tfrac{1}{2}+\mu-\kappa,2\mu+1,z). \tag{5.20}$$

Since

$$R^*_{\lambda}(2ipr)=\frac{\Gamma(1-i\eta)}{\Gamma(1+i\eta)}R_{\lambda}(2ipr), \tag{5.21}$$

we can write the arguments of the three confluent hypergeometric functions in (5.15) as $2ikr$, $2ipr$, and $2ip'r$. In terms of formula (5.16), we have

$$\kappa_1=i\eta,\quad \kappa_2=i\eta',\quad \kappa_3=0, \tag{5.22}$$

$$\mu_1=\lambda+1,\quad \mu_2=\lambda'+1,\quad \mu_3=L+\tfrac{3}{2}, \tag{5.23}$$

$$\alpha_1=2ip,\quad \alpha_2=2ip',\quad \alpha_3=2ik, \tag{5.24}$$

$$p_1=0,\quad \nu=-\tfrac{1}{2}, \tag{5.25}$$

$$A=i(p+p'+k), \tag{5.26}$$

$$M=\lambda+\lambda'+L+\tfrac{7}{2}, \tag{5.27}$$

$$I=2^{\lambda+\lambda'+L+7/2}p^{\lambda}p'^{\lambda'}k^{L}[i(p+p'+k)]^{-\lambda-\lambda'-L-3}$$

$$\times\Gamma(\lambda+\lambda'+L+3)F_A(\lambda+\lambda'+L+3;$$

$$\times\lambda+1+i\eta,\lambda'+1+i\eta',L+3/2;$$

$$\times2\lambda+2,2\lambda'+2,2L+3,2p(p+p'+k),$$

$$\times2p'/(p+p'+k),2k/(p+p'+k)). \tag{5.28}$$

Without loss of generality, we can assume $p>p'$, $p>k$. Then we can distinguish between two cases. (a) $p<p'+k$. (b) $p>p'+k$. In case (a) the series F_A in (5.28) is already convergent since all three arguments are less than unity. Moreover, the summation over λ,λ', and L is also convergent, since the fractions $2p/(p+p'+k)$, $2p'/(p+p'+k)$, and $2k/(p+p'+k)$, being identical to the three arguments in F_A, are also less than unity.

In case (b) we write F_A as follows:

$$F_A(\alpha\beta,\beta_2\beta_3\gamma_1\gamma_2\gamma_3 x_1 x_2 x_3) = \sum_{m_1,m_2,m_3} \frac{(\alpha,m_1+m_2+m_3)(\beta_1,m_1)(\beta_2,m_2)(\beta_3,m_3)}{(\gamma_1,m_1)(\gamma_2,m_2)(\gamma_3,m_3)m_1!m_2!m_3!} x_1^{m_1}x_2^{m_2}x_3^{m_3}$$

$$= \sum_{m_1,m_2} {}_2F_1(\alpha+m_2+m_3,\beta_1,\gamma_1,x_1) \frac{(\alpha,m_2+m_3)(\beta_2,m_2)(\beta_3,m_3)}{(\gamma_2,m_2)(\gamma_3,m_3)} x_2^{m_2}x_3^{m_3}, \tag{5.29}$$

where

$$(\alpha,m) = \Gamma(\alpha+m)/\Gamma(\alpha). \tag{5.30}$$

Then we use Kummer's relation to transform ${}_2F_1$ in (5.29) from argument x_1 to $1/x_1$ according to (4.49). With this transformation, the integral I is expressed as a convergent series for the following reasons. The function F_A is now convergent since all its arguments are less than unity. The factor $(2p/(p+p'+k))^\lambda$, which was the only divergent term originally, is now canceled by the terms $(-z)^{-a}$ and $(-z)^{-b}$ in Eq. (4.49), since

$$(-z)^{-a} = \left(-\frac{2p}{p+p'+k}\right)^{-\lambda-\lambda'-L-3} \tag{5.31}$$

and

$$(-z)^{-b} = \left(-\frac{2p}{p+p'+k}\right)^{-\lambda-1-i\eta}. \tag{5.32}$$

Thus we have obtained a convergent series for the radial integral in all cases.

Let us now briefly compare our results with those of Rozics and Johnson. For the angular part, Rocizs and Johnson obtained a very complicated result because they rotated all three vectors $\mathbf{k}$, $\mathbf{p}_1$, and $\mathbf{p}_2$ to their respective directions. We are able to avoid this complication by choosing $\mathbf{p}_1$ as the z axis, and rotating $\mathbf{k}$ and $\mathbf{p}_2$ with respect to $\mathbf{p}_1$. We then obtain a single term $Y^{\mu'-\tau}_{j'+(1/2)\tilde{\omega}'}(\theta_2,\phi_2)$in the matrix element M with regard to the variables θ_2 and ϕ_2. This term can be easily integrated over $d\Omega_2$, because of the orthonormality property of Y. Thus our final result is expressed in terms of θ_k and ϕ_k, which are both contained in the rotation matrix $\mathscr{D}^L_{MP}(\phi_k,\theta_k,0)$ in Eq. (5.6). This result is much simpler than that of Rocizs and Johnson.

For the radial integral, our result is basically the same as Rozics and Johnson, but with two additional remarks. The first remark is that mathematically this integral can be extended to the product of n confluent hypergeometric functions. This result will be useful when one considers higher-order perturbation using the exact Coulomb wave function. The second remark is that in the present case we have obtained a convergent series valid for all conditions.

Thus the relativistic Kepler problem has been solved in a form which is mathematically exact.

[1] M. H. Han, Nuovo Cimento B **42**, 367 (1966); I. A. Malkin and V. I. Man'ko, JETP Lett. **2**, 146 (1965).
[2] A. O. Barut and G. L. Bornzin, J. Math. Phys. **12**, 841 (1971).
[3] M. K. F. Wong and H. Y. Yeh, Phys. Rev. D **25**, 3396 (1982).
[4] H. Bacry and J. L. Richard, J. Math. Phys. **8**, 2230 (1967).
[5] W. Miller, *Lie Theory and Special Functions* (Academic, New York, 1968).
[6] L. Armstrong, Phys. Rev. A **3**, 1546 (1971); J. Phys. Coll. C4, Suppl. **31** (11-12), c4-17 (1970).
[7] A. O. Barut and G. L. Bornzin, Phys. Rev. D **7**, 3018 (1973).
[8] A. O. Barut and R. Wilson, Phys. Rev. A **13**, 918 (1976).
[9] A. O. Barut, W. Rasmussen, and S. Salamó, Phys. Rev. D **10**, 622 (1974).
[10] A. O. Barut, W. Rasmussen, and S. Salamo, Phys. Rev. D **10**, 630 (1974).
[11] M. K. F. Wong and H. Y. Yeh, Phys. Rev. A **27**, 2305 (1983).
[12] F. A. Babushkin, Opt. Spectrosc. **13**, 77 (1962); Acta Phys. Polon. **25**, 749 (1964); Opt. Spectrosc. **19**, 1 (1965).
[13] J. H. Scofield, Phys. Rev. **179**, 9 (1968).
[14] H. R. Rosner and C. P. Bhalla, Z. Phys. **231**, 347 (1970).
[15] H. E. Moses, Phys. Rev. A **11**, 387 (1975).
[16] M. E. Rose, *Multipole Fields* (Wiley, New York, 1955).
[17] Z. P. Grant, J. Phys. B: Atom. Molec. Phys. **7**, 1458 (1974).
[18] M. K. F. Wong and H. Y. Yeh, Phys. Rev. A **27**, 2300 (1983).
[19] M. E. Rose, G. H. Goertzel, B. I. Spinrad, J. Harr, and P. Strong, Phys. Rev. **83**, 79 (1951).
[20] T. A. Green and M. E. Rose, Phys. Rev. **110**, 105 (1958).
[21] H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Atoms* (Springer, Berlin, 1957).
[22] M. E. Rose, *Relativistic Electron Theory* (Wiley, New York, 1961).
[23] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. 1, p. 106.
[24] R. H. Pratt, Phys. Rev. **117**, 1017 (1960).
[25] M. K. F. Wong, Phys. Rev. D **26**, 927 (1982).
[26] W. Gordon, Z. Phys. **48**, 180 (1928).
[27] W. A. McKinley and H. Feshbach, Phys. Rev. **74**, 1759 (1948).
[28] J. B. Blatt and V. F. Weisskopf, *Theoretical Nuclear Physics* (Wiley, New York, 1952).
[29] D. M. Brink and G. R. Satchler, *Angular Momentum* (Oxford U. P., London, 1968), 2nd ed.
[30] A. I. Akhiezer and V. B. Berestetskii, *Quantum Electrodynamics* (Interscience, New York, 1965).
[31] W. R. Johnson and R. T. Deck, J. Math. Phys. **3**, 319 (1962).
[32] Reference 23, Vol. 1, p. 240.
[33] J. D. Rozics and W. R. Johnson, Phys. Rev. **135**, B56 (1964).
[34] A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U. P., Princeton, 1957).
[35] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Tables of Integral Transforms* (McGraw-Hill, New York, 1954), Vol. 1, p. 216.

The direct linearization of a class of nonlinear evolution equations

P. M. Santini,[a] M. J. Ablowitz, and A. S. Fokas
Department of Mathematics and Computer Science, Clarkson College of Technology, Potsdam, New York 13676

(Received 8 December 1983; accepted for publication 3 February 1984)

The paper deals with the direct linearization, an approach used to generate particular solutions of the partial differential equations that can be solved through the inverse scattering transform. Linear integral equations are presented which enable one to find broad classes of solutions to certain nonlinear evolution equations in 1 + 1 and 2 + 1 dimensions.

PACS numbers: 02.30.Jr

I. INTRODUCTION

The partial differential equations (PDE's) associated with the inverse scattering transform (IST) (see, for instance, Ref. 1 for details) are structurally rich. It is clear from the work done in this field that these equations admit many kinds of approaches and studies. Broadly speaking (see, for example, Ref. 2), it is possible to group these approaches in two different classes: "algebraic properties" and "methods of solution."

Among the algebraic properties one can associate with each of these PDE's are the existence of an infinite hierarchy of equations characterized by the same linear problem; the existence of infinitely many conserved quantities and of a Hamiltonian (sometimes bi-Hamiltonian) structure; the possibility of associating with these equations a so-called Bäcklund transformation (BT)—i.e., a nonlinear transformation connecting different solutions, etc.

The methods of solution developed so far depend of course on the specific problem that one has to solve: the IST for instance is the appropriate tool to solve the initial value problem associated with these PDE's.

In order to generate particular solutions there exist other methods: e.g., the BT; the Hirota approach[1]; the Dressing method[3]; and the Riemann–Hilbert direct approach,[4] introduced by Zakharov and Shabat (ZS); etc. The Dressing method has been formulated via an integral equation of the Gel'fand–Levitan–Marchenko (GLM) type, and the Riemann–Hilbert (RH) direct approach is based on a local homogeneous RH problem, used to generate solutions of the PDE. Later we will discuss in some detail the RH method, used often as a reference point of our analysis.

In this paper we will concentrate on a particular method of solution: the direct linearization (DL), an approach used to generate particular solutions of the PDE's that can be solved through the IST. We will (a) discuss earlier work and will give a natural generalization, which captures a significantly larger class of solutions; (b) stress the connections between this method and some of the main features of the IST; and (c) compare this linearization with the RH direct approach introduced by ZS, showing their connections and differences.

II. THE DIRECT LINEARIZATION

The DL was introduced by Fokas and Ablowitz[5] in connection with the Korteweg–de Vries (KdV) equation

$$u_t + u_{xxx} + 6uu_x = 0, \quad u = u(x,t). \tag{1}$$

It is based essentially on the existence of an integral equation

$$\phi(x,t;k) + i\int_L \frac{\phi(x,t;l)}{l+k} e^{ilx + il^3 t}\, d\lambda(l) = 1, \tag{2}$$

involving an arbitrary contour L and measure $d\lambda(l)$ which linearizes Eq. (1). In fact, under the assumption that the homogeneous version of (2) has only the trivial solution, the solution ϕ of (2) provides a solution $u(x,t)$ of the KdV equation through the formula

$$u(x,t) = -\partial_x \int_L \phi(x,t;l) e^{ilx + il^3 t}\, d\lambda(l). \tag{3}$$

The original motivation for this result is associated with the, by now classical, IST (corresponding to $u \to 0$ sufficiently rapidly as $|x| \to \infty$) of the KdV equation. Specifically the integral equation (2), with contour and measure fixed and given in terms of the scattering data, is the integral formulation of the matrix RH problem,[2]

$$\begin{pmatrix} \tilde{\phi}(x,t;k) \\ \phi(x,t;k) \end{pmatrix} = G(x,t;k) \begin{pmatrix} \tilde{\phi}(x,t;-k) \\ \phi(x,t;-k) \end{pmatrix}. \tag{4}$$

In (4) ϕ is the same as in (2), $\tilde{\phi}$ is another eigenfunction with appropriate analytic properties, and the matrix G is given in terms of the scattering data.

Another motivation is based on the Rosales perturbation approach[6]; in fact the solution (3) can be interpreted as the sum of the perturbation series solution of the KdV equation around the solution $u = 0$.

The arbitrariness of contour and measure in (2) allows one to capture a wider class of solutions than the one given by the GLM equation; as an example in Ref. 5 it was shown for instance that using (2) it is possible to find a three parameter family of solutions of the self-similar reduction of (1):

$$u''' + 6uu' - (2u + xu') = 0, \quad u = u(x). \tag{5}$$

The GLM equation is able to provide just one parameter family of solutions of (5).

Another suggestive argument is associated with the linear limit of (3); in this case, Eq. (3) becomes

$$u(x,t) = -\partial_x \int_L e^{i(kx + k^3 t)}\, d\lambda(k). \tag{6}$$

Equation (6) is *the general solution* ("Ehrenpreis principle") of the linearized KdV equation

$$u_t + u_{xxx} = 0. \tag{7}$$

The linear limit of (3) provides the most general solution of

[a] Permanent address: Dipartimento di Fisica, Universita di Roma, 00183 Roma, Italy.

 0022-2488/84/092614-06$02.50

Eq. (7), whereas it is known that the linear limit of the GLM provides just those solutions of (7) obtainable using the Fourier transform.

These considerations are very far from implying that this DL provides the most general solution of (1); on the contrary recent studies on the equation of Painlevé II (PII),

$$v'' - xv - 2v^3 = \alpha, \tag{8}$$

which is intimately connected to Eq. (5) (see Ref. 7), have shown[8] that it is not the case, since the solutions of (5) obtained through (2) correspond just to the limited interval (0,1) of the parameter α in (8).

In other words, the perturbation solution (3) (in the Rosales language) of the KdV equation around $u = 0$ corresponds only to the solution of PII in the interval $0 < \alpha < 1$. It is natural to consider an extension of the DL formulated above which would correspond to the perturbation solution of the KdV equation around any arbitrary solution u_0 of the KdV itself.

III. A GENERALIZATION OF THE DL

The essence of this more general linearization is given by the following proposition.

Proposition 1: Let $\psi(x,t;k)$ be a solution of

$$\psi(x,t;k) + \int_L \psi(x,t;l)h(x,t;k,l)d\lambda(l) = \psi_0^{(1)}(x,t;k), \tag{9}$$

where L and $d\lambda(l)$ are arbitrary contour and measure; $\psi_0^{(1)}(x,t;k)$ and $\psi_0^{(2)}(x,t;k)$ are two arbitrary solutions of the coupled systems

$$\psi_{0_{xx}} + (u_0 + k^2/4)\psi_0 = 0, \tag{10a}$$

$$\psi_{0_t} = u_{0_x}\psi_0 + (k^2 - 2u_0)\psi_{0_x}; \tag{10b}$$

$u_0 = u_0(x,t)$ is any given solution of the KdV equation (1); and $h(x,t;k,l)$ is defined in terms of $\psi_0^{(1,2)}$ by

$$h(x,t;k,l) \doteq [2/(l^2 - k^2)]\left[\psi_0^{(2)}(x,t;l)\psi_{0_x}^{(1)}(x,t;k) - \psi_{0_x}^{(2)}(x,t;l)\psi_0^{(1)}(x,t;k)\right]. \tag{11}$$

Assuming that the homogeneous version of (9) has only the zero solution, then

$$u(x,t) = u_0(x,t) + \partial_x \int_L \psi(x,t;k)\psi_0^{(2)}(x,t;k)d\lambda(k) \tag{12}$$

solves the KdV equation.

The proof is direct and it is obtained operating on Eq. (9) with the operators P and M defined by

$$P \doteq \partial_{xx} + u + k^2/4, \quad M \doteq -\partial_t + u_x + (k^2 - 2u)\partial_x. \tag{13}$$

The result of this operation gives

$$(P\psi)(x,t;k) + \int_L (P\psi)(x,t;l)h(x,t;k,l)d\lambda(l) = 0, \tag{14a}$$

$$(M\psi)(x,t;k) + \int_L (M\psi)(x,t;l)h(x,t;k,l)d\lambda(l) = \int_L (P\psi)(x,t;l)\psi_0^{(1)}(x,t;k)\psi_0^{(2)}(x,t;l)d\lambda(l), \tag{14b}$$

and now if we assume that the homogeneous version of Eq. (9) has only the zero solution, Eq. (14a) implies that $P\psi = 0$ and then Eq. (14b) implies that $M\psi = 0$. The compatibility between these two equations finally implies that u solves the KdV equation (1).

The linearization given here obviously contains the special cases in which $u_0 = 0$ and $u_0 = -2/x^2$, which are explicit solutions of Eq. (1); in these cases the DL was previously given .[5,9]

In the Appendix we give a constructive procedure that, starting with the general assumption (9), enables one to characterize the kernel h in terms of ψ_0 as in (11) and, at the same time, to fix the integral representation of $u - u_0$ in terms of ψ and ψ_0 as in (12). Such a systematic procedure, whose main steps are essentially the same for all the PDE's solvable via the IST, will be the basis of the results of this paper.

We remark that we could have given the DL of the KdV equation for the function $\phi(x,t;k) \doteq \psi(x,t;k)/\psi_0^{(1)}(x,t;k)$, instead of $\psi(x,t;k)$. In this case the corresponding integral equation

$$\phi(x,t;k) + \int_L \phi(x,t;l)g(x,t;k,l)d\lambda(l) = 1, \tag{15a}$$

$$g(x,t;k,l) \doteq \psi_0^{(1)}(x,t;l)h(x,t;k,l)/\psi_0^{(1)}(x,t;k) \tag{15b}$$

has 1 as forcing term and apparently would be the more appropriate formulation for investigating analyticity properties in k, in view of the solution of the IST. As far as the DL is concerned, the two formulations are completely equivalent and here and in the following we will use either one or the other, according to the convenience and to the elegance of the associated formulas.

The explicit formula (11) for the kernel h of Eq. (9) shows the singular character of the integral equation and strongly suggests that, as in the case $u_0 = 0$, some type of RH problem is going to be the natural structure underlying the IST of the KdV equation for solutions u, as a finite perturbation of a given solution u_0.

As we wrote above, the essence of this method is related to the existence of a linear integral equation like (9) [or (15a)] which provides solutions of the KdV equation. On the other hand, we know that the KdV equation is one of the many PDE's that can be solved through the IST. Hence it is natural to ask ourselves if and how the DL, in the generalized form introduced here, applies to other equations.

For this purpose, let us consider the $n \times n$ matrix equation

$$\Psi_x = zJ\Psi + Q\Psi, \quad \Psi = \Psi(x,t;k), \tag{16}$$

where the scalar constant z plays the role of spectral parameter, J is a constant diagonal matrix, and $Q = Q(x,t)$ is an off-diagonal matrix. Equation (16) is the natural $n \times n$ generalization of the generalized ZS problem (see Refs. 10 and 11) and its IST has been recently rigorously studied by Beals and Coifman.[12]

In order to give the DL associated with (16) it is convenient to introudce the matrix function $\Phi(x,t;k)$ defined by

$$\Phi(x,t;k) \doteq \Psi(x,t;k)\Psi_0^{(1)-1}(x,t;k), \tag{17}$$

where Ψ and $\Psi_0^{(1)}$ solve Eq. (16) corresponding to the two potentials Q and Q_0.

The linearization of the class of evolution equations associated with the spectral problem (16) is then given by the following.

Proposition 2: Let $\Phi(x,t;z)$ be a solution of

$$\Phi(x,t;z) + \int_L \Phi(x,t;l)G(x,t;z,l)d\lambda(l) = I, \tag{18}$$

when L and $d\lambda(l)$ are arbitrary contour and measure, I is the identity matrix, G is defined by

$$G(x,t;z,l) \doteq (z-l)^{-1}\Psi_0^{(1)}(x,t;l) \times G_0(l)\Psi_0^{(2)-1}(x,t;l), \tag{19}$$

where $G_0(l)$ is an arbitrary constant matrix function, and the $\Psi_0^{(1,2)}$ are two arbitrary solutions of Eq. (16) for $Q_0(x,t)$. Assuming that the homogeneous version of (18) has only the trivial solution, then $\Psi(x,t)$ defined by

$$\Psi(x,t;z) = \Phi(x,t;z)\Psi_0^{(1)}(x,t;z) \tag{20}$$

solves Eq. (16) if

$$Q(x,t) = Q_0(x,t) + \left[J, \int_L \Phi(x,t;l)\Psi_0^{(1)}(x,t;l) \times G_0(l)\Psi_0^{(2)-1}(x,t;l)d\lambda(l)\right]. \tag{21}$$

In this proposition and in the following ones we often introduce arbitrary functions assuming that they satisfy suitable regularity properties in order to give sense to the integral formulas in which they appear.

Again the proof is direct and is obtained by applying the operator Ω,

$$(\Omega F)(x,t;z) \doteq -F_x + z[J,F] + QF - FQ_0, \tag{22}$$

on Eq. (18) to get

$$(\Omega\Phi)(x,t;z) + \int_L (\Omega\Phi)(x,t;l)G(x,t;z,l)d\lambda(l) = 0; \tag{23}$$

the result follows from the uniqueness assumption. In the Appendix we give a sketch of how to constructively obtain the linearization contained in Proposition 2, since the procedure does not differ in spirit from the one used for obtaining Proposition 1.

Problem (16) allows us to easily discuss the connections between the DL and the RH direct approach, indeed it will turn out that, if used on Eq. (16), then the two approaches are equivalent.

The RH direct approach introduced by ZS is based on the solution of the following matrix homogeneous RH problem:

$$\Phi^+(x,t;z) = \Phi^-(x,t;z)[I + R(x,t;z)], \tag{24}$$

where z lies on an arbitrary contour L in the complex-z plane, $\Phi^+(z)$ and $\Phi^-(z)$ are the boundary values on L of functions analytic inside and outside, respectively, of L, $\Phi^-(z) \to I$ as $|z| \to \infty$, and R is defined by

$$R(x,t;z) \doteq \Psi_0(x,t;z)G_0(z)\Psi_0^{-1}(x,t;z), \tag{25}$$

where $G_0(z)$ is an arbitrary constant matrix and Ψ_0 solves Eq. (16) with the potential Q_0. Then it is easy to verify [using (24), (25), and (16)] that $\Psi^{\pm}(x,t,z)$ and $Q(x,t)$, defined by

$$\Psi^{\pm}(x,t;z) = \Phi^{\pm}(x,t;z)\Psi_0(x,t,z), \tag{26}$$

$$zJ + Q(x,t) = [\Phi_x^{\pm} + \Phi^{\pm}(zJ + Q_0(x,t))](\Phi^{\pm})^{-1}, \tag{27}$$

solve Eq. (16).

The equivalence between the DL given in Proposition 2 and the RH direct method is immediate and obtains by comparing (18) and (19) with the $(-)$ projection of (24):

$$\Phi^-(x,t;z) - \frac{1}{2\pi i}\int_L \Phi^-(x,t;l)\frac{R(x,t;l)}{l-z}dl = I, \tag{28}$$

where $z \to L$ from outside the contour.

The equivalence of the two approaches shows that the homogeneous RH problem (24) on which the ZS method is based, is the natural analytic structure underlying the linearization of the PDE's associated with the spectral problem (16). The particular z-dependence of the kernel G of Eq. (18), given in (19), indicates that the integral equation (18) comes from the $(-)$ projection of a homogeneous RH problem of the type (24). Vice versa, if the z-dependence of G appeared in a different way, we would infer that (24) is not an adequate analytic structure for describing the problem. We will show in the following that this phenomenon is not exceptional, being a common feature of the PDE's related to the IST in $2+1$ dimensions.

While the RH approach (due to its restrictive basis) cannot in general be applied, the DL, based on a linear integral equation of the type (18), where the z-dependence of the kernel G is determined *a posteriori*, case by case [through direct algebraic calculations and is in general different from the one given in (19)], turns out to be a viable approach for characterizing a wide class of solutions of the PDE under investigation.

IV. THE DL IN 2 + 1 DIMENSIONS

The DL in $2+1$ dimensions is again based on a linear integral equation

$$\Phi(x,y,t;k) + \int_L \Phi(x,y,t;\beta(l,\nu)) \times G(x,y,t;k,l,\nu)d\zeta(l,\nu) = I. \tag{29}$$

Now the integration is in two variables l and ν, a reflection of the higher dimensionality of the configurational space, L and $d\zeta(l,\nu)$ are arbitrary contour and measure, $\beta = \beta(l,\nu)$ is an arbitrary function of l and ν, and the kernel G will be characterized in terms of certain linear PDE's whose coefficients are given in terms of the unperturbed solution $u_0(x,y,t)$.

As an example, let us apply the DL to the Kadomtsev–Petviashvili (KP) equation[13]

$$(u_t + 6uu_x + u_{xxx})_x = -3\sigma^2 u_{yy}, \quad \sigma \in \mathbb{C} \tag{30}$$

that can be obtained as a compatibility condition of the system

$$P\psi \doteq \sigma\psi_y + \psi_{xx} + u\psi = 0, \tag{31a}$$

$$M\psi \doteq \psi_t + 4\psi_{xxx} + 6u\psi_x + 3\left(u_x - \sigma\int_{-\infty}^{x} u_y(x')dx'\right)\psi = 0. \tag{31b}$$

In our analysis σ can be thought of as an arbitrary complex parameter, including then the two cases $\sigma = i$ and -1 (KPI and KPII) in which Eq. (30) describes the propagation of quasi-one-dimensional waves in a nonlinear weakly dispersive medium and the sign of σ^2 coincides with the sign of the dispersion.

We have the following proposition.

Proposition 3: Let $\psi(x,y,t;k)$ be a solution of

$$\psi(x,y,t;k)+\int_L\int \psi(x,y,t;\beta(l,\nu))\times h(x,y,t;k,l,\nu)d\zeta(l,\nu)=\psi_0(x,y,t;k), \tag{32}$$

where ψ_0 solves the coupled system (31) corresponding to a given solution $u_0(x,y,t)$ of Eq. (30) and h is given by the formula

$$h(x,y,t;k,l,\nu)=\frac{1}{2}\int_\alpha^x f(x',y,t;l,\nu)\psi_0(x',y,t;k)dx'+\omega(y,t;k,l,\nu,\alpha), \tag{33}$$

where ω is a solution of the coupled system

$$\sigma\omega_y=\tfrac{1}{2}[f_x(\alpha)\psi_0(\alpha)-f(\alpha)\psi_{0_x}(\alpha)], \tag{34a}$$

$$\omega_t=-2[f_{xx}(\alpha)\psi_0(\alpha)-f_x(\alpha)\psi_{0_x}(\alpha)+f(\alpha)\psi_{0_{xx}}(\alpha)]-3u_0f(\alpha)\psi_0(a), \tag{34b}$$

with

$$f(\alpha)\doteq f(\alpha,y,t;l,\nu),\quad \psi_0(\alpha)=\psi_0(\alpha,y,t;k),$$

and f solves

$$\sigma f_y-f_{xx}-u_0f=0, \tag{35}$$

$$f_t+4f_{xxx}+6u_0f_x+3\left(u_{0_x}+\sigma\int_{-\infty}^x u_{0_y}dx'\right)f=0. \tag{36}$$

Assuming that the homogeneous version of (32) has only the trivial solution, then $u(x,y,t)$, defined through

$$u(x,y,t)=u_0(x,y,t)+\partial_x\int_L\int\psi(x,y,t;\beta(l,\nu))\times f(x,y,t;l,\nu)d\zeta(l,\nu), \tag{37}$$

is a solution of the KP equation.

Again the proof is direct and it is based on the application of the operators P and M on Eq. (32). In the Appendix we show how the constructive procedure used to get Propositions 1 and 2 generalizes naturally to this (2 + 1)-dimensional example, hence enriching itself of new features and properties.

The solutions of the IST for KPI and KPII (see Refs. 14 and 15) can be easily recovered by choosing $u_0=0$ and $\beta=l$ for $\sigma=i$, and $u_0=0\,\beta=l-i\nu$ for $\sigma=-1$.

The formulas (33) and (34) or, equivalently, the system of linear PDE's (A10) satisfied by h, have a rather complicated k-dependence. However, when $u_0=0$, the situation simplifies enormously; in order to see that, let us introduce the functions g and v defined as

$$g(x,y,t;k,l,\nu)\doteq h(x,y,t;k,l,\nu)\times\psi_0(x,y,t;\beta(l,\nu))/\psi_0(x,y,t;k), \tag{38a}$$

$$v(x,y,t;l,\nu)\doteq\psi_0(x,y,t;\beta(l,\nu))f(x,y,t;l,\nu). \tag{38b}$$

Rewriting the system (A10) (including also the t-equation) for the function g, and considering the case $u_0=0$ (and $\psi_0(x,y,t;k)=\exp[ikx+(k^2/\sigma)y+4ik^3t]$), one obtains the overdetermined system

$$g_x+i(k-\beta)g=\tfrac{1}{2}v, \tag{39a}$$

$$\sigma g_y+(k^2-\beta^2)g=\tfrac{1}{2}[v_x-i(k+\beta)v], \tag{39b}$$

$$g_t+4i(k^3-\beta^3)g=2[(k^2+\beta^2+k\beta)v+i(k+2\beta)v_x-v_{xx}]. \tag{39c}$$

The compatibility condition for the system (39) fixes the k dependence of g in the form

$$g(x,y,t;k,l,\nu)=-iv(x,y,t;l,\nu)/[2(k-c(l,\nu))], \tag{40}$$

where $c=c(l,\nu)$ is an arbitrary function of l and ν and, correspondingly, v solves the equations

$$v_x=i(\beta-c)v, \tag{41a}$$

$$\sigma v_y=(\beta^2-c^2)v, \tag{41b}$$

$$v_t=4i(\beta^3-c^3)v. \tag{41c}$$

The k-dependence of g (and then of h) implies that the integral equation (32) can be derived from a RH problem of the type introduced by Manakov[16] in a work in which he has generalized and adapted the RH direct approach of Ref. 4 to 2 + 1 dimensions. He postulates a nonlocal RH problem,

$$\phi^+(x,y,t;k)=\phi^-(x,y,t;k)+\int\phi^-(x,y,t;l)G(x,y,t;k,l)dl, \tag{42}$$

in order to detect and generate PDE's solvable via the IST. The existence of explicit cases, associated with $u_0=0$ (and briefly discussed above), in which a RH structure is recovered, is a confirmation of the validity of Manakov's approach (for $u_0=0$) in finding a connection between the KP equation and the nonlocal RH problem (42). Such a connection was also proven via the solution of the IST (see Refs. 14, 15, and 17). In Ref. 15 in particular, for the first time it was shown that the KPII equation is related to a "$\bar\partial$" problem, whose integral representation also gives rise to the k-dependence presented in (40). But the nongenericity of the above-mentioned examples corresponding to the case $u_0=0$ indicates at the same time that the RH problem (42) is not adequate to capture a wide class of solutions of the KP equation.

We will show in the following that essentially the same situation arises when one writes the DL of a class of PDE's associated with the 2 + 1 dimensional generalization of the spectral problem (16). Such a generalization is[18]

$$\Psi_x=J\Psi_y+Q\Psi, \tag{43}$$

where $\Psi=\Psi(x,y,t;k)$, $Q=Q(x,y,t)$ and J are $n\times n$ matrices, and J is constant and diagonal while Q is off diagonal. Physically relevant equations such as the so-called Davey–Stewartson equation, the n-wave interaction in 2 + 1 dimensions, and the modified KP equation are related to (43). The IST associated with this linear problem has been recently investigated and solved in Refs. 19–21.

The DL corresponding to (43) is formulated in the following way.

Proposition 4: Let $\phi(x,y,k)$ be a solution of

$$\Phi(x,y,;k)+\int_L\int\Phi(x,y,;\beta(l,\nu))\times G(x,y;k,l,\nu)d\zeta(l,\nu)=I, \tag{44}$$

where L and $d\zeta(l,\nu)$ are arbitrary contour and measure,

$\beta = \beta(l,\nu)$ is an arbitrary function of l and ν, G is given by the expression

$$G(x,y;k,l,\nu) \doteq \Psi_0(x,y;\beta(l,\nu))\Big(\int_\alpha^x \Psi_0^{-1}(x',y;\beta(l,\nu)) \times R(x',y;l,\nu) J \Psi_0(x',y,k)\,dx' + g(y,k,l,\nu,\alpha)\Big)\Psi_0^{-1}(x,y;k), \tag{45}$$

where

$$g_y(y,k,l,\nu,\alpha) = \Psi_0^{-1}(\alpha,y;\beta(l,\nu))R(\alpha,y;l,\nu)\Psi_0(\alpha,y;k), \tag{46}$$

$R = R(x,y;l,\nu)$ solves

$$-R_x + R_y J + [J, S(x,y;\beta(l,\nu))R] + [Q_0(x,y), R] = 0, \tag{47}$$

with

$$S(x,y;\beta) \doteq \Psi_{0_y}(x,y;\beta)\Psi_0^{-1}(x,y;\beta), \tag{48}$$

and Ψ_0 is a solution of (43) corresponding to the potential $Q_0(x,y)$. Then

$$\Psi(x,y;k) = \Phi(x,y;k)\Psi_0(x,y;k) \tag{49}$$

solves Eq. (43) if $Q(x,y)$ is given by

$$Q(x,y) = Q_0(x,y) + \Big[J, \int_L \int \Phi(x,y;\beta(l,\nu)) \times R(x,y;l,\nu) d\zeta(l,\nu)\Big]. \tag{50}$$

Again we refer to the Appendix for the derivation of this proposition. Formulas (45) and (46) imply that the kernel G satisfies the following set of (compatible) linear PDE's:

$$G_y + GS(k) - S(\beta)G = R, \tag{51a}$$

$$G_x + G(JS(k) + Q_0) - (JS(\beta) + Q_0)G = RJ. \tag{51b}$$

When $Q_0 = 0$ (and $\psi_0 = \exp[ik(Jx + y)]$), the compatibility condition for the system (51) fixes the k-dependence of G in the form

$$G(x,y;k,l,\nu) = -iR(x,y;l,\nu)/[k - c(l,\nu)], \tag{52}$$

and, correspondly, R is defined through the equations

$$R_x = i(\beta JR - cRJ), \tag{53a}$$

$$R_y = i(\beta - c)R \tag{53b}$$

postulated by Manakov in its nonlocal RH approach. This shows again how the nonlocal RH problem (42) is an appropriate tool to detect the PDE's in 2 + 1 dimensions corresponding to the linear problem (43), but, unlike the case of its associated 1 + 1 analog, it is able to capture a restricted class of solutions only (e.g., the ones obtained perturbing off of the zero solution).

Concluding this paper, we would like to remark that the DL has been studied here in connection with a certain selection of relevant eigenvalue problems associated with the IST theory, showing that the general assumptions on which the DL is based are consistent with the general features of the IST theory.

APPENDIX

In this Appendix we will illustrate the constructive procedure used in this paper in order to obtain the DL contained in Propositions 1–4. Since the main steps of this procedure are essentially the same for all of the PDE's related to the IST, we will discuss them in some detail for the KdV example, limiting our discussion of the other cases to those situations in which the procedure introduced needs to be modified or exhibits new features.

(1) The first step consists in writing the integral equation for ψ,

$$\psi(x;k) + \int_L \psi(x;l)h(x;k,l)d\lambda(l) = \psi_0^{(1)}(x;k), \tag{A1a}$$

or for $\phi = \psi\psi_0^{(1)-1}$,

$$\phi(x;k) + \int_L \phi(x;l)g(x;k,l)d\lambda(l) = 1. \tag{A1b}$$

The kernel h has to be characterized *a posteriori* as is indicated in the following steps.

(2) In the second step one applies the spectral operator P to (A1) [or Ω to (A2)]. In the KdV case $P = \partial_{xx} + u(x) + k^2/4$ and Eq. (A1a) implies

$$\psi_{xx}(x;k) + \int_L [\psi_{xx}(x;l)h(x;k,l) + 2\psi_x(x;l)h_x(x;k,l) + \psi(x;l)h_{xx}(x;k,l)]d\lambda(l) = \psi_{0_{xx}}(x;k), \tag{A2a}$$

$$u(x)\psi(x;k) + \int_L u(x)\psi(x;l)d\lambda(l) = u_0(x)\psi_0(x;k) + [u(x) - u_0(x)]\psi_0(x;k), \tag{A2b}$$

$$\frac{k^2}{4}\psi(x;k) + \int_L \Big(\frac{l^2}{4}\psi(x;l)h(x;k,l) + \frac{k^2 - l^2}{4}\psi(x;l)h(x;k,l)\Big)d\lambda(l) = \frac{k^2}{4}\psi_0(x;k). \tag{A2c}$$

Adding these three equations up, one obtains

$$(P\psi)(x;k) + \int_L (P\psi)(x;l)h(x;k,l)d\lambda(l) + \int_L \Big[2\psi_x(x;l)h_x(x;k,l) + \psi(x;l) \times \Big(h_{xx}(x;k,l) + \frac{k^2 - l^2}{4}h(x;k,l)\Big)\Big]d\lambda(l) = (P_0\psi_0^{(1)})(x;k) + (u - u_0)\psi_0^{(1)}(x;k). \tag{A3}$$

Then the requirement $P\psi = 0$ ($P_0\psi_0$ is already zero by hypothesis) isolates an equation for $u - u_0$ which, in the KdV case, reads

$$(u - u_0)\psi_0^{(1)}(k) = 2\int_L \psi_x(l)f(k,l)d\lambda + \int_L \psi(l)\Big(h_{xx} + \frac{k^2 - l^2}{4}h\Big)d\lambda \tag{A4}$$

here and below we omit for convenience the x dependence.

(3) The analysis of Eq. (A4) suggests the structure of the integral representation of $u - u_0$, in our case Eq. (A4) implies that $u - u_0$ *must* have the form

$$u - u_0 = \int_L [\psi_x(l)f_1(l) + \psi(l)f_2(l)]\,d\lambda, \tag{A5}$$

where the functions f_1 and f_2 are characterized in the next step.

(4) Evaluate the consequences of the assumed structure (A5). For instance in this specific case, Eqs. (A4) and (A5) imply

$$2h_x = f_1(l)\psi_0^{(1)}(k), \tag{A6a}$$

$$h_{xx} + [(k^2 - l^2)/4]h = f_2(l)\psi_0^{(1)}(k), \tag{A6b}$$

and one can verify that the compatibility condition for this system implies that $f_2 = f_{1_x} = \psi_{0_x}^{(2)}$, where $\psi_0^{(2)}$ is an arbitrary solution of the Schrödinger equation (10), and also that h is given by formula (11).

When applied to other examples, the procedure above repeats exactly for the first two steps, while steps 3 and 4 adjust to the specific problem under investigation. If, for instance, we deal with Eq. (16), after steps 1 and 2 we have

$$Q - Q_0 = \int_L \{(z - l)J\Phi G + \Phi[-G_x + (lJ + Q_0)G - G(zJ + Q_0)]\}d\lambda(l), \tag{A7}$$

and now taking into account that $Q - Q_0$ is a k-independent off-diagonal matrix, on analogy of (A5) we necessarily find the structure

$$Q - Q_0 = \left[J, \int_L \phi(l)R(l)\right]d\lambda, \tag{A8}$$

where again R has to be characterized. Substituting (A8) into (A7) we then obtain

$$(z - l)G(z,l) = R(l), \tag{A9a}$$

$$G_x + G(zJ + Q_0) - (lJ + Q_0)G = R(l)J. \tag{A9b}$$

System (A9) has the solution

$$G(k,l) = R(l)/(z - l), \quad R(l) = \psi_0^{(1)}(l)G_0(l)\psi_0^{(2)-1}(l).$$

The application of our procedure to equations in 2 + 1 dimensions leaves essentially unchanged the first three steps, and leads to the integral representations (A5) for KP and (A8) for the spectral problem (43). Equation (A5) implies for KP the following system:

$$2h_x = f_1(l)\psi_0(k), \tag{A10a}$$

$$\sigma h_y + h_{xx} = f_2(l)\psi_0(k), \tag{A10b}$$

whose compatibility condition implies that $f_{1_x} = f_2 = f_x$, and formula (34). Equation (A8) for the linear problem (43) implies the system (51), whose compatibility condition is given by Eq. (47).

[1] M. J. Ablowitz and H. Segur, "Solitons and the Inverse Scattering Transform," in *SIAM Studies in Applied Mathematics*, No. 4 (SIAM, Philadelphia, 1981).
[2] A. S. Fokas and M. J. Ablowitz, "Lectures on the Inverse Scattering Transform for Multidimensional (2 + 1) Problems," INS No. 28, December 1982, *Lecture Notes in Physics*, edited by K. B. Wolf, No. 189 (Springer, Berlin, 1983).
[3] V. E. Zakharov and P. B. Shabat, Funct. Anal. Appl. **8**, 226 (1974).
[4] V. E. Zakharov and P. B. Shabat, Funct. Anal. Appl. **13**, 166 (1979).
[5] A. S. Fokas and M. J. Ablowitz, Phys. Rev. Lett. **47**, 1096 (1981).
[6] R. Rosales, Stud. Appl. Math. **59**, 117 (1978).
[7] A. S. Fokas and M. J. Ablowitz, J. Math. Phys. **23**, 2033 (1982).
[8] A. S. Fokas and M. J. Ablowitz, "On the Initial Value Problem of the Second Painlevé Transcendent," Commun. Math. Phys. **91**, 381 (1983).
[9] A. S. Fokas and M. J. Ablowitz, in *AIP Conference Proceedings, No. 88*, edited by M. Tabor and Y. M. Treve (AIP, New York, 1982), pp. 237.
[10] V. E. Zakharov and P. B. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[11] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974); M. J. Ablowitz and R. Haberman, J. Math. Phys. **16**, 2301 (1975).
[12] R. Beals and R. R. Coifman, "Scattering and Inverse Scattering for First Order Systems," Commun. Pure Appl. Math. (to be published).
[13] B. B. Kadomtsev and V. I. Petviashvili, Sov. Phys. Dokl. **15**, 539 (1970).
[14] A. S. Fokas and M. J. Ablowitz, Phys. Lett. A **94**, 67 (1983).
[15] M. J. Ablowitz, D. Bar Yaacov, and A. S. Fokas, "On the IST for KPII," INS No. 21, Nov. 1982, Stud. Appl. Math. **69**, 135 (1983).
[16] S. V. Manakov (private communication); and lecture by P. Kulish at the "Solitons '82" Conference and Workshop, Edinburgh, Scotland, August 1982.
[17] S. V. Manakov, Physica D **3**, 420 (1981).
[18] M. J. Ablowitz, R. Haberman, Phys. Rev. Lett. **35**, 1185 (1975).
[19] A. S. Fokas, Phys. Rev. Lett. **51**, 3 (1983).
[20] A. S. Fokas and M. J. Ablowitz, Phys. Rev. Lett. **51**, 7 (1983).
[21] A. S. Fokas and M. J. Ablowitz, J. Math. Phys. **25**, 2494 (1984).

On the old–new method of solving nonlinear equations[a)]

M. W. Kalinowski[b)]
Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada

(Received 17 November 1981; accepted for publication 27 January 1984)

A method for solving a quasilinear nonelliptical equation of the second order is developed. We give classification and parametrization of simple elements of the equation. An equation of potential stationary flow of compressible gas in a supersonic region is considered as an example. A new exact solution is obtained which may be treated as a nonlinear analog of stationary wave. A gauge structure for the equation and an analog of Bäcklund transformation are introduced.

PACS numbers: 02.30.Ks, 02.70. + d, 02.40. + m, 47.40.Ki

I. INTRODUCTION

In this paper we present a method of solving nonlinear partial and nonelliptical differential equations. The paper is devoted to theory of a differential equation of second order with coefficients depending on first derivatives. The method is old, for we employ Riemann invariants which have been known, really, for a long time.[1-3] Nevertheless, we have found a new feature of this old theory related to the problem of algebraization of a differential equation. We exactly solve a characteristic cone for a Riemann wave and classify all simple elements for such an equation. All these simple elements depend on parameters. We proved that these parameters are parameters of some orthogonal or pseudo-orthogonal groups. [It is very important in later analysis (see Sec. VI).] Having a classification of simple elements we may classify simple waves. Every simple element belongs to some submanifold $F_i \subset \mathscr{E}^*$. The dimension of this submanifold equals the number of parameters. These parameters are arbitrary functions of dependent variables φ_i. Since we are dealing with Riemann invariants, parameters become functions of R (Riemann invariant). Due to these arbitrary functions we may integrate the equations for simple waves. It indicates that it is possible to introduce a gauge structure for every class of simple elements.

This paper is divided into six sections. In Sec. II, we describe a parametrization and a classification of simple elements and we write down all possible elements. In Sec. III we deal with a physical example from hydrodynamics. It is the equation of potential stationary flow for a perfect gas in a supersonic region. In Sec. IV we calculate all possible simple elements for this equation and we write them down in the Appendices. In Sec. V, we calculate the simple wave corresponding to one of the simple elements. We find a new exact solution of the equation and examine its properties. It depends on three arbitrary functions. We shift the arbitrariness from the parameters (depending on R—Riemann invariant) to more convenient functions. It results in some restrictions on the arbitrary functions and on the range of the Riemann invariant R. New freedom connected to the parameters of the characteristic cone, and restrictions imposed on them are new points in the Riemann-invariants method. Simultaneously we have obtained a restriction on the range of the dependent variable R. This seems to be a new point also. We have got the solution for the equation of potential flow of the perfect gas (in three dimensions) which depends on three arbitrary functions. Restrictions which we had to impose on the functions lead to physical effects. The solution may be considered as an analog of a nonlinear stationary wave. We have obtained planes of "density and pressure nodes" and planes of "magnitude of velocity antinodes." In Sec. VI, we deal with a "gauge" structure for the equation and its simple waves. We get a transformation of gauge type connecting two simple waves (from the same class). The transformation may be treated as a nonlinear representation of a gauge group originating from the orthogonal or pseudo-orthogonal group. The transformation is similar to the classical Bäcklund transformation.

II. THE ALGEBRAIZATION PROCEDURE

Let us consider a system of partial differential equations

$$a_j^{s\nu}(u^1,u^2,\ldots,u^l)\frac{\partial}{\partial x^\nu}u^j(x^1,x^2,\ldots,x^n)=0 \quad \begin{matrix}\nu=1,2,\ldots,n.\\ s=1,2,\ldots,m.\\ j=1,2,\ldots,l.\end{matrix} \quad m\geqslant l \tag{2.1}$$

$$x=(x^1,x^2,\ldots,x^n)\in\mathscr{E}\ ,\quad u(x)=(u^1(x),u^2(x),\ldots,u^l(x))\in\mathscr{H}\ ,$$

which is a quasilinear homogeneous system of the first order with coefficients depending only on the unknown functions. This system may be overdetermined, i.e., $m\geqslant l$. Let us suppose that this system is a nonelliptical one. This indicates that there exist some nontrivial solutions of the algebraic system of equations

$$a_j^{s\nu}\gamma^j\lambda_\nu=0, \quad \text{where } \operatorname{rank}\|a_j^{s\nu}\lambda_\nu\|<l \tag{2.2}$$

for vectors $\gamma\in R^{\,l}$ and $\lambda\in R^{\,n}$.

The above algebraic system of equations adequately specifies so-called knotted characteristic vectors in hodograph space, $\mathscr{H}=R^{\,l}$ (i.e., the space of the values of the functions u^j) and in physical space $\mathscr{E}=R^{\,n}$ (i.e., the space of independent variables). The pair γ and λ will be called a knotted pair iff it obeys Eq. (2.2). This fact will be marked by $\gamma\sim\lambda$. Matrix $L^{\,j}_{\,\nu}=\gamma^j\lambda_\nu$ created by a pair of knotted vectors will be a simple integral element, because rank $\|L^{\,j}_{\,\nu}(u_0)\|=1$, where $u_0\in\mathscr{H}$.

It is convenient to consider λ as an element of space $\mathscr{E}^*$ which is the space of linear forms, $\mathscr{E}^*\ni\lambda:\mathscr{E}\to R^{\,1}$. On the other hand, in the terminology of tensor calculus if we consider $x\in\mathscr{E}$ as a contravariant vector, then $\lambda\in\mathscr{E}^*$ is a covariant vector. In these terms the element L is an element of tensor space $T_u\mathscr{H}\otimes\mathscr{E}^*$ of the form $L=\gamma\otimes\lambda$. Now we introduce a simple wave, which will suggest a separation of simple integral elements from a set of integral elements.

a) Partially supported by NSF Grant No. INT 73-20002 A01, formerly GF-41958.

b) On leave of absence from the Institute of Philosophy and Sociology, Polish Academy of Science, 00-330 Warsaw, Nowy Swiat 72, Poland.

 0022-2488/84/092620-13$02.50

Let the mapping $u:D\to\mathscr{H}$, $D\subset\mathscr{E}$ be any solution of the system (2.1). We call u a simple wave for a homogeneous system if the tangent mapping du is a simple element at any point $x_0\in D$. Let us consider the smooth curve: $\Gamma:R\to f(R)$ in hodograph space $\mathscr{H}$ parametrized by R, so the tangent vector

$$\frac{df(R)}{dR}=\gamma(f(R)) \tag{2.3}$$

is a characteristic vector. Then there exists a field of characteristic covectors $\lambda(u)$ connected with $\gamma(f(R))$ specified on the curve $\Gamma:\lambda=\lambda(f(R))$. Hence we may state the following. If the curve $\Gamma\subset\mathscr{H}$ obeys the condition (2.3) and if $\varphi(\cdot)$ is any differentiable function of one variable, then the function $u=u(x)$ specified in an implicit way by

$$u=f(R),\qquad R=\varphi(\lambda_\nu(f(R))x^\nu),\ \text{where } a_j^{s\nu}\gamma^j\lambda_\nu=0 \tag{2.4}$$

is the solution of basic system (2.1). The solution is called a simple wave or Riemann wave.

A proof may be obtained by direct differentiation of the implicit relations (2.4). The vector λ in (2.4) specifies the velocity and direction of the wave propagation. The curve Γ fulfilling the condition (2.3) is called a characteristic curve in hodograph space $\mathscr{H}$. It means that if the mapping $u:\mathscr{E}\to\mathscr{H}$ is a simple wave, then a characteristic curve in space $\mathscr{H}$ is an image of mapping u. Parameter R, specified on this curve, is called a Riemann invariant.

Now let us consider a nonelliptical equation of the second order,

$$\sum_{i,j=1}^{n}a_{ij}\left(\frac{\partial\phi}{\partial x^1},\frac{\partial\phi}{\partial x^2},\ldots,\frac{\partial\phi}{\partial x^n}\right)\frac{\partial^2\phi}{\partial x^i\partial x^j}=0. \tag{2.5}$$

Equation (2.5) may be transformed, by introduction of new dependent variables, to the system of equations of first order,

$$\sum_{i,j=1}^{n}a_{ij}(u^1,\ldots,u^n)\frac{\partial u^i}{\partial x^j}=0, \tag{2.6}$$

$$\frac{\partial u^i}{\partial x^j}-\frac{\partial u^j}{\partial x^i}=0,$$

where $u^i=(\partial\phi/\partial x^i)$, $i=1,2,\ldots,n$.

Rewriting the system of equations (2.6) by means of simple integral elements we have

$$\sum_{i,j=1}^{n}a_{ij}\gamma^i\lambda_j=0, \tag{2.7a}$$

$$\gamma^i\lambda_j-\gamma^j\lambda_i=0,\quad i,j=1,2,\ldots,n. \tag{2.7b}$$

From Eq. (2.7b) we obtain that vector λ is proportional to vector γ. Thus from Eq. (2.7a) we get a quadratic form with respect to variables $\lambda_1,\lambda_2,\ldots,\lambda_n$,

$$Q(\lambda_1,\lambda_2,\ldots,\lambda_n)=\sum_{i,j=1}^{n}a_{ij}\lambda_i\lambda_j=0. \tag{2.8}$$

Equation (2.8) is the equation of a cone of the characteristic vector λ, specifying the velocity and direction of propagation of the simple wave.

Our aim will be to find the parametrical equations of this cone and at the same time to parametrize covector λ.

To do this we transform the quadratic form Q to a canonical form, i.e., we diagonalize matrix of the form Q, $A=(a_{ij})$. Hence we look for eigenvalues of this matrix and write the secular equation

$$\det(A-\omega J)=0. \tag{2.9}$$

We also search for a matrix B which diagonalizes the matrix A,

$$B^TAB=\begin{pmatrix}\omega_1 & & 0\\ & \omega_2 & \\ & & \ddots\\ 0 & & \omega_n\end{pmatrix},\qquad B^T=B^{-1}, \tag{2.10}$$

where $\omega_1,\omega_2,\ldots,\omega_n$ are the eigenvalues of matrix A. The matrix A is a symmetrical real matrix and it has real eigenvalues. There always exists at least one such orthogonal matrix B responsible for a rotation in n-dimensional Euclidian space from variables $v=(v^1,\ldots,v^n)$ to variables $\lambda=(\lambda^1,\ldots,\lambda^n)$,

$$\lambda=B\begin{pmatrix}v^1\\ v^2\\ \vdots\\ v_n\end{pmatrix}=Bv\quad\text{and} \tag{2.11}$$

$$Q(\lambda_1,\ldots,\lambda_n)=\omega_1(v_1)^2+\omega_2(v_2)^2+\cdots+\omega_n(v_n)^2$$

$$=(B^TAB)_{ij}vv_j.$$

We assume that $\omega_i\neq0$, $i=1,2,\ldots,n$. The case with $\omega_i=0$ is examined separately.

Now we aim to parametrize matrix B, i.e., we find all B's obeying (2.10).

Let a matrix A possess K different eigenvalues with orders l_j, $j=1,2,\ldots,K$, respectively, so that $\Sigma_{j=1}^{K}l_j=n$. Thus we obtain the following formula:

$$Q(\lambda_1,\lambda_2,\ldots,\lambda_n)=\sum_{i=1}^{K}\omega_i\left(\sum_{\mu=1}^{l_i}v_{p_i+\mu}^2\right),\quad\text{where } p_i=\sum_{\nu=1}^{i=1}l_\nu,\ \text{and}$$

$$B^TAB=\begin{pmatrix}\omega_1 & & & & & 0\\ & \ddots & & & & \\ & & \omega_1 & & & \\ & & & \ddots & & \\ & & & & \omega_K & \\ & & & & & \ddots\\ 0 & & & & & \omega_K\end{pmatrix}\quad(l_1,\ldots,l_K) \tag{2.12}$$

Let a matrix B_0 be any matrix diagonalizing the matrix A. Let us observe that any orthogonal transformation of variables $v_1,v_2,...,v_{l_1}$ does not destroy the diagonalization, i.e., any orthogonal $l_1\times l_1$ matrix C_1, acting on the first l_1 variables, does not change the diagonal form.

So a matrix B is defined modulo the following matrix:

$$B_1=\begin{pmatrix} \underbrace{C_1}_{l_1}\Big\}l_1 & & & & & 0 \\ & \underbrace{1}_{1} & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & 0 & & & & 1 \end{pmatrix}\Big\}n\,, \tag{2.13}$$

where $C_1^T=C_1^{-1}$. The same may be said about the remaining subsets of variables v_i, i.e., sequences of l_i elements.

The action of the $l_i\times l_i$ matrix $C_i^T=C_i^{-1}$ (the orthogonal one) cannot destroy the diagonalization. It is easy to see that matrix C_i corresponds to the following $n\times n$ matrix B_i,

$$B_i=\begin{pmatrix} 1 & & & & & & & 0 \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & \ddots & & & & \\ & q_1 & & & 1 & \underbrace{\quad}_{l_i} & & \\ & & & & l_i\{ & \boxed{C_i} & & \\ & & & & & & 1 & \\ 0 & & & & & & & 1 \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & q_2 & 1 \end{pmatrix}\Big\}n\,, \tag{2.14}$$

where $\underbrace{\qquad\qquad}_{n}$,

$$q_1=\sum_{j=1}^{i-1}l_i\,,\quad q_2=\sum_{j=i+1}^{k}l_i\,.$$

Hence matrix B is defined modulo the product of K matrices B_i of the form of (2.14). From that we obtain

$$B=B_0\prod_{i=1}^{K}B_i\,, \tag{2.15}$$

where B_0 is an arbitrary, but established, matrix obeying (2.12). The order of multiplication of matrices B_i is irrelevant, because all matrices commutate one with another. So, we see that the problem of parametrization of B is reduced to parametrization of each of the K orthogonal matrices C_i. Each of these matrices is responsible for an arbitrary rotation or reflection in l_i-dimensional space, thus it depends on $\frac{1}{2}l_i(l_i-1)$ parameters which are generalized Euler angles and some set of discrete parameters.

So we have $C_i=C_i(\alpha_{j_i}^i)$, where $j_i=1,2,...,l_i(l_i-1)/2$, where $i=1,2,...,K$.

By enumerating them in the order of their occurrence, i.e., first parameters of matrix C_1, then C_2, etc., we can write

$$B=B(\alpha_1,\alpha_2,...,\alpha_m,K_1,K_2,...,K_l)\,,\ \text{where}$$

$$m=\frac{1}{2}\sum_{i=1}^{K}l_i(l_i-1) \tag{2.16}$$

and $K_1,K_2,...,K_l$ correspond to reflection.

Finally, we have

$$\gamma\sim\lambda=B(\alpha_1,\alpha_2,...,\alpha_m,K_1,K_2,...,K_l)v\,. \tag{2.17}$$

After having matrix B parametrized we deal with form Q in coordinates v, and consequently with the equation of the characteristic cone in these coordinates. So we have

$$\sum_{i=1}^{K}\omega_i\left(\sum_{\mu=1}^{l_i}v_{p_i+\mu}^2\right)=0 \tag{2.18}$$

[compare (2.12)].

We find a parametrical equation of this cone. In order to do this we write (2.18) in the following form:

$$\sum_{j=1}^{n}z_iv_j^2=0,\quad \text{where } z_j \text{ is one of } \omega_i\,. \tag{2.18'}$$

Equation (2.18') is the equation of a $(n-l)$-dimensional quadric in projective coordinates in the canonical form.[4]

Let us suppose that $v_{j_0}\neq 0$ and divide both sides of (2.18') by $v_{j_0}^2$ and introduce new coordinates

$$\overset{(1)}{\sigma}_j=v_j/v_{j_0}\,,\quad j=1,2,...,n\,,\quad j\neq j_0\,.$$

One gets

$$\sum_{\substack{j=1\\ j\neq j_0}}^{n}z_j\overset{(1)}{\sigma}_j^2+z_{j_0}=0\,. \tag{2.19}$$

This equation can be written in a parametric form,

$$\overset{(1)}{\sigma}_j=\overset{(1)}{\sigma}_j(\overset{(1)}{\tau}_1,\overset{(1)}{\tau}_1,...,\overset{(1)}{\tau}_{n-2})\,,\quad j=1,2,...,n,\quad j\neq j_0\,. \tag{2.20}$$

$\overset{(1)}{\tau}_1,\overset{(1)}{\tau}_2,...,\overset{(1)}{\tau}_{n-2}$ are internal parameters of the quadric whereas functions $\overset{(1)}{\sigma}_j(\overset{(1)}{\tau}_1,\overset{(1)}{\tau}_2,...,\overset{(1)}{\tau}_{n-2})$ are expressed by either trigonometric or hyperbolic functions with respect to a type of quadric (2.19) according to F. Klein's classification.[4] Thus, we parametrize vector v,

$$v_j=\overset{(1)}{\sigma}_j\left(\overset{(1)}{\tau}_1,\overset{(1)}{\tau}_2,...,\overset{(1)}{\tau}_{n-2}\right)\overset{(1)}{\tau}_{n-1}\,,\quad j\neq j_0\,, \tag{2.21}$$

$$v_{j_0}=\overset{(1)}{\tau}_{n-1}\,,\quad \overset{(1)}{\tau}_{n-1}\neq 0\,,$$

where $\overset{(1)}{\tau}_{n-1}$ is a new parameter.

In a case when $v_{j_0}=0$ we have

$$\sum_{\substack{j=1\\ j\neq j_0}}^{n}z_iv_j^2=0\,, \tag{2.22}$$

so in this case the dimension of (2.22) is smaller than (2.18').

Similarly we choose $v_{j_1}\neq 0$, $j_1\neq j_0$, $j_1=1,2,...,n$, and introduce new variables

$$\overset{(2)}{\sigma}_j=v_j/v_{j_1}\,,\quad j\neq j_1,\ j\neq j_0\,. \tag{2.23}$$

We obtain Eq. (2.22) in the form

$$\sum_{\substack{j=1\\ j\neq j_0\\ j\neq j_1}}^{n}z_j\overset{(2)}{\sigma}_j^2+z_{j_1}=0\,. \tag{2.24}$$

Now we write its parametrical form,

$$\overset{(2)}{\sigma}_j=\overset{(2)}{\sigma}_j\left(\overset{(2)}{\tau}_1,\overset{(2)}{\tau}_2,...,\overset{(1)}{\tau}_{n-3}\right)$$

(one parameter less than above) and get

$$v_j = \overset{(2)}{\sigma_j}\Big(\overset{(2)}{\tau_1},\overset{(2)}{\tau_2},\ldots,\overset{(2)}{\tau_{n-3}}\Big)\overset{(2)}{\tau_{n-2}}, \quad j\neq j_1,\ j\neq j_0,$$
$$v_{j_1} = \overset{(2)}{\tau_{n-2}}, \quad \overset{(2)}{\tau_{n-2}}\neq 0, \tag{2.25}$$
$$v_{j_0} = 0,$$

and we proceed likewise, i.e., we consider the case $v_{j_0}=0=v_{j_1}$ and choose $v_{j_2}\neq 0$, etc. After K steps of this procedure we get the following equations for v:

$$v_j = \overset{(K)}{\sigma_j}\Big(\overset{(K)}{\tau_1},\overset{(K)}{\tau_2},\ldots,\overset{(K)}{\tau_{n-(K+1)}}\Big)\overset{(K)}{\tau_{n-K}}, \ \overset{(K)}{\tau_{n-K}}\neq 0,$$
$$\bigwedge_{\nu=0,1,\ldots,K-2} v_{j_\nu} = 0, \tag{2.26}$$
$$v_{j_{K-1}} = \overset{(K)}{\tau_{n-K}},$$
$$\bigwedge_{\nu=0,1,2,\ldots,K-1} j\neq j_\nu .$$

It is easy to see that we exhaust all possibilities if $n = K_{\max}+1$. Then we have

$$v_r = \overset{(n-1)}{\sigma}\cdot\overset{(n-1)}{\tau_1}, \quad \overset{(n-1)}{\tau_1}\neq 0,$$
$$v_{j_{n-1}} = \overset{(n-1)}{\tau_1}, \tag{2.27}$$
$$r\neq j_\nu$$
$$\bigwedge_{\nu=0,1,\ldots,n-2}$$

and the rest of v_i equal zero. The last case where $v=0$ is uninteresting.

At this stage we can present full parametrization of convector λ, which divides a set of λ's into $(n-1)$ disconnected types. Since we have

$$\lambda = B(\alpha_1,\alpha_2,\ldots,\alpha_m,K_1,K_2,\ldots,K_l)v,$$

where $\alpha_1,\alpha_2,\ldots,\alpha_m$ are generalized Euler angles, and $K_1,K_2,\ldots,K_l$ equal 1 or (-1) and correspond to reflections

$$\overset{(1)}{\gamma}\sim\overset{(1)}{\lambda} = Bv = B\begin{bmatrix} \overset{(1)}{\sigma_1}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \\ \overset{(1)}{\sigma_2}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \\ \vdots \\ \overset{(1)}{\sigma_{j_0-1}}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \\ 1 \quad \leftarrow j_0 \\ \overset{(1)}{\sigma_{j_0+1}}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \\ \vdots \\ \overset{(1)}{\sigma_n}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \tau_{n-1} \end{matrix}$$

$$\sim B(\alpha_1,\alpha_2,\ldots,\alpha_m,K_1,K_2,\ldots,K_l)\cdot\begin{bmatrix} \overset{(1)}{\sigma_1}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \\ \overset{(1)}{\sigma_2}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \\ \vdots \\ \overset{(1)}{\sigma_{j_0-1}}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \\ 1 \quad \leftarrow j_0 \\ \overset{(1)}{\sigma_{j_0+1}}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \\ \vdots \\ \overset{(1)}{\sigma_n}\Big(\overset{(1)}{\tau_1},\overset{(1)}{\tau_2},\ldots,\overset{(1)}{\tau_{n-2}}\Big) \end{bmatrix} \tag{2.28}$$

j_0 is the number of the row in which 1 appears. It is easy to see that λ depends on $(m+n-2)$ arbitrary parameters and on l integers equal to (± 1). For γ we have

$$\overset{(2)}{\gamma}\sim\overset{(2)}{\lambda}\sim B(\alpha_1,\ldots,a_m,K_1,\ldots,K_l)\cdot\begin{bmatrix} \overset{(2)}{\sigma_1}\Big(\overset{(2)}{\tau_1},\overset{(2)}{\tau_2},\ldots,\overset{(2)}{\tau_{n-3}}\Big) \\ \overset{(2)}{\sigma_2}\Big(\overset{(2)}{\tau_1},\overset{(2)}{\tau_2},\ldots,\overset{(2)}{\tau_{n-3}}\Big) \\ \vdots \\ \overset{(2)}{\sigma_{j_0-1}}\Big(\overset{(2)}{\tau_1},\overset{(2)}{\tau_2},\ldots,\overset{(2)}{\tau_{n-3}}\Big) \\ 0 \quad -j_0 \\ \overset{(2)}{\sigma_{j_0+1}}\Big(\overset{(2)}{\tau_1},\overset{(2)}{\tau_2},\ldots,\overset{(2)}{\tau_{n-3}}\Big) \\ \vdots \\ \overset{(2)}{\sigma_{j_1-1}}\Big(\overset{(2)}{\tau_1},\overset{(2)}{\tau_2},\ldots,\overset{(2)}{\tau_{n-3}}\Big) \\ 1 \quad -j_1 \\ \overset{(2)}{\sigma_{j_1+1}}\Big(\overset{(2)}{\tau_1},\overset{(2)}{\tau_2},\ldots,\overset{(2)}{\tau_{n-3}}\Big) \\ \vdots \\ \overset{(2)}{\sigma_n}\Big(\overset{(2)}{\tau_1},\overset{(2)}{\tau_2},\ldots,\overset{(2)}{\tau_{n-3}}\Big) \end{bmatrix} \tag{2.29}$$

It depends on $(m+n-3)$ arbitrary parameters and l integers equal to ± 1. Note that the 0 appears in the jth row and that the 1 appears in the j_1th row.

In this way we may get that $\overset{(n-k)}{\gamma}$ depends on $(m+n-(K+1))$ free parameters and l integers equal to (± 1).

Then in the column on which $B(\alpha_1,\ldots,\alpha_m,K_1,\ldots,K_l)$ acts, zeros appear at the places of numbers $j_0 j_1,\ldots j_{K-2}$ and at the place of number j_{K-1} the integer 1. The last element

$$\overset{(n-1)}{\gamma} \sim \overset{(n-1)}{\lambda} \sim B(\alpha_1,\alpha_2,\ldots,\alpha_m,K_1,\ldots,K_l)\cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ (n-1) \\ \sigma \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \leftarrow r \\ \\ \\ \\ \\ \leftarrow j_{n-2} \end{matrix} \tag{2.30}$$

$\overset{(n-1)}{\sigma} \neq 0$) has only zeros in column v except the place of number r and the place of number j_{n-2} and it depends on m arbitrary parameters and l integers equal to (± 1).

The division and parametrization of simple elements is not unique. It depends on the choice of sequence $j_0, j_1, \ldots, j_{n-2}$. For any other choice it will be, in general, different.

Now, let us consider the case where matrix A has a zero eigenvalue. Let $\omega_1 = 0$ and have an order l_1. In such a case we may proceed as in the $(n - l_1)$-dimensional case starting with a parametrization of quadric (in projective coordinates)

$$\sum_{j=l_1+1}^{n} z_j v_j^2 = 0 \tag{2.31}$$

assuming that $v_1, v_2, \ldots, v_{l_1}$ are arbitrary, i.e., $v_j = \mu_j$, $j = 1,2,\ldots,l_1$.

Thus we get a classification of simple elements of Eq. (2.5). Let $F_i \subset \mathscr{E}^*$ be a set of all simple elements belonging to one of these classes. If $\lambda \in F_i$, then λ is given by one of formulas (2.28), (2.29), etc. Thus we have found many types of simple elements depending on various numbers of free parameters. Each of them has its own type of simple wave. Hence, according to Eq. (2.4) we have

$$\overset{(i)}{u} = \overset{(i)}{f}(R^i), \quad \text{where } \frac{d\overset{(i)}{f}}{dR^i} = \overset{(i)}{\lambda}(f^i(R^i)),$$
$$R^i = \varphi(\lambda^i_\nu x^\nu), \quad \overset{(i)}{\lambda} \in F_i, \tag{2.32}$$

where i enumerates the types of simple waves admissible by (2.5). Due to free parameters, which are functions of u, and consequently of R^i we may integrate (2.32) and get its exact solutions. It is easy to see that F_i is a submanifold of $\mathscr{E}^*$ whose dimension is equal to the number of parameters.

III. EQUATION OF POTENTIAL FLOW OF A COMPRESSIBLE, PERFECT GAS

Now we consider the equation of potential of velocity for a stationary flow of perfect gas,

$$(c^2 - \phi_x^2)\phi_{xx} + (c^2 - \phi_y^2)\phi_{yy} + (c^2 - \phi_z^2)\phi_{zz} - 2(\phi_x\phi_y\phi_{xy} + \phi_x\phi_z\phi_{xz} + \phi_y\phi_z\phi_{yz}) = 0, \tag{3.1}$$

where

$$c^2 = c_0^2 - [(\mathscr{H}-1)/2](\phi_x^2 + \phi_y^2 + \phi_z^2) \text{ and } \mathbf{v} = \nabla\phi \tag{3.2}$$

(ϕ is the velocity potential and c the velocity of sound). Equation (3.1) is interesting for us in the supersonic region

$$0 \leqslant c^2 \leqslant \phi_x^2 + \phi_y^2 + \phi_z^2. \tag{3.3}$$

In the region the equation is nonelliptical and we have[5]

$$c_0^2\left(\frac{2}{\mathscr{H}+1}\right) \leqslant \phi_x^2 + \phi_y^2 + \phi_z^2 \leqslant c_0^2\left(\frac{2}{\mathscr{H}-1}\right). \tag{3.4}$$

Now we look for the solutions of (3.1) provided that (3.4) is satisfied. Using the fact that for a perfect gas $c^2 = \mathscr{H}p/\rho_0$ and from the adiabatic equation $p = a\rho_0^{\mathscr{H}}$, $a = \text{const}$, we can calculate ρ_0 and p from (3.2),

$$\rho_0 = \left[\frac{1}{a\mathscr{H}}\left(c_0^2 - \frac{\mathscr{H}-1}{2}(\phi_x^2 + \phi_y^2 + \phi_z^2)\right)\right]^{1/(\mathscr{H}-1)},$$
$$p = \left[\frac{1}{a\mathscr{H}}\left(c_0^2 - \frac{\mathscr{H}-1}{2}(\phi_x^2 + \phi_y^2 + \phi_z^2)\right)\right]^{\mathscr{H}/(\mathscr{H}-1)}. \tag{3.5}$$

We are interested in the solutions of Eq. (3.1) that are given in terms of the Riemann invariants. Following the method we transform (3.1), by introducing some new dependent variables, into a quasilinear system of equations of the first order:

$$(c^2 - \varphi_1^2)\varphi_{1,x} + (c^2 - \varphi_2^2)\varphi_{2,y} + (c^2 - \varphi_3^2)\varphi_{3,z} - 2(\varphi_1\varphi_2\varphi_{1,y} + \varphi_1\varphi_3\varphi_{1,z} + \varphi_2\varphi_3\varphi_{2,z}) = 0, \tag{3.6}$$
$$\varphi_{1,y} - \varphi_{2,x} = 0,$$
$$\varphi_{1,z} - \varphi_{3,x} = 0,$$
$$\varphi_{2,z} - \varphi_{3,y} = 0,$$

where

$$\phi_{,x} = \varphi_1, \phi_{,y} = \varphi_2, \phi_{,z} = \varphi_3. \tag{3.7}$$

Thus Eq. (3.1) is transformed into the overdetermined system of four equations for three functions.

IV. SIMPLE INTEGRAL ELEMENTS

Now we write the system (3.7) in the form (2.8) introducing covector λ,

$$Q(\lambda_1,\lambda_2,\lambda_3) = (c^2 - \varphi_1^2)\lambda_1^2 + (c^2 - \varphi_2^2)\lambda_2^2 + (c^2 - \varphi_3^2)\lambda_3^2 - 2(\varphi_1\varphi_2\lambda_1\lambda_2 + \varphi_1\varphi_3\lambda_1\lambda_3 + \varphi_2\varphi_3\lambda_2\lambda_3) = 0. \tag{4.1}$$

Following the procedure described in Sec. II, we parametrize covector λ. To do this we diagonalize the form (4.1) and write the secular equation

$$\det(A - \mu J) = 0,$$

where

$$A = \begin{pmatrix} (c^2 - \varphi_1^2) & -\varphi_1\varphi_2 & -\varphi_1\varphi_3 \\ -\varphi_1\varphi_2 & (c^2 - \varphi_2^2) & -\varphi_2\varphi_3 \\ -\varphi_1\varphi_3 & -\varphi_2\varphi_3 & (c^2 - \varphi_3^2) \end{pmatrix}. \tag{4.2}$$

From (4.2) we get the third-order equation for a value μ,

$$(c^2 - \mu^2)^2(c^2 - (\varphi_1^2 + \varphi_2^2 + \varphi_3^2) - \mu) = 0. \tag{4.3}$$

Hence one obtains two different eigenvalues

$$\mu_1 = c^2, \quad \mu_2 = c^2 - (\varphi_1^2 + \varphi_2^2 + \varphi_3^2) \tag{4.4}$$

from which the first has an order 2. Thus the quadratic form Q reduced to the canonical form is

$$Q(\lambda,\lambda) = c^2(y_1^2 + y_2^2) + (c^2 - (\varphi_1^2 + \varphi_2^2 + \varphi_3^2))y_3^2 = 0. \tag{4.5}$$

Now we find parametrical equations for (4.5). Supposing that $y_l \neq 0$, we simply get

$$\frac{X^2}{(a/c)^2} - Y^2 = 1, \quad \text{where } X = \frac{y_2}{y_1}, \quad Y = \frac{y_3}{y_1}, \tag{4.6}$$

where $a^2 = \varphi_1^2 + \varphi_2^2 + \varphi_3^2 - c^2 > 0$ (supersonic flow).

Then we have

$$Y = \sinh\rho, \quad X = (a/c)\cosh\rho, \tag{4.7}$$

i.e., the parametrical equation of a hyperbola.

If $y_1 = 0$, then one gets

$$c^2y_2^2 - a^2y_3^2 = 0, \tag{4.8}$$

$$y_2 = \mathscr{E}(a/c)y_3, \quad \mathscr{E} = \pm 1. \tag{4.9}$$

In our case there is only one eigenvalue of order larger than 1, i.e., 2. Thus the diagonalizing matrix B will depend only on one parameter α, $B = B(\alpha)$. Matrix B may be easily built from eigenvectors of matrix A and one obtains

$$B(\alpha,l) = \begin{pmatrix} \frac{-1}{\chi_2}\left(\varphi_2\sin\alpha + \frac{\varphi_1\varphi_3}{\chi_1}\cos\alpha\right), & \frac{1}{\chi_2}\left((-1)^K\varphi_2\cos\alpha + (-1)^{K+1}\frac{\varphi_1\varphi_3}{\chi_1}\sin\alpha\right), & \frac{\varphi_1}{\chi_1} \\ \frac{1}{\chi_2}\left(\varphi_1\sin\alpha - \frac{\varphi_2\varphi_3}{\chi_1}\cos\alpha\right), & \frac{(-1)^{K+1}}{\chi_2}\left(\varphi_1\cos\alpha + \frac{\varphi_2\varphi_3}{\chi_1}\sin\alpha\right), & \frac{\varphi_2}{\chi_1} \\ \frac{\chi_2}{\chi_1}\cos\alpha, & \frac{(-1)^K\chi_2}{\chi_1}\sin\alpha, & \frac{\varphi_3}{\chi_1} \end{pmatrix}, \tag{4.10}$$

$K = 0,1$, $l = (-1)^K$, where α is a parameter depending on $\varphi_1,\varphi_2,\varphi_3$ and

$$\chi_1^2 = \varphi_1^2 + \varphi_2^2 + \varphi_3^2, \quad \chi_2^2 = \varphi_1^2 + \varphi_2^2.$$

Integer l equals (± 1) and is connected to a reflection in two-dimensional space, whereas α is connected to a rotation. According to Sec. II we have

$$\overset{(1)}{\gamma} \sim \overset{(1)}{\lambda} = B(\alpha,l)\begin{pmatrix}1\\X\\Y\end{pmatrix} \tag{4.11}$$

and

$$\overset{(2)}{\gamma} \sim \overset{(2)}{\lambda} = B(\alpha,l)\begin{pmatrix}0\\1\\\sigma\end{pmatrix}, \quad \sigma = \mathscr{E}a/c. \tag{4.12}$$

Then by inserting (4.10) into (4.11) and (4.12) we get explicitly the form of simple elements. The elements are presented in Appendix A (F_1 and F_2).

According to results given in Sec. II, $\overset{(1)}{\lambda}$ depends on two parameters α and ρ, whereas $\overset{(2)}{\lambda}$ only depends upon α.

Let us consider form (4.5) once more, now assuming that $y_3 \neq 0$, and let us introduce

$$X = y_2/y_3, \quad Y = y_1/y_3. \tag{4.13}$$

Then we have

$$X^2 + Y^2 = (a/c)^2 \tag{4.14}$$

so

$$X = (a/c)\sin\rho, \quad Y = (a/c)\cos\rho, \tag{4.15}$$

i.e., a parametrical equation of a circle, ρ is a function of φ_i, $i = 1,2,3$. If $y_3 = 0$, we obtain $y_1 = y_2 = 0$, thus a zero case. We have

$$\gamma_1 \sim \overset{(1')}{\lambda} = B(\alpha,l)\begin{pmatrix}Y\\X\\1\end{pmatrix}. \tag{4.16}$$

By inserting (4.10), (4.15) into (4.16) we obtain the explicit form of simple elements. They are presented in $F_{1'}$ (Appendix B). $\overset{(1')}{\lambda}$ depends on two parameters: α and ρ.

However, in a certain case a degeneration occurs and for $K = 0$ we get only one parameter $\beta = \rho + \alpha$, whereas for $K = 1$, $\rho - \alpha = \omega$. In this case $\overset{(1')}{\lambda}$ depends, really, on only one parameter.

Let us consider (4.5) again, assuming that $y_2 \neq 0$. We introduce new coordinates

$$Y = y_3/y_2, \quad X = y_1/y_2 \tag{4.17}$$

and obtain from (4.5)

$$Y^2/(a/c)^2 - X^2 = 1, \tag{4.18}$$

and in a parametrical form

$$Y = (a/c)\cosh\rho, \quad X = \sinh\rho. \tag{4.19}$$

Thus we have

$$\gamma_{1''} \sim \overset{(1'')}{\lambda} = B(\alpha,l)\begin{pmatrix}X\\1\\Y\end{pmatrix}. \tag{4.20}$$

If $y_2 = 0$ we obtain from (4.5)

$$c^2y_1 - a^2y_3 = 0, \tag{4.21}$$

$$y_1 = \mathscr{E}(a/c)y_3, \quad \mathscr{E}^2 = 1 \tag{4.22}$$

and

$$\gamma_{2''} \sim \overset{(2'')}{\lambda} = B(\alpha,l)\begin{pmatrix}\sigma\\0\\1\end{pmatrix}, \text{ where } \sigma = \frac{\mathscr{E}a}{c}. \tag{4.23}$$

Inserting (4.10) into (4.20) and (4.23) we achieve the explicit form of simple elements $\overset{(1'')}{\lambda}$ and $\overset{(2'')}{\lambda}$, which are presented in Appendix C, ($F_{1''}$, $F_{2''}$).

Thus we have found here several types of simple elements which are used to construct solutions, i.e., simple waves and their interactions, so-called double and multiple waves. In Appendices A–C we present all classes of simple elements—$F_1, F_2, F_{1'}, F_{2'}, F_{1''}, F_{2''}$ for (3.1).

V. SIMPLE WAVES

Now we present the simplest solutions of the system (3.1), namely those that have been constructed on the basis of homogeneous simple integral elements. The method of find-

ing these solutions are presented in Refs. 1–3. In this section we deal with a method of solving Eq. (2.3). We use the results of Sec. II concerning parametrization of simple elements. The crucial point of our method is using the freedom of parameters occurring in simple elements. According to terminology presented in Refs. 1 and 2, the elementary solution of the homogeneous system has been called a simple wave. Those solutions may be interpreted as waves since they show moving disturbances, the profile of which changes in the course of propagation [a sign of this is the implicit form of the relation (2.4) for the $R(x)$ function]. The form of solution of (2.4) suggests that covector λ may be regarded as equivalent to the wave vector $(\omega,\mathbf{k})$ which specifies the velocity and direction of propagation of the wave. The specific profile of a simple wave is explicitly determined by its initial data. A certain amount of freedom is connected with the freedom of choosing one free function, which is a function of one variable. The above remarks concern all simple waves which have been found. There exists also another freedom, connected to parameters in simple elements in our method. This freedom is of another kind and origin. Due to it we may integrate Eq. (2.3) and obtain solutions with $(q+1)$ arbitrary function of one variable, where q is a number of independent parameters in the simple elements. For all functions we obtain a certain restriction and it seems to be an interesting point in the method. The simple wave obeys conditions (2.3) and (2.4). By substitution of the simple integral element (4.11) into Eq. (2.3), we obtain for $K=0$

$$\begin{aligned}
\frac{d\varphi_1}{dR} &= \frac{(\chi_1^2-c^2)^{1/2}}{\chi_2}\left[-\left(\varphi_2\sin\alpha+\frac{\varphi_1\varphi_3}{\chi_1}\cos\alpha\right)+\left(\varphi_2\cos\alpha-\frac{\varphi_1\varphi_3}{\chi_1}\sin\alpha\right)\sinh\rho\right]+\frac{c\varphi_1}{\chi_1}\cosh\rho\,,\\
\frac{d\varphi_2}{dR} &= \frac{(\chi_1^2-c^2)^{1/2}}{\chi_2}\left[\left(\varphi_1\sin\alpha-\frac{\varphi_2\varphi_3}{\chi_1}\cos\alpha\right)-\left(\varphi_1\cos\alpha+\frac{\varphi_2\varphi_3}{\chi_1}\sin\alpha\right)\sinh\rho\right]+\frac{c\varphi_2}{\chi_1}\cosh\rho\,,\\
\frac{d\varphi_3}{dR} &= \frac{\chi_2(\chi_1-c^2)^{1/2}}{\chi_1}\left[\cos\alpha+\sin\alpha\sinh\rho\right]+\frac{c\varphi_3}{\chi_1}\cosh\rho\,.
\end{aligned}\tag{5.1}$$

Now we introduce the new dependent variables χ_1, χ_2, and $\mu_0=\varphi_1/\varphi_2$. We get from the system (5.1)

$$\begin{aligned}
\frac{d\chi_1}{dR} &= c\cosh\rho \geqslant c\,,\\
\frac{d\chi_2}{dR} &= \frac{1}{\chi_1}\left[-\mathscr{E}_1(\chi_1^2-\chi_2^2)^{1/2}(\chi_1^2-c^2)^{1/2}\right.\\
&\quad\left.\times(\cos\alpha+\sin\alpha\sinh\rho)+c\chi_2\cosh\rho\right]\,,\\
\frac{d}{dR}\arctan\mu_0 &= ((\chi_1^2-c^2)^{1/2}/\chi_2)(\cos\alpha-\sin\alpha)\,.
\end{aligned}\tag{5.2}$$

Let us observe that quantities α and ρ, existing in Eqs. (5.2) are arbitrary functions of R. Hence, it is convenient to shift the arbitrariness from α and ρ to χ_1 and χ_2. The right-hand sides of the equations for χ_1 and χ_2 depend on α and ρ but they do not contain derivatives of these functions. Thus assuming that χ_1 and χ_2 are arbitrary, we obtain equations for α and ρ. These equations will express α and ρ in terms of functions χ_1 and χ_2 and their derivatives with respect to R. The condition of solvability of algebraic (or transcendental) equations for α and ρ provide us with restrictions on χ_1 and χ_2. In such a way the problem of solving the system of equations for χ_1 and χ_2 is reduced to the problem of solving algebraic equations and searching for restrictions for the arbitrary functions χ_1 and χ_2. We will derive a set of such restrictions. After expressing α and ρ in terms of new arbitrary functions with restrictions we insert these relations into the third equation of (5.2) and to a covector (4.11). In this way, on the right-hand side of this equation we have the functions of R and we can integrate the equation

$$\mu_0=\tan\left(\int_{R_0}^{R}\frac{(\chi_1^2-c^2)^{1/2}}{\chi_2}(\cos\alpha-\sin\alpha)\,dR'+d_0\right),\tag{5.3}$$

$d_0=\text{const}$. Solving equations

$$\mu_0=\varphi_1/\varphi_2\,,\quad \chi_1^2=\varphi_1^2+\varphi_2^2+\varphi_3^2\,,\quad \chi_2^2=\varphi_1^2+\varphi_2^2\,,$$

we find the solution, which is a simple wave,

$$\begin{aligned}
&\varphi_1=\varphi_1(R)\,,\\
&\varphi_2=\varphi_2(R)\,,\\
&\varphi_3=\varphi_3(R)\,,\\
&R=\Psi(\overset{(1)}{\lambda}_1x+\overset{(1)}{\lambda}_2y+\overset{(1)}{\lambda}_3z)\,.
\end{aligned}\tag{5.4}$$

It is easy to see that the conditions for χ_1 and χ_2 imply some restrictions for the range of parameter R and consequently [see (5.4)] a restriction for the function Ψ. This restriction is responsible for the fact that the function Ψ has its range in a certain subset of the real axis.

Now let us realize this program. For convenience of calculating we assume that

$$\chi_1^2=e^{2H}\,,\quad \chi_2^2=e^{2G}\,,\quad \text{where } e^{2G}\leqslant e^{2H}\tag{5.5}$$

and we find restrictions for H and G (H and G are functions of R).

The first restriction for e^{2H} is of course (3.4) and we have

$$c_0^2(2/\mathscr{H}+1)\leqslant e^{2H}\leqslant c_0^2(2/\mathscr{H}-1)\,.\tag{5.6}$$

By substituting (5.5) into equations one and two of (5.2) and then using the relations between trigonometric and hyperbolic functions we get

$$\begin{aligned}
&\cosh\rho=\frac{e^H}{c}\frac{dH}{dR}\,,\\
&t^2-\frac{2AB}{B^2+1}t+\frac{A^2-1}{B^2+1}=0\,,
\end{aligned}\tag{5.7}$$

where

$$A=\frac{e^{(G+H)}\dfrac{d}{dR}(G-H)}{(e^{2H}-e^{2G})^{1/2}(e^{2H}-c^2)^{1/2}}\,,$$

$$B = \sinh\rho = \mathscr{E}_2\left(\frac{e^{2H}}{c^2}\left(\frac{dH}{dR}\right)^2 - 1\right)^{1/2}, \tag{5.8}$$

$$t = \sin\alpha\,,\ \cos\alpha = \mathscr{E}(1-t^2)^{1/2}\,,$$

$$\mathscr{E}_2^2 = \mathscr{E}_3^2 = 1\,.$$

Thus, we get the first restrictions

$$\text{A.}\quad 1 \leqslant \cosh\rho = \frac{e^H}{c}\frac{dH}{dR}\,, \tag{5.9}$$

$$\text{B.}\quad |t| \leqslant 1\,.$$

At the same time the quadratic equation (5.7) (second equation) must have real roots, so its Δ cannot be negative. The latter condition yields

$$B^2 - A^2 + 1 \geqslant 0\,. \tag{5.10}$$

By inserting A and B into (5.10) we obtain the following condition:

$$\alpha_1 z_1^2 - \alpha_2 z_2^2 + \alpha_3 z_1 z_2 \geqslant 0\,, \tag{5.11}$$

$$z_1 = \frac{dH}{dR}\,,\quad z_2 = \frac{dG}{dR}\,,$$

where

$$\alpha_1 = e^{4H} - e^{2(H+G)} - c^2 e^{2H}\,,$$

$$\alpha_2 = c^2 e^{2G} \geqslant 0\,,\quad \alpha_3 = 2c^2 e^{2G} \geqslant 0\,.$$

Supposing that $G \neq \text{const}$, we introduce the new variable $z_3 = z_1/z_2$ and we obtain

$$\alpha_1 z_3^2 + \alpha_3 z_3 - \alpha_2 \geqslant 0\,. \tag{5.12}$$

Now we present the conditions for which (5.12) will be always satisfied regardless of the value z_3. By doing this we avoid differential inequalities that are hard to satisfy. These conditions are the following: Δ_1 of the quadratic equation (5.12) must be nonpositive, whereas the coefficient of z_3^2 must be positive. So, we have

$$0 \geqslant \Delta_1 = -e^{4H} + c^2 e^{2G} + e^{2(G+H)} - c^2 e^{2H}\,, \tag{5.13}$$

$$0 < \alpha_1 = e^{4H} - e^{2(H+G)} - c^2 e^{2H}\,. \tag{5.14}$$

The condition (5.13) is stronger than (5.14) and it is sufficient to fulfill only (5.13). By inserting (5.5) and (3.2) into (5.13) we obtain

$$0 < (\mathscr{H}+1)e^{4H} + (\mathscr{H}-3)e^{2(H+G)} - 2c_0^2(e^{2G} + e^{2H})\,. \tag{5.15}$$

Using (5.6) we easily conclude that $2c_0^2 - (\mathscr{H}-3)e^{2H} > 0$, so finally we have

$$e^{2G} \geqslant \left[\frac{(\mathscr{H}-1)e^{2H} - 2c_0^2}{2c_0^2 - (\mathscr{H}-3)e^{2H}}\right] e^{2H}\,. \tag{5.16}$$

Let us turn now to condition B. We solve the quadratic equation with respect to t and we get

$$t_{1,2} = \frac{AB \pm (B^2 - A^2 + 1)^{1/2}}{B^2 + 1}\,. \tag{5.17}$$

We choose the root with the smaller modulus,

$$t = \mathscr{E}_2\frac{|A|\cdot|B| - (B^2 - A^2 + 1)^{1/2}}{B^2 + 1}\,,\quad \mathscr{E}_2^2 = 1. \tag{5.17'}$$

Now we will prove that

$$\frac{||A|\cdot|B| - (B^2 - A^2 + 1)^{1/2}|}{B^2 + 1} \leqslant 1 \tag{5.18}$$

when $B^2 - A^2 + 1 \geqslant 0$.

Let us consider the two cases

$$|A|\cdot|B| - (B^2 - A^2 + 1)^{1/2} \leqslant 0\,, \tag{5.19a}$$

$$|A|\cdot|B| - (B^2 - A^2 + 1)^{1/2} \geqslant 0\,. \tag{5.19b}$$

(a) We start from an obvious inequality $(|A| + |B|)^2 \geqslant 0$ and we have

$$(A^2 + B^2 + 2|A|\cdot|B|)(B^2 + 1) \geqslant 0\,, \tag{5.20}$$

i.e.,

$$A^2B^2 + A^2 + B^4 + B^2 + 2|A|\cdot|B| + 2|A|\cdot|B|B^2 \geqslant 0 \tag{5.21}$$

or

$$\begin{aligned} B^2 - A^2 + 1 &\leqslant B^4 + 1 + A^2B^2 + 2|A|\cdot|B| + 2B^2 + 2|A|\cdot|B|^3 \\ &= (B^2 + 1 + |A| + |B|)^2 \end{aligned} \tag{5.22}$$

and from that $(B^2 - A^2 + 1)^{1/2} - |A|\cdot|B| \leqslant (B^2 + 1)$ or

$$|(B^2 - A^2 + 1)^{1/2} - |A|\cdot|B||/B^2 + 1 \leqslant 1\,. \tag{5.23}$$

(b) We have $B^2 - A^2 + 1 \geqslant 0$, that is

$$A^2 \leqslant B^2 + 1\,. \tag{5.24}$$

So,

$$A^2B^2 \leqslant B^2(B^2 + 1) \leqslant (B^2 + 1)^2\,, \tag{5.25}$$

$$|A\cdot B| \leqslant (B^2 + 1)\,. \tag{5.25'}$$

And from that $|AB| \geqslant (B^2 - A^2 + 1)^{1/2}$ we have

$$\left|\frac{|AB| - (B^2 - A^2 + 1)^{1/2}}{B^2 + 1}\right| \leqslant 1\,. \tag{5.26}$$

Summing up, we see that $|t| \leqslant 1$, so condition B is satisfied.

Now let us turn to condition A. This is a differential inequality which may be solved by applying well-known results.[6] We have

$$\frac{e^H}{c(e^H)}\frac{dH}{dR} \geqslant 1\,,\quad H(R_0) = H_0 \tag{5.27}$$

and the function P obeys the equation

$$\frac{1}{c(P)}\frac{dP}{dR} = 1 \text{ and } P(R_0) = e^{H_0}\,. \tag{5.27a}$$

Using the definition of c we can write (5.17) and (5.27a) in the following form:

$$\frac{d}{dR}\left[\sqrt{\frac{2}{\mathscr{H}-1}}\arcsin\left(\frac{\sqrt{(\mathscr{H}-1)/2}\,e^H}{c_0}\right)\right] \geqslant 1 \tag{5.28}$$

and $H(R_0) = H_0$,

$$\frac{d}{dR}\left[\sqrt{\frac{2}{\mathscr{H}-1}}\arcsin\left(\frac{\sqrt{(\mathscr{H}-1)/2}\,P}{c_0}\right)\right] = 1\,, \tag{5.28a}$$

$$P(R_0) = e^{H_0}\,.$$

From (5.28a) we get

$$P(R) = c_0\sqrt{\frac{2}{\mathscr{H}-1}}\sin\left[\sqrt{\frac{2}{\mathscr{H}-1}}(R - R_0) + \arcsin\left(\sqrt{\frac{\mathscr{H}-1}{2}}\frac{e^{H_0}}{c_0}\right)\right]. \tag{5.29}$$

We have

$$\sqrt{\frac{2}{\mathscr{H}-1}}\arcsin\left(\frac{\sqrt{(\mathscr{H}-1)/2}\,e^{H(R)}}{c_0}\right) \geqslant \sqrt{\frac{2}{\mathscr{H}-1}}\arcsin\left(\frac{\sqrt{(\mathscr{H}-1)/2}\,P(R)}{c_0}\right) \tag{5.30}$$

and

$$e^{H(R_0)} = P(R_0) = e^{H_0} \tag{5.31}$$

(see, e.g., Ref. 6), hence

$$e^H \geqslant c_0\sqrt{\frac{2}{\mathscr{H}-1}}\sin\left[\sqrt{\frac{2}{\mathscr{H}-1}}(R-R_0) + \arcsin\left(\frac{\sqrt{(\mathscr{H}-1)/2}e^{H_0}}{c_0}\right)\right], \tag{5.31a}$$

$e^{H(R_0)} = e^{H_0}$.

Simultaneously we have

$$c_0\sqrt{\frac{2}{\mathscr{H}+1}} \leqslant e^{H(R)} \leqslant c_0\sqrt{\frac{2}{\mathscr{H}-1}}, \tag{5.32}$$

$$c_0\sqrt{\frac{2}{\mathscr{H}+1}} \leqslant e^{H_0} \leqslant c_0\sqrt{\frac{2}{\mathscr{H}-1}}. \tag{5.33}$$

Since $P(R_K) = c_0\sqrt{2/(\mathscr{H}-1)}$, for

$$R_K = R_0 + \sqrt{\frac{\mathscr{H}-1}{2}}\left(\frac{\pi}{2} + 2K\pi - \arcsin\left(\frac{1}{c_0}\sqrt{\frac{\mathscr{H}-1}{2}}\,e^{H_0}\right)\right), \tag{5.34}$$

$$K = 0, \pm 1, \pm 2, \ldots,$$

we have

$$e^{H(R_K)} = c_0\sqrt{\frac{2}{\mathscr{H}-1}}. \tag{5.35}$$

We have, of course,

$$|R_{K+1} - R_K| = \sqrt{2(\mathscr{H}-1)}\,\pi. \tag{5.36}$$

Hence, function $e^{H(R)}$ is defined in the interval $(R_0, +\infty)$ for an established but arbitrary R_0.

At points R_K, $c^2 = 0$, the magnitude of the vector of velocity of flow is maximal, and simultaneously pressure and density become equal to zero. Thus, the solution of the system (5.1) is

$$\varphi_1 = \eta_2 e^G \sin K(R), \quad \varphi_2 = \eta_2 e^G \cos K(R), \quad \eta_1^2 = \eta_2^2 = 1, \quad \varphi_3 = \eta_1\sqrt{e^{2H} - e^{2G}}, \tag{5.37}$$

where

$$K(R) = \int_{R_0}^{R}(\cos\alpha - \sin\alpha)\frac{(e^{2H} - c^2)^{1/2}}{e^G}\,dR' + c',$$

$c' = \text{const}$, $R = \Psi(\overset{(1)}{\lambda}_1 x + \overset{(1)}{\lambda}_2 y + \overset{(1)}{\lambda}_3 z)$, $R \in \langle R_0, +\infty) = L$.

The function Ψ takes values only from the interval L. The conditions (5.16), (5.31a), (5.32), (5.33) are implied on functions G and H.

For $\cos\alpha$ and $\sin\alpha$ we have the following expressions:

$$\sin\alpha = \mathscr{E}_1\left\{e^G\frac{d}{dR}(H-G)\left(\frac{e^{2H}}{c^2}\left(\frac{dH}{dR}\right)^2 - 1\right)^{1/2} - \left[\left(\frac{dH}{dR}\right)^2(e^{2H} - e^{2G})\frac{(e^{2H}-c^2)}{c^2} - e^{2G}\left(\frac{d}{dR}(H-G)\right)^2\right]^{1/2}\right\} \times\left[\frac{1}{c^2}e^H\left(\frac{dH}{dR}\right)^2(e^{2H}-e^{2G})^{1/2}\cdot(e^{2H}-c^2)^{1/2}\right]^{-1},$$

$$\cos\alpha = \mathscr{E}_3\left(1 - \left\{\left[-e^{2G}\left(\frac{d}{dR}(H-G)\right)^2\left(\frac{e^{2H}}{c^2}\left(\frac{dH}{dR}\right) - 1\right) + \left(\frac{dH}{dR}\right)^2(e^{2H}-e^{2G})(e^{2H}-c^2)\frac{1}{c^2}\right] \times\left[\frac{1}{c^4}e^{2H}\left(\frac{dH}{dR}\right)^4(e^{2H}-e^{2G})\cdot(e^{2H}-c^2)\right]^{-1}\right\} - e^{2G}\left(\frac{d}{dR}(H-G)^2 + 2e^G\frac{d}{dR}(H-G)\right)\left(\frac{e^{2H}}{c^2}\left(\frac{dH}{dR}\right)^2 - 1\right)^{1/2} \times\left[\left(\frac{dH}{dR}\right)^2(e^{2H}-c^2)(e^{2H}-c^2)\frac{1}{c^2} - e^{2G}\left(\frac{d}{dR}(H-G)\right)^2\right]^{1/2}\left[\frac{1}{c^4}e^{2H}\left(\frac{dH}{dR}\right)^4(e^{2H}-e^{2G})(e^{2H}-c^2)\right]^{-1/2}\Bigg)^{1/2}, \tag{5.38}$$

$$\cosh\rho = \frac{e^H}{c}\frac{dH}{dR}, \quad \sinh\rho = \mathscr{E}_2\left(\frac{e^{2H}}{c^2}\left(\frac{dH}{dR}\right)^2 - 1\right)^{1/2}, \tag{5.39}$$

where

$$\mathscr{E}_1^2 = \mathscr{E}_2^2 = \mathscr{E}_3^2 = 1.$$

Of course, expressions (5.38) and (5.39) should be inserted into covector λ, like expressions (5.37).

Now let us consider the following equation:

$$R_K = \Psi(\overset{(1)}{\lambda}_1(R_K)x + \overset{(1)}{\lambda}_2(R_K)y + \overset{(1)}{\lambda}_3(R_K)z), \quad R_K \in L, \tag{5.40}$$

where R_K is given by (5.34). For a given K, Eq. (5.40) describes one or several planes with a normal vector $\overset{(1)}{\lambda}(R_K) = (\overset{(1)}{\lambda}_1(R_K), \overset{(1)}{\lambda}_2(R_K), \overset{(1)}{\lambda}_3(R_K))$. If the equation $R_K = \Psi(r)$ possesses n roots r_i, $i = 1,2,\ldots,n$, $\Psi(r_i) = R_K$, then there are n planes:

$$\lambda_1(R_K)x + \lambda_2(R_K)y + \lambda_3(R_K)z = r_i, \quad i = 1,2,\ldots,n. \tag{5.41}$$

Thus, they are parallel planes. When the function Ψ is single valued we have only one plane. For different $K_1 \neq K_2$ the planes belonging to K_i do not have to be parallel and in general they cross each other. The sections of planes may also cross.

Let us notice that (5.41) is a place at which both density and pressure disappear and the magnitude of the velocity reaches its maximal value. So in fact they are planes of "nodes" of density and pressure and of "antinodes" of the magnitude of the velocity vector. Because planes may cross each other we have also straight lines and points of "nodes" and "antinodes." Thus, the above solution may be treated as a nonlinear analog of a standing wave.

Now let us turn to the case $G = G_0 = \text{const}$, $dG/dR = 0$. We must specify the condition for $\Delta \geqslant 0$ [compare (5.11)]. Then we have $z_2 = 0$ and (5.11) is reduced to

$$\alpha_1 z_1^2 \geqslant 0 . \tag{5.42}$$

Supposing that $z_1 \neq 0$, $H \neq \text{const}$ we obtain $\alpha_1 \geqslant 0$, which by using (5.14) is reduced to

$$e^{2H} \geqslant \left(\frac{2}{\mathscr{H}-1}\right)(c_0^2 + e^{2G_0}) . \tag{5.43}$$

And now we may repeat all the considerations concerning the conditions A and B and achieve the same results as before with the substitution $G = G_0 = \text{const}$. The only difference will be another restriction for the lowest value of e^H and e^{H_0} and the absence of restriction (5.16).

The case $H = \text{const}$ leads to a nonphysical solution. From the equation $e^H(dH/dR) = c \cosh \rho \geqslant c$ we have that $c = 0$, which leads to vanishing of density and pressure everywhere, thus to the absence of gas.

For all simple elements from Appendix A it is possible to repeat the considerations and it was done in Ref. 7. But the most interesting case is described in this section.

VI. GAUGE STRUCTURE AND "BÄCKLUND TRANSFORMATION"

Now we construct some geometrical structures for simple (Riemann) waves of (2.5). These structures establish relations between Riemann waves of (2.5) and allow us to introduce nonlinear transformations connecting two Riemann waves [exact solutions of (2.5)]. The nonlinear transformations are of gauge type, and may be treated as "Bäcklund transformation" [8] for (2.5).

Namely, let matrix $A = (a_{ij})$ have K eigenvalues ω_i, $i = 1,2,...,K$, each of order l_i, $\Sigma_{i=1}^K l_i = n$.

In such a case we have a natural group acting on simple elements λ. $\Pi_{i=1}^K \otimes \mathrm{O}(l_i)$ each of the group $\mathrm{O}(l_i)$ acts on the coordinates $v_j, j = \Sigma_{r=1}^{i-1} l_r, \Sigma_{r=1}^{i-1} l_r + 1, ..., \Sigma_{r=1}^{i} l_r$ without destroying the diagonalization of matrix A. It is easy to see that $\alpha_1, \alpha_2, ..., \alpha_m$ are parameters of the group $\Pi_{i=1}^K \otimes \mathrm{O}(l_i)$, $m = \Sigma_{i=1}^K \frac{1}{2} l_i(l_i - 1)$.

Simultaneously we can connect the group $\mathrm{O}(p,q)$ with the cone

$$\sum_{j=1}^{\eta} \zeta_j v_j^2 = 0 \quad (\zeta_j \text{ is equal to one of } \omega_i) . \tag{6.1}$$

$$\text{Let } \zeta_j = \mathscr{E}_j |\zeta_j| , \text{ where } \mathscr{E}_j = \operatorname{sgn} \zeta_j . \tag{6.2}$$

By transforming v_j to v_j', $j = 1,2,...,n$,

$$v_j' = \sqrt{|\zeta_j|}\, v_j , \tag{6.3}$$

we transform (6.1) into

$$\sum_{j=1}^{\eta} \mathscr{E}_j v_j'^2 = 0 , \tag{6.4}$$

i.e., into a canonical cone. The group $\mathrm{O}(p,q)$ conserves a quadratic form

$$Q(v',v') = \sum_{i=1}^{\eta} \mathscr{E}_i v_i'^2 , \tag{6.5}$$

where p = number of integers $\mathscr{E}_i$ equal to 1, q = number of integers $\mathscr{E}_i$ equal to (-1) in the sum (6.5). Obviously $p + q = n$ (we assume there exist no zero eigenvalues).

Now, let us notice that classes of simple elements, and, in consequence, simple waves, which are constructed according to Sec. II, are related to the choice of a concrete chain of subgroups $\mathrm{O}(p,q)$. This chain ends on the two-element group $\{e, -e\}$ or the trivial group $\{e\}$, hence

$$\mathrm{O}(p,q) \supset \mathrm{O}(p_1,q_1) \supset \mathrm{O}(p_2,q_2) \supset \cdots \begin{matrix} \supset \{e, -e\} \\ \supset \{e\} \end{matrix} , \tag{6.6}$$

where for $p_i, q_i, p_{i+1}, q_{i+1}$ we have the following relations: either $p_i = p_{i+1}, q_i = q_{i+1} + 1$

$$\text{or } p_i = p_{i+1} + 1 , \quad q_i = q_{i+1} , \tag{6.7}$$

$p_0 = p$, $q_0 = q$.

In this way the dimension of the space in which the group operates, diminishes to 1 according to Sec. II. The choice of the sequence of subscripts $j_0 j_1 \cdots j_{K-2}$ corresponds to one of the possible chains of subgroups (6.6). Thus with each simple element we can connect in a natural way the following group:

$$L_i = \left[\otimes \prod_{r=1}^{K} \mathrm{O}(l_i) \right] \otimes \mathrm{O}(p_i,q_i) . \tag{6.8}$$

The origin of each factor of the simple product is, of course, different. In general, we connect with Eq. (2.5) the group

$$L = \left[\otimes \prod_{r=1}^{K} \mathrm{O}(l_i) \right] \otimes \mathrm{O}(p,q) . \tag{6.9}$$

Group L_i acts on a submanifold $F_i \subset \mathscr{E}^*$ (the manifold of simple elements of a chosen class according to the classification from Sec. II). Since a simple element is a function of a point of hodograph space $\mathscr{H}$ (the space of values of the solutions of the equation) we may connect with the equation very natural fiber bundles. For every class of simple elements we have a fiber bundle P_i over the base space $\mathscr{H}$ with structural group L_i, typical fiber F_i, and projection $\pi_i : P_i \to \mathscr{H}$.

It is easy to see that $\dim(L_i) = \dim(F_i)$ and for every simple element, $\lambda \in F_i$ we have

$$\lambda = g \cdot \lambda_0 , \quad \text{where } g \in L_i \text{ and } \begin{matrix} \lambda_0 \in F_i , \\ \lambda_0 = \text{const} . \end{matrix} \tag{6.10}$$

Taking a local section of P_i we get

$$\lambda(u) = g(\alpha(u))\lambda_0 , \quad u \in \mathscr{D} \subset \mathscr{H} , \tag{6.11}$$

where α is the set of all parameters of the group L_i. But in the case of simple waves we have Riemann invariant R (parametrization in hodograph space $\mathscr{H}$). Thus we obtain a special structure, a bundle Π_i over the base space $\mathscr{R}$ (a straight line of the Riemann invariant) with structural group L_i, typical fiber F_i, and with a projection $\overline{\Pi}_i : \Pi_i \to \mathscr{R}$. For every local section of Π_i we get

$$\lambda(R) = g(\alpha(R))\lambda_0 , \quad R \in V \subset \mathscr{R} . \tag{6.12}$$

Every local section f gives us a simple wave belonging to a chosen class of simple waves (simple elements). If we have

two local sections f and g we have two different simple waves of the same type. If we change the section from f to g, we change functions $\alpha(R)$ to $\beta(R)$ and we get

$$g(\alpha(R)) = h(R)g(\beta(R)), \quad \lambda(\alpha(R)) = h(R)\lambda(\beta(R)), \quad h(R) \in L. \tag{6.13}$$

Thus we see that the action of the "gauge group" of L_i over straight line $\mathscr{R}$ (Riemann invariant) on a simple wave creates a new simple wave of the same type ("gauge group" means that the parameters of L_i depend on R). In Sec. V we solved Eq. (2.4) using arbitrary functions $\alpha(R)$. We shift the freedom from the α's to new, more convenient functions. We should do this for functions α and β independently. For α and β one gets algebraic (or transcendental) equations. These equations will express α (or β) in terms of new functions and their first derivatives with respect to R. The condition of solvability of the algebraic (or transcendental) equations provides us with restrictions for the new functions. Varying $\alpha(R)$ to $\beta(R)$ we change these new functions and their first derivatives. Thus we get the gauge transformation connecting two exact solutions (simple waves of the same type). This transformation is very similar to classical Bäcklund transformation.[8]

For Eq. (3.1) we have the following situation:

$$L = \mathrm{O}(2) \otimes \mathrm{O}(1,2). \tag{6.14}$$

And we used the following chains of subgroups of O(1,2):

$$\begin{aligned} &\mathrm{O}(1,2) \supset \overset{(1)}{\mathrm{O}}(1,1) \supset \{e, -e\}, \\ &\mathrm{O}(1,2) \supset \mathrm{O}(2) \supset \{e, -e\}, \\ &\mathrm{O}(1,2) \supset \overset{(2)}{\mathrm{O}}(1,1) \supset \{e, -e\}. \end{aligned} \tag{6.15}$$

All these chains correspond to examined simple elements and simple waves and we obtain the following gauge groups:

$$\begin{aligned} &L_1 = \mathrm{O}(2) \otimes \overset{(1)}{\mathrm{O}}(1,1), \quad L_{1'} = \mathrm{O}(2) \otimes \mathrm{O}(2), \\ &L_{2'} = \mathrm{O}(2) \otimes \{e, -e\}, \\ &L_2 = \mathrm{O}(2) \otimes \{e, -e\}, \quad L_{1''} = \mathrm{O}(2) \otimes \overset{(2)}{\mathrm{O}}(1,1), \\ &L_{2''} = \mathrm{O}(2) \otimes \{e, -e\}. \end{aligned} \tag{6.16}$$

The case with $L_{1'} = \mathrm{O}(2) \otimes \mathrm{O}(2)$ is very interesting because we have simple elements corresponding to that group with only one arbitrary parameter. Two parameters of $\mathrm{O}(2) \otimes \mathrm{O}(2)$ become one parameter of the "diagonal group O(2)."

Now let us write down the explicit form of the action of the gauge group on functions H and G for the case (5.1). Let us suppose that there exists one exact solution—a simple wave with parameters α and ρ and corresponding arbitrary functions H and G. In this case we have $L_i = \mathrm{O}(2) \otimes \mathrm{O}(1,1) = L_1$, α is a parameter of O(2) and ρ of O(1,1).

We change functions ρ and α into $\rho + \Delta\rho$ and $\alpha + \Delta\alpha$. It is a change of gauge by means of functions $\Delta\rho$ and $\Delta\alpha$. And we look at how G and H will change. In this way we obtain the explicit action of the gauge group on the manifold of functions H and G and their first derivatives with respect to R.

We have

$$\cosh\rho = \frac{e^H}{c(e^H)}\frac{dH}{dR}, \quad \cosh(\rho + \Delta\rho) = \frac{e^{H_1}}{c(e^{H_1})}\frac{dH_1}{dR} \tag{6.17}$$

but

$$\cosh(\rho + \Delta\rho) = \cosh\rho\cosh\Delta\rho + \sinh\rho\sinh\Delta\rho. \tag{6.18}$$

Inserting (6.17) and (5.39) into (6.18) we get

$$\frac{e^{H_1}}{c(e^{H_1})}\frac{dH_1}{dR} = \frac{e^H}{c(e^H)}\frac{dH}{dR}\cosh(\Delta\rho) + \mathscr{E}_2\left(\frac{e^{2H}}{c^2(e^{2H})}\left(\frac{dH}{dR}\right)^2 - 1\right)^{1/2}\sinh(\Delta\rho). \tag{6.19}$$

(6.19) is the nonlinear representation of the gauge group originating from O(1,1) on the manifold of functions H and G and their first derivatives with respect to R ($H, H_1, G, G_1, \alpha, \rho, \Delta\alpha, \Delta\rho$ are functions of R).

For O(2) we get similarly from (6.20)

$$\sin(\alpha + \Delta\alpha) = \sin\alpha\cos\Delta\alpha + \sin\Delta\alpha\cos\alpha, \tag{6.20}$$

$$\begin{aligned} &\mathscr{E}_1\left\{e^{G_1}\frac{d}{dR}(H_1 - G_1)\left(\frac{e^{2H_1}}{c_1^2}\left(\frac{dH_1}{dR}\right)^2 - 1\right)^{1/2} - \left[\left(\frac{dH_1}{dR}\right)^2(e^{2H_1} - e^{2G_1})\frac{(e^{2H_1} - c^2)}{c^2} - e^{2G_1}\left(\frac{d}{dR}(H_1 - G_1)\right)^2\right]^{1/2}\right\} \\ &\quad\times\left[\frac{1}{c_1^2}e^{H_1}\left(\frac{dH_1}{dR}\right)^2(e^{2H_1} - e^{2G_1})^{1/2}\cdot(e^{2H_1} - c_1^2)^{1/2}\right]^{-1} \\ &= \cos(\Delta\alpha)\cdot\mathscr{E}_2\left\{e^G\frac{d}{dR}(H - G)\left(\frac{e^{2H}}{c^2}\left(\frac{dH}{dR}\right)^2 - 1\right)^{1/2} - \left[\left(\frac{dH}{dR}\right)^2(e^{2H} - e^{2G})\frac{(e^{2H} - c^2)}{c^2} - e^{2G}\left(\frac{d}{dR}(H - G)\right)^2\right]^{1/2}\right\} \\ &\quad\times\left[\frac{1}{c^2}e^H\left(\frac{dH}{dR}\right)^2(e^{2H} - e^{2G})^{1/2}\cdot(e^{2H} - c^2)^{1/2}\right]^{-1} \\ &\quad + \sin(\Delta\alpha)\cdot\mathscr{E}_3\left(1 - \left\{\left[e^{2G}\left(\frac{d}{dR}(H - G)\right)^2\left(\frac{e^{2H}}{c^2}\left(\frac{dH}{dR}\right)^2 - 1\right) + \left(\frac{dH}{dR}\right)^2(e^{2H} - e^{2G})(e^{2H} - c^2)\frac{1}{c^2}\right]\right. \right. \\ &\quad\left.\left.\times\left[\frac{1}{c^4}e^{2H}\left(\frac{dH}{dR}\right)^4(e^{2H} - e^{2G})\cdot(e^{2H} - c^2)\right]^{-1}\right\}\right. \\ &\quad + \left\{e^{2G}\left(\frac{d}{dR}(H - G)\right)^2 + 2e^G\frac{d}{dR}(H - G)\left(\frac{e^{2H}}{c^2}\left(\frac{dH}{dR}\right)^2 - 1\right)^{1/2}\cdot\left[\left(\frac{dH}{dR}\right)^2(e^{2H} - e^{2G})(e^{2H} - c^2)\frac{1}{c^2}\right.\right. \\ &\quad\left.\left. + e^{2G}\left(\frac{d}{dR}(H - G)\right)^2\right]^{1/2}\right\}\left[\frac{1}{c^4}e^{2H}\left(\frac{dH}{dR}\right)^4(e^{2H} - e^{2G})(e^{2H} - c^2)\right]^{-1}\Bigg)^{1/2}. \end{aligned} \tag{6.21}$$

Restrictions on functions H, H_l, G, G_l and a range of the parameter R given in Sec. V have been already imposed (see Sec. V).

Relations (6.19), (6.21) are analogs of the Bäcklund transformation for Eq. (5.1). At the same time this action is a certain representation of the gauge group (a local one) on the manifold of arbitrary functions and their first derivatives parametrizing the solution. In this way the "Bäcklund" transformation for Eq. (5.1) is a nonlinear representation of the gauge group (a local one) which has originated from the group L_l.

ACKNOWLEDGMENTS

I thank Professor R. Rączka and Professor R. Żelazny for continuous support during preparation of the present version of the paper.

APPENDIX A

Covectors λ, (F_1,F_2), $K=0$

$$\overset{(1)}{\underset{(K=0)}{\lambda}} = \begin{pmatrix} \frac{(\chi_1^2-c^2)^{1/2}}{\chi_2}\left[-\left(\varphi_2\sin\alpha+\frac{\varphi_1\varphi_3}{\chi_1}\cos\alpha\right)+\left(\varphi_2\cos\alpha-\frac{\varphi_1\varphi_3}{\chi_1}\sin\alpha\right)\sinh\rho\right]+\frac{c\varphi_1}{\chi_1}\cosh\rho \\ \frac{(\chi_1^2-c^2)^{1/2}}{\chi_2}\left[\left(\varphi_1\sin\alpha-\frac{\varphi_2\varphi_3}{\chi_1}\cos\alpha\right)-\left(\varphi_1\cos\alpha+\frac{\varphi_2\varphi_3}{\chi_1}\sin\alpha\right)\sinh\rho\right]+\frac{c\varphi_2}{\chi_1}\cosh\rho \\ \frac{\chi_2(\chi_1^2-c^2)^{1/2}}{\chi_1}(\cos\alpha+\sin\alpha\sinh\rho)+\frac{c\varphi_3}{\chi_1}\cosh\rho \end{pmatrix},$$

$$\overset{(2)}{\underset{(K=0)}{\lambda}} = \begin{pmatrix} \frac{\mathscr{E}(\chi_1^2-c^2)^{1/2}}{\chi_2}\left(\varphi_2\cos\alpha-\frac{\varphi_1\varphi_3}{\chi_1}\sin\alpha\right)+\frac{c\varphi_1}{\chi_1} \\ \frac{-\mathscr{E}(\chi_1^2-c^2)^{1/2}}{\chi_2}\left(\varphi_2\cos\alpha+\frac{\varphi_2\varphi_3}{\chi_1}\sin\alpha\right)+\frac{c\varphi_2}{\chi_1} \\ \frac{\mathscr{E}(\chi_1^2-c^2)^{1/2}}{\chi_1}\chi_2\sin\alpha+\frac{c\varphi_3}{\chi_1} \end{pmatrix}.$$

Covectors λ, $K=1$

$$\overset{(1)}{\underset{(K=1)}{\lambda}} = \begin{pmatrix} \frac{(\chi_1^2-c^2)^{1/2}}{\chi_2}\left[-\left(\varphi_2\sin\alpha+\frac{\varphi_1\varphi_3}{\chi_1}\cos\alpha\right)+\left(-\varphi_2\cos\alpha+\frac{\varphi_1\varphi_3}{\chi_1}\sin\alpha\right)\sinh\rho\right]+\frac{c\varphi_1}{\chi_1}\cosh\rho \\ \frac{(\chi_1^2-c^2)^{1/2}}{\chi_2}\left[\left(\varphi_1\sin\alpha-\frac{\varphi_1\varphi_3}{\chi_1}\cos\alpha\right)+\left(\varphi_1\cos\alpha+\frac{\varphi_2\varphi_3}{\chi_1}\sin\alpha\right)\sinh\rho\right]+\frac{c\varphi_2}{\chi_1}\cosh\rho \\ \frac{\chi_2(\chi_1^2-c^2)^{1/2}}{\chi_1}(\cos\alpha-\sin\alpha\sinh\rho)+\frac{c\varphi_3}{\chi_1}\cosh\rho \end{pmatrix},$$

$$\overset{(2)}{\underset{(K=1)}{\lambda}} = \begin{pmatrix} \frac{\mathscr{E}(\chi_1^2-c^2)^{1/2}}{\chi_2}\left(-\varphi_2\cos\alpha+\frac{\varphi_1\varphi_3}{\chi_1}\sin\alpha\right)+\frac{c\varphi_1}{\chi_1} \\ \frac{\mathscr{E}(\chi_1^2-c^2)^{1/2}}{\chi_2}\left(\varphi_1\cos\alpha+\frac{\varphi_2\varphi_3}{\chi_1}\sin\alpha\right)+\frac{c\varphi_2}{\chi_1} \\ \frac{-\mathscr{E}(\chi_1^2-c^2)^{1/2}}{\chi_1}\chi_2\sin\alpha+\frac{c\varphi_3}{\chi_1} \end{pmatrix}.$$

APPENDIX B

Covectors λ', $K=0$, $(F_{1'},F_{2'})$

$$\overset{(1')}{\underset{(K=0)}{\lambda}} = \begin{pmatrix} \frac{(\chi_1^2-c^2)^{1/2}}{c\chi_2}\left(\varphi_2\cos\beta-\frac{\varphi_1\varphi_3}{\chi_1}\sin\beta\right)+\frac{\varphi_1}{\chi_1} \\ \frac{-(\chi_1^2-c^2)^{1/2}}{c\chi^2}\left(\varphi_1\cos\beta+\frac{\varphi_2\varphi_3}{\chi_1}\sin\beta\right)+\frac{\varphi_2}{\chi_1} \\ \frac{(\chi_1^2-c^2)^{1/2}}{c\chi_1}\sin\beta+\frac{\varphi_3}{\chi_1} \end{pmatrix}.$$

Covectors λ', $K=1$

$$\overset{(1')}{\lambda} = \begin{pmatrix} \dfrac{(\chi_1^2 - c^2)^{1/2}}{c\chi_2}\left(-\varphi_2 \cos\omega - \dfrac{\varphi_1\varphi_3}{\chi_1}\sin\omega\right) + \dfrac{\varphi_1}{\chi_1} \\ \dfrac{(\chi_1^2 - c^2)^{1/2}}{c\chi_2}\left(\varphi_1 \cos\omega - \dfrac{\varphi_2\varphi_3}{\chi_1}\sin\omega\right) + \dfrac{\varphi_2}{\chi_1} \\ \dfrac{(\chi_1^2 - c^2)^{1/2}}{c\chi_1}\sin\omega + \dfrac{\varphi_3}{\chi_1} \end{pmatrix}.$$

APPENDIX C

Covectors λ'', $(F_{1''}, F_{2''})$

$$\overset{(1'')}{\underset{K=0}{\lambda}} = \begin{pmatrix} \dfrac{(\chi_1^2 - c^2)^{1/2}}{\chi_2}\left[-\dfrac{\varphi_1\varphi_3}{\chi_1}(\cos\alpha\sinh\rho + \sin\alpha) - \varphi_2(\sinh\rho\sin\alpha + \cos\alpha)\right] + \dfrac{c\varphi_1}{\chi_1}\cosh\rho \\ \dfrac{(\chi_1^2 - c^2)^{1/2}}{\chi_2}\left[-\dfrac{\varphi_2\varphi_3}{\chi_1}(\sinh\rho\cos\alpha + \sin\alpha) + \varphi_1(\sinh\rho\sin\alpha - \cos\alpha)\right] + \dfrac{c\varphi_2}{\chi_1}\cosh\rho \\ \dfrac{(\chi_1^2 - c^2)^{1/2}}{\chi_1}\chi_2(\sinh\rho\cos\alpha + \sin\alpha) + \dfrac{c\varphi_3}{\chi_1}\cosh\rho \end{pmatrix},$$

$$\overset{(1'')}{\underset{K=1}{\lambda}} = \begin{pmatrix} \dfrac{(\chi_1^2 - c^2)^{1/2}}{\chi_2}\left[-\dfrac{\varphi_1\varphi_3}{\chi_1}(\sinh\rho\cos\alpha - \sin\alpha) - \varphi_2(\sinh\rho\sin\alpha - \cos\alpha)\right] + \dfrac{c\varphi_1}{\chi_1}\cosh\rho \\ \dfrac{(\chi_1^2 - c^2)^{1/2}}{\chi_2}\left[\dfrac{\varphi_2\varphi_3}{\chi_1}(-\sinh\rho\cos\alpha + \sin\alpha) + \varphi_1(\cos\alpha + \sinh\rho\sin\alpha)\right] + \dfrac{c\varphi_2}{\chi_1}\cosh\rho \\ \dfrac{(\chi_1^2 - c^2)^{1/2}}{\chi_1}\chi_2(\sinh\rho\cos\alpha - \sin\alpha) - \dfrac{c\varphi_3}{\chi_1}\cosh\rho \end{pmatrix},$$

$$\overset{(2'')}{\lambda} = \begin{pmatrix} \dfrac{-\mathscr{E}(\chi_1^2 - c^2)^{1/2}}{c\chi_2}\left(\varphi_2 \sin\alpha + \dfrac{\varphi_1\varphi_3}{\chi_1}\cos\alpha\right) + \dfrac{\varphi_1}{\chi_1} \\ \dfrac{\mathscr{E}(\chi_1^2 - c^2)^{1/2}}{c\chi_2}\left(\varphi_1 \sin\alpha - \dfrac{\varphi_2\varphi_3}{\chi_1}\cos\alpha\right) + \dfrac{\varphi_2}{\chi_1} \\ \dfrac{\mathscr{E}(\chi_1^2 - c^2)^{1/2}}{c\chi_1}\chi_2\cos\alpha + \dfrac{\varphi_3}{\chi_1} \end{pmatrix}, \quad K = 0,1,$$

where $\alpha, \rho, \beta, \omega$ are arbitrary functions of φ_i, $i = 1,2,3$ and $\chi_1^2 = \varphi_1^2 + \varphi_2^2 + \varphi_3^2$, $\chi_2^2 = \varphi_1^2 + \varphi_2^2$, $\mathscr{E}^2 = 1$.

[1] C. E. B. Riemann, "Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite," Abhandl. Königl. Ges. Wiss. Göttingen **8**, 1869.

[2] M. Burnat, "Theory of simple waves for nonlinear systems of partial differential equations and applications to gas dynamics," Arch. Mech. Stosow. **18**, 4 (1966).

[3] Z. Peradzyński, "Nonlinear plane K-waves and Riemann invariants," Bull. Acad. Pol. Sci. Ser. Sci. Tech. **19**, 9 (1971).

[4] K. Borsuk, *Multidimensional analytic geometry* (PWN, Warsaw, 1969).

[5] L. Landau and E. Lifshits, *Mechanics of continuous media* (Mir, Moscow, 1952) (in Russian).

[6] J. Szarski, *Differential inequalities* (PWN, Warsaw, 1967).

[7] M. W. Kalinowski, "On the Old-New Method of Solving Nonlinear Equations," Warsaw University preprint IFT/13/80, Warsaw, 1980.

[8] F. A. E. Pirani, D. C. Robinson, and W. F. Shadwick, *Local Jet Bundle formulation of Bäcklund transformation* (Reidel, London, 1979).

V^*-algebras: A particular class of unbounded operator algebras

Giuseppina Epifanio and Camillo Trapani
Istituto di Fisica dell'Università di Palermo, Palermo, Italy

(Received 22 July 1982; accepted for publication 30 March 1984)

We consider a weak unbounded commutant for a set of unbounded operators and we examine op*-algebras which coincide with some of their bicommutant. This class of op*-algebras, called V^*-algebras, shows some properties close to those which hold true for bounded operator algebras.

PACS numbers: 02.30.Tb, 03.65.Db

I. INTRODUCTION

Algebras of unbounded operators have been the subject, in recent years, of many studies (Refs. 1–9), particularly in view of their applications to quantum theories.

From some point of view, these algebras seemed to be the best generalization of algebras of bounded operators in Hilbert space. But a deeper study has shown that a large number of pathologies arises and that many properties of algebras of bounded operators may fail to be true for them.

For this reason it seems to be convenient to select, among the most general unbounded operator algebras, classes showing a better behavior. In this spirit, many interesting works have been published (see, for instance, Refs. 10–12).

The present paper may be considered as a further attempt in this direction.

After having summarized, in Sec. II, notations and basic definitions, in Sec. III we introduce a weak unbounded commutant for a set of unbounded operators and discuss some simple topological properties of it.

As is natural, a definition of a commutant leads us to consider op*-algebras which coincide with some of their bicommutants. Thus we introduce, in Sec. IV, V^*-algebras and SV^*-algebras and show that in these algebras it is possible to overcome some difficulties which arise in the study of the most general unbounded operator algebras. Actually, some properties which hold true for bounded operator algebras can be proved for V^*-algebras.

In Sec. V we show that the bounded part of a V^*-algebra is a von Neumann algebra and that a V^*-algebra is not necessarily an EW^*-algebra (Refs. 10 and 11). In fact, in a V^*-algebra a symmetric element need not be essentially self-adjoint.

On the other hand we prove that an arbitrary von Neumann algebra is allowed to be the bounded part of a nontrivial V^*-algebra.

Finally, in Sec. VI, we briefly discuss some situations where an algebra $\mathscr{A}$ generates a V^*-algebra and examine, furthermore, the abelian case.

For what concerns applications, it has been proved elsewhere (Ref. 13) that V^*-algebras provide a natural framework when discussing the problem of the completeness of a set of compatible observables, a problem whose classical formulation is given in terms of von Neumann algebras.

As in the above case, we guess in the mathematical description of many physical theories (such as quantum field theory, quantum statistical mechanics, etc.) that V^*-algebras could appropriately be substituted for von Neumann algebras, as far as the use of bounded operator algebras is not something intrinsic in those theories but is only needed for technical reasons. (With respect to this point see, e.g., Haag's discussion in Ref. 14.)

II. NOTATION AND PRELIMINARY DEFINITIONS

Let $\mathscr{D}$ be a scalar product space. We will denote with $C_{\mathscr{D}}$ the $*$-algebra of all linear operators in $\mathscr{D}$ which have an adjoint in $\mathscr{D}$; or, equivalently, the $*$-algebra of all $\sigma(\mathscr{D}, \mathscr{D})$-continuous operators. The $\sigma(\mathscr{D}, \mathscr{D})$-topology is understood to be that defined by the set of seminorms

$$\{\varphi \to |(\varphi,\psi)|, \psi \in \mathscr{D}\}.$$

The algebra $C_{\mathscr{D}}$ is often called $\mathscr{L}^+(\mathscr{D})$ by other authors.

We will denote with $\mathscr{H}$ the Hilbert space which is the norm-completion of $\mathscr{D}$. Then $C_{\mathscr{D}}$ can be meant to be the set of all closable operators A in $\mathscr{H}$ having $\mathscr{D}$ as a dense common invariant domain and such that $A^*\mathscr{D} \subseteq \mathscr{D}$. The involution in $C_{\mathscr{D}}$ is then defined by $A \to A^+$ with $A^+ = A^* \restriction \mathscr{D}$.

An op*-algebra $\mathscr{A}$ in $\mathscr{D}$ is an involutive subalgebra, with unity, of $C_{\mathscr{D}}$.

An op*-algebra $\mathscr{A}$ on $\mathscr{D}$ is said to be (i) closed: if $\mathscr{D} = \tilde{\mathscr{D}}(\mathscr{A}) = \cap_{A\in\mathscr{A}} D(\overline{A})$ or, equivalently, if $\mathscr{D}$ is complete under the $\mathscr{A}$-graph topology defined in $\mathscr{D}$ by the set of seminorms $\varphi \to \|A\varphi\|$, $A \in \mathscr{A}$; (ii) self-adjoint: if $\mathscr{D} = \mathscr{D}^*(\mathscr{A}) = \cap_{A\in\mathscr{A}} D(A^*)$; (iii) symmetric: if $\forall A \in \mathscr{A}$ $(1 + A^+A)^{-1} \in \mathscr{A}_b = \mathscr{A} \cap \mathscr{B}(\mathscr{H})$; (iv) standard: if each symmetric element A of $\mathscr{A}$ (i.e., $A = A^+$) is essentially self-adjoint in $\mathscr{D}$ or, equivalently, if $\forall A \in \mathscr{A}$ $\overline{A^+} = A^*$ results.

III. COMMUTANTS

In the following we will be concerned with the concept of a commutant of a set of operators. We will give a definition of a commutant in such a way that also unbounded operators may belong to it.

Definition 3.1: Let $\mathscr{H}$ be a Hilbert space and $\mathscr{D}$ a dense linear manifold of $\mathscr{H}$. With $C(\mathscr{D},\mathscr{H})$ we will indicate the set of all closable operators A in $\mathscr{H}$ such that $\mathscr{D} \subseteq D(A) \cap D(A^*)$.

It is clear that $B(\mathscr{H}) \subseteq C(\mathscr{D},\mathscr{H})$ and $C_{\mathscr{D}} \subseteq C(\mathscr{D},\mathscr{H})$.

It is easily seen that $C(\mathscr{D},\mathscr{H})$ is a $*$-invariant linear space. (A set $\mathscr{O}$ of operators is said to be $*$-invariant if $A \in \mathscr{O}$ implies $A^* \in \mathscr{O}$.)

It is possible to define in $C(\mathscr{D},\mathscr{H})$ many different topologies. We will use the following ones.

 0022-2488/84/092633-05$02.50

(i) $\mathscr{D}$-weak topology is defined by the set of seminorms

$$A \in C(\mathscr{D},\mathscr{H}) \to |(A\varphi,\psi)| \varphi,\psi \in \mathscr{D}.$$

(ii) $\mathscr{D}$-strong topology is defined by the set of seminorms

$$A \in C(\mathscr{D},\mathscr{H}) \to \|A\varphi\| \varphi \in \mathscr{D}.$$

The map $A \to A^*$, which is continuous with respect to the $\mathscr{D}$-weak topology, is not, in general, $\mathscr{D}$-strongly continuous.

Let $\mathscr{O}$ be a $*$-invariant subset of $C(\mathscr{D},\mathscr{H})$. The weak bounded commutant $\mathscr{O}'_w$ is understood to be the set

$$\mathscr{O}'_w = \{B \in \mathscr{B}(\mathscr{H}): (Af,B^*g) = (Bf,A^*g) \quad \forall f,g \in \mathscr{D}, \quad \forall A \in \mathscr{O}\}.$$

The set $\mathscr{O}'_w$ is a $*$-invariant linear space, but it is not, in general, an algebra.

We define a weak unbounded commutant as the set

$$\mathscr{O}'_\sigma = \{X \in C(\mathscr{D},\mathscr{H}): (Af,X^*g) = (Xf, A^*g) \quad \forall f,g \in \mathscr{D}, \quad \forall A \in \mathscr{O}\}.$$

It is clear that $\mathscr{O}'_w = \mathscr{O}'_\sigma \cap B(\mathscr{H})$.

The second commutant is now defined as $\mathscr{O}''_{\sigma\sigma} = (\mathscr{O}'_\sigma)'_\sigma$ and the commutants of higher order in a similar way; it is easy to see that $\mathscr{O}'_\sigma = \mathscr{O}'''_{\sigma\sigma\sigma}$, etc.

Both $\mathscr{O}'_\sigma$ and $\mathscr{O}''_{\sigma\sigma}$ are $*$-invariant linear subspaces of $C(\mathscr{D},\mathscr{H})$.

It is also possible to define a commutant in $C_\mathscr{D}$ for a $*$-invariant subset of $C(\mathscr{D},\mathscr{H})$. We will call it $\mathscr{O}'_C$ and define it as $\mathscr{O}'_C = \mathscr{O}'_\sigma \cap C_\mathscr{D}$.

If $\mathscr{O}$ is a $*$-invariant subset of $C(\mathscr{D},\mathscr{H})$, in analogy to what is usually made for op*-algebras, we indicate with $\mathscr{D}^*(\mathscr{O})$ the subset of $\mathscr{H}$: $\mathscr{D}^*(\mathscr{O}) = \cap_{A \in \mathscr{O}} D((A \restriction \mathscr{D})^*)$. Clearly $\mathscr{D} \subseteq \mathscr{D}^*(\mathscr{O})$.

We will now give some topological properties of the commutant.

Proposition 3.2: Let $\mathscr{O}$ be a $*$-invariant subset of $C(\mathscr{D},\mathscr{H})$ such that $\mathscr{O}\mathscr{D} \subseteq \mathscr{D}$. Then (i) $\mathscr{O}'_\sigma$ is $\mathscr{D}$-weakly closed; (ii) $\mathscr{O}'_\sigma \mathscr{D} \subseteq \mathscr{D}^*(\mathscr{O})$.

Proof: (i) is an easy consequence of the fact that for $A \in \mathscr{O}$, $X \in C(\mathscr{D},\mathscr{H})$ and $f, g \in \mathscr{D}$, the linear functional in $C(\mathscr{D},\mathscr{H})$

$$\omega(X) = (Xf,A^*g) - (Af,X^*g)$$

is $\mathscr{D}$-weakly continuous in $C(\mathscr{D},\mathscr{H})$. (ii) If $A \in \mathscr{O}$, $X \in \mathscr{O}'_\sigma$, $f, g \in \mathscr{D}$ we have

$$|(Af,X^*g)| = |(Xf,A^*g)| = |(f,X^*A^*g)| \leqslant k\|f\| \quad \forall f \in \mathscr{D}, \quad (k = \|X^*A^*g\|).$$

Therefore, $X^*g \in D((A \restriction \mathscr{D})^*)$.

Remark 3.3: Since $\mathscr{O}'_\sigma$ need not, in general, leave $\mathscr{D}$ invariant $\mathscr{O}''_{\sigma\sigma}$ may fail to be $\mathscr{D}$-weakly closed.

Corollary 3.4: Let $\mathscr{A}$ be an op*-algebra. We have (i) $\mathscr{A}'_\sigma$ is $\mathscr{D}$-weakly closed; (ii) $\mathscr{A}'_\sigma = \{X \in C(\mathscr{D},\mathscr{H}): X\mathscr{D} \subseteq \mathscr{D}^*(\mathscr{A}), XAf = A^+{}^*Xf, \forall f \in \mathscr{D}, \forall A \in \mathscr{A}\}$; (iii) if $\mathscr{A}$ is self-adjoint, $\mathscr{A}'_\sigma = \mathscr{A}'_C$ and $\mathscr{A}''_{\sigma\sigma}$ is $\mathscr{D}$-weakly closed.

The proof is straightforward.

In contrast with the bounded case the bicommutant of an op*-algebra $\mathscr{A}$ may fail to be the closure of $\mathscr{A}$. However, we have

Proposition 3.5: Let $\mathscr{A}$ be an op*-algebra such that the $\mathscr{D}$-strong closure $\overline{\mathscr{A}}^s$ of $\mathscr{A}$ in $C(\mathscr{D},\mathscr{H})$ is contained in $\mathscr{A}''_{w\sigma}$. If one of the following conditions is fulfilled, then $\mathscr{A}''_{w\sigma}$ coincides with the $\mathscr{D}$-strong closure of $\mathscr{A}$ in $C(\mathscr{D},\mathscr{H})$.

(i) $\mathscr{A}$ consists only of bounded operators.

(ii) $\mathscr{R}'_w = \mathscr{A}'_w$ for some op*-algebra $\mathscr{R} \subseteq \mathscr{A}_b = \mathscr{A} \cap \mathscr{B}(\mathscr{H})$.

(iii) For each $\varphi \in \mathscr{D}$ the norm-closure in $\mathscr{D}$ of the subspace $\mathscr{A}\varphi$ is orthocomplemented in $\mathscr{D}$.

Proof: (i) We need only to prove that $\mathscr{A}''_{w\sigma} \subseteq \overline{\mathscr{A}}^s$. Let $\varphi \in \mathscr{D}$, $\varphi \neq 0$. We consider the subspace $\overline{\mathscr{A}\varphi}$ of $\mathscr{H}$ (the closure will be meant in the norm-topology of $\mathscr{H}$). Let P be the projection operator onto this subspace. Clearly, each $A \in \mathscr{A}$ leaves $\overline{\mathscr{A}\varphi}$ invariant, and then $PA = AP$ results. Hence, $P \in \mathscr{A}'_w$. If $B \in \mathscr{A}''_{w\sigma}$ and $\psi \in \mathscr{D}$ we get

$$((1-P)B\varphi,\psi) = (B\varphi,(1-P)\psi) = ((1-P)\varphi,B^*\psi) = 0.$$

Hence, $B\varphi = PB\varphi$, for the density of $\mathscr{D}$, or, equivalently, $B\varphi \in \overline{\mathscr{A}\varphi}$. This is turn implies that $B \in \overline{\mathscr{A}}^s$.

(ii) Follows easily from (i).

(iii) The argument is almost the same used in the proof of (i). Notice that any $A \in \mathscr{A}$ leaves $\overline{\mathscr{A}\varphi}$ invariant because $\overline{\mathscr{A}\varphi}$ being orthocomplemented, coincides also with the $\sigma(\mathscr{D},\mathscr{D})$-closure of $\mathscr{A}\varphi$ and A is $\sigma(\mathscr{D},\mathscr{D})$-continuous. Therefore, the projection P onto $\overline{\mathscr{A}\varphi}$ is an element of $\mathscr{A}'_w$. If we denote by $\overline{P}$ the continuous extension of P to $\mathscr{H}$, in the same way as in (i), it can be proved that $B\varphi = \overline{P}B\varphi$, $\forall B \in \mathscr{A}''_{w\sigma}$. This means that $B\varphi$ is an element of the closure of $\mathscr{A}\varphi$ with respect to the norm-topology of $\mathscr{H}$. Therefore, $B \in \overline{\mathscr{A}}^s$.

Remark 3.6: In the proof of (i) we did not use the fact that $\mathscr{A}$ leaves $\mathscr{D}$ invariant. Thus the statement remains true for an arbitrary $*$-algebra with unity of bounded operators.

Remark 3.7: Under the stronger assumption $\overline{\mathscr{A}}^s \subseteq \mathscr{A}''_{\sigma\sigma}$ we get, as is obvious, the same statement for $\mathscr{A}''_{\sigma\sigma}$ and moreover $\mathscr{A}''_{\sigma\sigma} = \mathscr{A}''_{w\sigma}$ results.

Notice that if we require that the $\mathscr{D}$-weak closure $\overline{\mathscr{A}}^w$ of $\mathscr{A}$ is contained in $\mathscr{A}''_{w\sigma}$ we get also $\mathscr{A}''_{w\sigma} = \overline{\mathscr{A}}^w = \overline{\mathscr{A}}^s$.

Corollary 3.8: If $\mathscr{A}$ is a closed and symmetric op*-algebra, then $\mathscr{A}''_{\sigma\sigma} = \mathscr{A}''_{w\sigma} = \overline{\mathscr{A}}^s = \overline{\mathscr{A}}^w$.

Proof: Since a closed and symmetric op*-algebra is self-adjoint (Ref. 11, Proposition 2.6), by Corollary 3.4, both $\mathscr{A}''_{\sigma\sigma}$ and $\mathscr{A}''_{w\sigma}$ are $\mathscr{D}$-weakly closed. Furthermore, for a symmetric op*-algebra $(\mathscr{A}_b)'_w = \mathscr{A}'_w$ results (Ref. 12, Lemma 2). The statement follows from (ii) of Proposition 3.5.

The assumptions (ii) and (iii) of Proposition 3.5 seem to be independent; however, we have

Proposition 3.9: Let $\mathscr{A}$ be a self-adjoint op*-algebra such that $\mathscr{A}'_w = (\mathscr{A}_b)'_w$. Then for each $\varphi \in \mathscr{D}$ the norm closure of the subspace $\mathscr{A}\varphi$ is orthocomplemented in $\mathscr{D}$.

Proof: Let $\overline{\mathscr{A}\varphi}$ be the norm-closure in $\mathscr{H}$ of the subspace $\mathscr{A}\varphi$ and let P be the corresponding projection operator. Since the bounded elements of $\mathscr{A}$ leave $\overline{\mathscr{A}\varphi}$ invariant, $P \in (\mathscr{A}_b)'_w = \mathscr{A}'_w$. Therefore, by Corollary 3.4, P leaves $\mathscr{D}$ invariant.

The converse of this proposition is not, in general, true. In fact, let T be a self-adjoint operator in $\mathscr{H}$ and $\mathscr{T}$ the algebra generated by the restriction of T to $\mathscr{D}^\infty(T) = \cap_{n>1} D(T^n)$. As is known $\mathscr{T}$ is self-adjoint on $\mathscr{D}^\infty(T)$. Let $\varphi \in \mathscr{D}^\infty(T)$ and $\mathscr{M} = \mathscr{T}\varphi$. As will be shown later (Proposition 6.1) the norm-closure $\overline{\mathscr{M}}$ of $\mathscr{M}$ in $\mathscr{D}^\infty(T)$ is orthocomplemented in $\mathscr{D}^\infty(T)$. On the other hand, in this case, $\mathscr{T}'_w \neq (\mathscr{T}_b)'_w$ because $\mathscr{T}_b = \mathbb{C}1$.

The assumption $(\mathscr{A}_b)'_w = \mathscr{A}'_w$ has also the following topological consequence:

Proposition 3.10: Let $\mathscr{A}$ be an op*-algebra such that $\overline{\mathscr{A}}^s \subseteq \mathscr{A}''_{w\sigma}$. If $(\mathscr{A}_b)'_w = \mathscr{A}'_w$ then $\mathscr{A}_b$ is $\mathscr{D}$-strongly dense in $\mathscr{A}$.

Proof: Because of Proposition 3.5 $(\mathscr{A}_b)''_{w\sigma} = \overline{\mathscr{A}}^s_b$. Therefore, $\mathscr{A}''_{w\sigma} = (\mathscr{A}_b)''_{w\sigma} \subseteq \overline{\mathscr{A}}^s_b \subseteq \overline{\mathscr{A}}^s$. Then $\mathscr{A}''_{w\sigma} = \overline{\mathscr{A}}^s = \overline{\mathscr{A}}^s_b$.

Remark 3.11: By a slight modification of Ref. 12, Lemma 2, for a symmetric op*-algebra $\mathscr{A}$ the equality $(\mathscr{A}_b)'_\sigma = \mathscr{A}'_\sigma$ can be proved.

IV. BICOMMUTANTS AND V^*-ALGEBRAS

We will now introduce the concept of V^*-algebra.

Definition 4.1: Let $\mathscr{A}$ be an op*-algebra on $\mathscr{D}$. We say that $\mathscr{A}$ is regular (respectively, completely regular) if $\mathscr{A}''_{\sigma\sigma} \subseteq C_{\mathscr{D}}$ (resp., $\mathscr{A}''_{w\sigma} \subseteq C_{\mathscr{D}}$).

A regular (resp., completely regular) op*-algebra is said to be a V^*-algebra (resp., an SV*-algebra) if $\mathscr{A}''_{\sigma\sigma} = \mathscr{A}$ (resp. $\mathscr{A}''_{w\sigma} = \mathscr{A}$).

Finally, we say that a V^*-algebra (resp., an SV^*-algebra) is a $\overline{V}^*$-algebra (resp., an $\overline{SV}^*$-algebra) if it is $\mathscr{D}$-weakly closed.

Clearly, if $\mathscr{A}$ is completely regular, it is regular and if it is an SV^*-algebra it is a V^*-algebra.

Remark 4.2: If $\mathscr{D} = \mathscr{H}$ [and therefore, $C(\mathscr{D},\mathscr{H}) = B(\mathscr{H})$] the concepts of V^*-algebra, $\overline{V}^*$-algebra, SV^*-algebra, and $\overline{SV}^*$-algebra coincide with the usual concept of von Neumann algebra.

As a consequence of Corollary 3.4 we have

Proposition 4.3: If $\mathscr{A}$ is a self-adjoint op*-algebra, then $\mathscr{A}'_\sigma$ is a $\overline{V}^*$-algebra.

Corollary 4.4: If $\mathscr{A}$ is a commutative self-adjoint op*-algebra, then both $\mathscr{A}'_\sigma$ and $\mathscr{A}''_{\sigma\sigma}$ are $\overline{V}^*$-algebras.

Remark 4.5: If, for a given op*-algebra $\mathscr{A}$, $\mathscr{A}'_\sigma$ is a V^*-algebra, then $\mathscr{A}'_w$ is a von Neumann algebra.

We already noticed that if the bicommutant of a given op*-algebra $\mathscr{A}$ is also an op*-algebra, then $\mathscr{A}''_{\sigma\sigma}$ is a V^*-algebra. Besides, it is clear that $\mathscr{A}''_{\sigma\sigma}$ is also minimal for this property. If $\mathscr{A}''_{\sigma\sigma}$ is an op*-algebra, we say that it is the V^*-algebra generated by $\mathscr{A}$. On the other hand, the condition $\mathscr{A}''_{\sigma\sigma} \subseteq C_{\mathscr{D}}$ being not always fulfilled, it is not always possible to find a V^*-algebra generated by a given op*-algebra $\mathscr{A}$.

In order to find conditions for the regularity of an op*-algebra we have to look for information about the range of the operators of $\mathscr{A}''_{\sigma\sigma}$. However, a difficulty arises from the fact that in an op*-algebra the left multiplication may fail to be continuous in the topology induced on $\mathscr{A}$ by the $\mathscr{D}$-strong topology of $C(\mathscr{D},\mathscr{H})$.

Proposition 4.6: Let $\mathscr{A}$ be an op*-algebra and $\overline{\mathscr{A}}^s$ its $\mathscr{D}$-strong closure in $C(\mathscr{D},\mathscr{H})$. If the left multiplication is $\mathscr{D}$-strongly continuous in $\mathscr{A}$, then for each $A \in \overline{\mathscr{A}}^s$, $A\mathscr{D} \subseteq \widetilde{\mathscr{D}}(\mathscr{A})$ results.

Thus, if $\mathscr{A}$ is a closed op*-algebra and $\mathscr{A}''_{\sigma\sigma} \subseteq \overline{\mathscr{A}}^s$ (resp., $\mathscr{A}''_{w\sigma} \subseteq \overline{\mathscr{A}}^s$) then $\mathscr{A}$ is regular (resp., completely regular).

Proof: Let $A \in \overline{\mathscr{A}}^s$. Then there exists a net $(A_\alpha) \subseteq \mathscr{A}$ such that $A_\alpha \overset{s}{\to} A$, i.e., $A_\alpha\varphi \to A\varphi \ \ \forall\varphi \in \mathscr{D}$. If $B \in \mathscr{A}$, by the hypothesis, $BA_\alpha\varphi$ is a Cauchy net with respect to the norm of $\mathscr{D}$, or, equivalently, $A_\alpha\varphi$ is a Cauchy net with respect to the $\mathscr{A}$-topology of $\mathscr{D}$. Then there is an element $\psi \in \widetilde{\mathscr{D}}(\mathscr{A})$ such that $A_\alpha\varphi \to \psi$ in the $\mathscr{A}$-topology. Of course, $\psi = A\varphi$. Therefore, $A\varphi \in \widetilde{\mathscr{D}}(\mathscr{A})$.

Notice that if the algebra $\mathscr{A}$ is dominated by a subalgebra of its center, the left multiplication is continuous.

Let us now give some properties of V^*-algebras.

In a previous paper,[7] we examined some spectral properties of the *-algebra $C_{\mathscr{D}}$. We gave, in particular, a sufficient condition in order that a self-adjoint operator of $C_{\mathscr{D}}$ admit a spectral decomposition with spectral measure with values in the same algebra. This proposition was, afterwards, proved by Antoine and Mathot (Ref. 8, Corollary 5.3) under weaker assumptions. We examine now this problem for a V^*-algebra.

Proposition 4.7: Let $\mathscr{A}$ be an op*-algebra and let $T \in \mathscr{A}$ be an essentially self-adjoint operator in $\mathscr{D}$ and $E(\lambda)$ the spectral family of $\overline{T}$. Then $E(\lambda) \in \mathscr{A}''_{w\sigma}$.

Consequently, if $\mathscr{A}$ is completely regular then $E(\lambda)\mathscr{D} \subseteq \mathscr{D}$; moreover, if $\mathscr{A}$ is an SV^*-algebra then $E(\lambda) \restriction \mathscr{D} \in \mathscr{A}$.

Proof: It is known, by the classical spectral theorem, that $E(\lambda)$ commutes with all bounded operators commuting with $\overline{T}$. Therefore, $E(\lambda) \in \{\overline{T}\}''$ (the usual bicommutant). But since $\mathscr{D}$ is an invariant core for $\overline{T}$ we have (Ref. 8, Lemma 5.1) $\{\overline{T}\}' = \{\overline{T}\}'_w$. Thus $\{\overline{T}\}'' = \{T\}''_{ww} \subseteq \{T\}''_{w\sigma} \subseteq \mathscr{A}''_{w\sigma}$.

Now let T be an essentially self-adjoint operator belonging to an op*-algebra $\mathscr{A}$. We will shortly discuss the behavior of the functions of T, when they exist (an analogous study was made in Ref. 15, from another point of view).

Proposition 4.8: Let $\mathscr{A}$ be an op*-algebra on $\mathscr{D}$ and T an essentially self-adjoint operator of $\mathscr{A}$.

Let u be a measurable function of a real variable, finite and determined almost everywhere with respect to the spectral family $E(\lambda)$ of $\overline{T}$. If

$$\mathscr{D} \subseteq \{\varphi \in \mathscr{H} : \int_{-\infty}^{\infty} |u(\lambda)|^2 d(E(\lambda)\varphi,\varphi) < \infty\}, \qquad (1)$$

then $u(\overline{T}) \in \mathscr{A}''_{w\sigma}$.

Consequently, if $\mathscr{A}$ is a completely regular op*-algebra, then $u(\overline{T})\mathscr{D} \subseteq \mathscr{D}$.

Furthermore, if $\mathscr{A}$ is an SV^*-algebra then $u(T) = u(\overline{T}) \restriction \mathscr{D} \in \mathscr{A}$.

Proof: As is known, $u(\overline{T})$ commutes, in the usual sense, with each element of $\{\overline{T}\}'$. As we saw before $\{\overline{T}\}' = \{T\}'_w$ and therefore, $u(\overline{T}) \in \{T\}''_{w\sigma} \subseteq \mathscr{A}''_{w\sigma}$.

Remark 4.9: If $u(\lambda)$ is a bounded function, then (1) is

automatically fulfilled, thus all bounded functions of T belong to $\mathscr{A}''_{w\sigma}$.

Remark 4.10: Proposition 4.8 extends immediately to functions of a family $\{A_1 \dots A_n\}$ of strongly commuting self-adjoint operators (i.e., with commuting spectral projections).

As a simple application of the previous propositions, let us now say a few words about one-parameter groups of unitary operators which play, as is known, an important role for applications to quantum theories.

Proposition 4.11: Let $\mathscr{A}$ be an SV^*-algebra on $\mathscr{D}$ and $U(t)$ a one-parameter continuous group of unitary operators in $\mathscr{H}$. Let A be the infinitesimal generator of the group $U(t)$. If $A \restriction \mathscr{D} \in \mathscr{A}$ and is essentially self-adjoint in $\mathscr{D}$, then $U(t) \restriction \mathscr{D} \in \mathscr{A}$.

The converse of the above proposition is not in general true because we do not know if $\mathscr{D}$ is contained in the domain of the infinitesimal generator. However, we can give the following:

Proposition 4.12: Let $\mathscr{A}$ be a $\overline{V}^*$-algebra on $\mathscr{D}$ and $U(t)$ a one-parameter continuous group of unitary operators such that $U(t) \restriction \mathscr{D} \in \mathscr{A}$ and whose infinitesimal generator A contains $\mathscr{D}$ in its domain. Then $A \restriction \mathscr{D} \in \mathscr{A}$.

Proof: If $\mathscr{D} \subseteq D(A)$ then $A \in C(\mathscr{D},\mathscr{H})$. Because

$$Af = \lim_{h\to 0} \frac{1}{ih}(U(h) - 1)f \quad \forall f \in D(A),$$

A belongs to the $\mathscr{D}$-strong closure of $\mathscr{A}$ in $C(\mathscr{D},\mathscr{H})$. But $\mathscr{A}$ is, by hypothesis, $\mathscr{D}$-strongly closed. Therefore, $A \in \mathscr{A}$.

V. THE BOUNDED PART OF A V^*-ALGEBRA

In this section we will shortly discuss the structure of the bounded part of a V^*-algebra and compare V^*-algebras with EW^*-algebras, introduced by Dixon and extensively studied by Inoue (Refs. 10 and 11).

Proposition 5.1: Let $\mathscr{A}$ be a $\overline{V}^*$-algebra. Its bounded part is a von Neumann algebra.

Proof: Since $\mathscr{A}$ is $\mathscr{D}$-weakly closed in $C(\mathscr{D},\mathscr{H})$, $\mathscr{A}_b$ is $\mathscr{D}$-weakly closed in $B(\mathscr{H})$ and therefore, it is closed in the usual weak topology of $B(\mathscr{H})$.

Proposition 5.1 does not imply, in general, that a V^*-algebra is an EW^*-algebra (i.e., a symmetric op*-algebra whose bounded part is a von Neumann algebra) for a V^*-algebra need not be symmetric, as the following example shows.

Example: It is shown in Ref. 5 (Theorem 7.1) that, for an abelian op*-algebra $\mathscr{A}$, standardness is equivalent to commutativity of $\mathscr{A}''_{ww}$ and it is also shown that there exists a self-adjoint abelian op*-algebra with trivial commutant, $\mathscr{A}'_w = \mathbb{C}1$ (Ref. 5, Example 3). This op*-algebra is therefore, nonstandard. But since it is self-adjoint and abelian, it is regular (Corollary 4.4), and thus $\mathscr{A}''_{\sigma\sigma}$ is a V^*-algebra. Were $\mathscr{A}''_{\sigma\sigma}$ an EW^*-algebra, then it would be standard (Ref. 11, Theorem 2.3) and so $\mathscr{A}$ would also be standard.

Proposition 5.2: Let $\mathscr{A}$ be a self-adjoint SV^*-algebra. Then $\mathscr{A}$ is an EW^*-algebra.

Proof: By the hypotheses, $\mathscr{A} = \mathscr{A}''_{cc} = \mathscr{A}''_{\sigma\sigma} = \mathscr{A}''_{w\sigma}$. By Ref. 12, Lemma 1, $\forall A \in \mathscr{A}''_{cc}$, $\overline{A}\eta\mathscr{A}''_{ww} = (\mathscr{A}''_{w\sigma})_b = (\mathscr{A}''_{\sigma\sigma})_b$. Therefore, $A^*\overline{A}\eta(\mathscr{A}''_{\sigma\sigma})_b$, and thus $(1 + A^*\overline{A})^{-1} \restriction \mathscr{D} \in (\mathscr{A}''_{\sigma\sigma})_b$. Hence, $\mathscr{A}$ is an EW^*-algebra.

Proposition 5.3: Let $\mathscr{A}$ be a closed EW^*-algebra and assume, moreover that $\mathscr{A}$ is $\mathscr{D}$-strongly closed in $C(\mathscr{D},\mathscr{H})$. Then $\mathscr{A}$ is a closed SV^*-algebra.

Proof: It is an easy consequence of Corollary 3.8.

Remark 5.4: As a consequence of the previous propositions a closed and symmetric op*-algebra which is $\mathscr{D}$-strongly closed is an EW^*-algebra.

An interesting question arises now: What kind of von Neumann algebra may occur as the bounded part of a V^*-algebra? We will answer by showing that any von Neumann algebra is the bounded part of a certain V^*-algebra.

Actually, we have

Proposition 5.5: Let $\mathscr{N}$ be a von Neumann algebra. Then there exists a pre-Hilbert space $\mathscr{D}$ and a V^*-algebra $\mathscr{V}$ on $\mathscr{D}$ such that $\mathscr{V}_b = \mathscr{N}$.

Proof: If $\mathscr{N} = \mathscr{B}(\mathscr{H})$, then $\mathscr{D} = \mathscr{H}$ and $\mathscr{N}''_{\sigma\sigma} = B(\mathscr{H})$. If $\mathscr{N} = \mathbb{C}1$, for an arbitrary $\mathscr{D}$, we have $\mathscr{N}'_\sigma = C(\mathscr{D},\mathscr{H})$ and thus $\mathscr{N}''_{\sigma\sigma} = \mathbb{C}1$; so $\mathscr{V} = \mathbb{C}1$.

Apart from the trivial cases, if $\mathscr{N}$ is a von Neumann algebra it is possible to find a self-adjoint unbounded operator T affiliated with $\mathscr{N}'$. Let $\mathscr{T}$ be the algebra generated by T on $\mathscr{D}^\infty(T)$. We get $\mathscr{N} \subseteq \mathscr{T}'_w = \mathscr{T}'$. Then $\mathscr{N} \subseteq \mathscr{T}'_\sigma$. This is turn implies $\mathscr{N}''_{\sigma\sigma} \subseteq \mathscr{T}'_\sigma \subseteq C_{\mathscr{D}}$.

Let $\mathscr{V} = \mathscr{N}''_{\sigma\sigma}$. $\mathscr{V}$ is a V^*-algebra. It remains only to prove that $\mathscr{V}_b = \mathscr{N}$.

Since $\mathscr{N} \subseteq \mathscr{N}''_{\sigma w}$ and $\mathscr{N}' = \mathscr{N}'_w \subseteq \mathscr{N}'_\sigma$ we get $\mathscr{N}''_{\sigma w} \subseteq (\mathscr{N}')'_w = \mathscr{N}'' = \mathscr{N}$. Therefore, $\mathscr{N} = \mathscr{N}''_{\sigma w} = \mathscr{V}_b$.

Remark 5.6: The V^*-algebra obtained in this way is, of course not unique.

VI. SOME SPECIAL SITUATIONS

Let us now discuss some situation where an algebra $\mathscr{A}$ generates a V^*-algebra. In particular, we will consider the case of countably dominated op*-algebras and the abelian case. We will produce an example.

Proposition 6.1: Let $\mathscr{A}$ be an op*-algebra on $\mathscr{D}$ and assume that there exists an essentially self-adjoint operator $T \in \mathscr{A}$ such that $\mathscr{D}^\infty(T) = \mathscr{D}$. Then (i) $\mathscr{A}$ is self-adjoint and, therefore, $\mathscr{A}'_\sigma$ is a $\overline{V}^*$-algebra; (ii) if there exists $z \in \rho(\overline{T})$ such that $(T - z)^{-1} \in \mathscr{A}$ then $\mathscr{A}''_{\sigma\sigma} = \mathscr{A}''_{w\sigma} = \overline{\mathscr{A}}^s = \overline{\mathscr{A}}^w$; (iii) if T is an element of the center of $\mathscr{A}$ then $\mathscr{A}$ is completely regular; (iv) in the hypothesis of (iii), both $\mathscr{A}'_\sigma$ and $\mathscr{A}''_{\sigma\sigma}$ are symmetric $\overline{SV}^*$-algebras and consequently $\mathscr{A}''_{w\sigma} = \mathscr{A}''_{\sigma\sigma}$.

Proof: (i) follows from the fact that, by the closed graph theorem, $\mathscr{A}$ is dominated by the self-adjoint subalgebra $\mathscr{T}$ generated by T.

(ii) We will prove that the norm-closure of the subspaces $\mathscr{A}f$ is orthocomplemented in $\mathscr{D}$ for each $f \in \mathscr{D}$. The statement will follow from Proposition 3.5 and 3.8.

Without loss of generality we can assume $z = 0$. The operator T^{-n} leaves $\overline{\mathscr{A}f}$ invariant; in fact let $\psi = \lim_{k\to\infty} A_k f$ (Hilbert norm), then $T^{-n}\psi = \lim_{k\to\infty} T^{-n}A_k f \in \overline{\mathscr{A}f}$. Let P be the projection onto $\overline{\mathscr{A}f}$, then $PT^{-n} = T^{-n}P$. Let $\varphi,\psi \in \mathscr{D}$, $\varphi = T^{-n}h$ for some $h \in \mathscr{D}$. Then we get

$$(T^n\varphi, P\psi)$$
$$= (h, P\psi) = (Ph,\psi) = (Ph,T^{-n}T^n\psi) = (T^{-n}h, PT^n\psi)$$
$$= (\varphi, PT^n\psi) = (P\varphi,T^n\psi).$$

Therefore $P \in \mathscr{T}'_w \subset C_{\mathscr{D}}$.

(iii) This can be carried out as a consequence of Proposition 4.6 (taking into account that $\mathscr{T}$ dominates $\mathscr{A}$ and thus the left multiplication is continuous) or by a simple direct calculation.

(iv) By the same argument used in the proof of Proposition 5.2, $\mathscr{A}''_{\sigma\sigma} = \mathscr{A}''_{w\sigma}$ is a symmetric op*-algebra. Then $(\mathscr{A}''_{w\sigma})'_w = \mathscr{A}'_w$ and therefore $\mathscr{A}''_{w\sigma} = (\mathscr{A}''_{w\sigma})''_{w\sigma}$. The statement for $\mathscr{A}'_\sigma$ follows from the symmetry of $\mathscr{A}''_{\sigma\sigma}$.

Example: Let $\mathscr{D}$ be the Schwartz space $\mathscr{S}(\mathbb{R}^3)$ and consider the operators

$$H = \sum_{i=1}^{3} p_i^2 + q_i^2,$$
$$L_1 = q_2 p_3 - q_3 p_2, \quad L_2 = q_3 p_1 - q_1 p_3,$$
$$L_3 = q_1 p_2 - q_2 p_1$$

with

$$q_i f = x_i f \text{ and } p_i f = i\frac{\partial}{\partial x_i} f.$$

As is known $[H, L_i] = 0$. Thus H is in the center of the algebra $\mathscr{A}$ generated by $\{H, L_1, L_2, L_3\}$. Moreover, H is essentially self-adjoint in $\mathscr{S}(\mathbb{R}^3)$ and $\mathscr{S}(\mathbb{R}^3) = \mathscr{D}^\infty(\overline{H})$ (Ref. 5, Sec. 5, Example 2). Therefore, the algebra $\mathscr{A}$ generates an $\overline{SV}$*-algebra.

Remark 6.2: Proposition 6.1 applies immediately to the op*-algebra $\mathscr{T}$ generated by a self-adjoint operator T in $\mathscr{D}^\infty(T)$. In this case some further information can be obtained by the following proposition.

Proposition 6.3: Let T be a self-adjoint operator in $\mathscr{H}$ and $\mathscr{T}$ the op*-algebra generated by its restriction to $\mathscr{D}^\infty(T)$. We have (i) if $u(T)$ is a function of T, defined in the usual way by the functional calculus, and $D(u(T)) \supseteq \mathscr{D}^\infty(T)$ then $u(T) \upharpoonright \mathscr{D}^\infty(T) \in \mathscr{T}''_{\sigma\sigma}$ and thus it leaves $\mathscr{D}^\infty(T)$ invariant; (ii) if $\mathscr{H}$ is separable, $\mathscr{T}''_{\sigma\sigma}$ consists only of functions of T.

Proof: We need only to prove (ii). We saw before that if $A \in \mathscr{T}''_{\sigma\sigma}$, then $\overline{A}\eta\mathscr{T}''_{ww} = \{E(\lambda)\}''$. Hence, $\overline{A}$ bicommutes in the usual sense with $\overline{T}$. From the classical theorem (see, for instance, Ref. 16, n. 129), under the assumption of separable $\mathscr{H}$, we get the statement.

Proposition 6.4: Let $\mathscr{A}$ be a closed abelian standard op*-algebra on $\mathscr{D}$.

(i) $\mathscr{A}$ is self-adjoint and therefore both $\mathscr{A}'_\sigma$ and $\mathscr{A}''_{\sigma\sigma}$ are $\overline{V}$*-algebras.

(ii) Both $\mathscr{A}'_\sigma$ and $\mathscr{A}''_{\sigma\sigma}$ are symmetric and, therefore, standard.

(iii) $\mathscr{A}''_{w\sigma} = \mathscr{A}''_{\sigma\sigma}$.

(iv) Both $\mathscr{A}'_\sigma$ and $\mathscr{A}''_{\sigma\sigma}$ are $\overline{SV}$*-algebras.

Proof: (ii) Since $\mathscr{A}$ is standard, $\mathscr{A}''_{ww} \subseteq \mathscr{A}'_w \subseteq C_{\mathscr{D}}$ (Ref. 5, Theorem 7.1). By the same argument used in the proof of Proposition 5.2 we get the statement.

(iii) follows from the fact that $\mathscr{A}'_\sigma$ is symmetric.

(iv) is analogous to (iv) of Proposition 6.1.

ACKNOWLEDGMENTS

We thank Professor J.-P. Antoine and Dr. F. Mathot for interesting discussions.

One of us (C.T.) acknowledges gratefully the hospitality of the Institut de Physique Théorique de l'Université Catholique de Louvain.

He also acknowledges interesting discussions with Professor D. A. Dubin and Professor K. Schmüdgen.

Research was supported by MPI, Università di Palermo.

[1] R. Ascoli, G. Epifanio, and A. Restivo, Commun. Math. Phys. **18**, 291 (1970).
[2] R. Ascoli, G. Epifanio, and A. Restivo, Riv. Mat. Univ. Parma **3**, 21 (1974).
[3] G. Lassner, Rep. Math. Phys. **3**, 279 (1972).
[4] D. Arnal and J. P. Jurzak, J. Funct. Anal. **24**, 397 (1977).
[5] R. T. Powers, Commun. Math. Phys. **21**, 397 (1971).
[6] S. Gudder and W. Scruggs, Pac. J. Math. **70**, 369 (1977).
[7] G. Epifanio and C. Trapani, J. Math. Phys. **22**, 974 (1981).
[8] J.-P. Antoine and F. Mathot, Ann. Inst. H. Poincaré **XXXVII**, 29 (1982).
[9] H. Araki and J. P. Jurzak, Publ. Res. Inst. Math. Sci. **18**, 1013 (1982).
[10] P. G. Dixon, Proc. London Math. Soc. **23**, 53 (1971).
[11] A. Inoue, Pac. J. Math. **65**, 77 (1976).
[12] A. Inoue, Proc. Am. Math. Soc. **69**, 97 (1978).
[13] J. P. Antoine, G. Epifanio, and C. Trapani, Helv. Phys. Acta **56**, 1175 (1983).
[14] R. Haag and B. Schroer, J. Math. Phys. **3**, 248 (1962).
[15] G. Epifanio and C. Trapani, J. Math. Phys. **23**, 39 (1982).
[16] F. Riesz and B. Sz. Nagy, *Leçons d'analyse fonctionnelle* (Gauthier–Villars, Paris, 1965).

On the center-of-mass motion of geometrically confined classical particles

C. Dullemond
Institute for Theoretical Physics, University of Nijmegen, Nijmegen, The Netherlands

(Received 8 June 1983; accepted for publication 6 April 1984)

By means of group-theoretical methods based on O(3,2) a description of center-of-mass motion is given of a set of harmonically oscillating classical particles which can attain relativistic velocities. The limitation $|\mathbf{v}| < c$ on the velocity leads to a limitation $|\mathbf{r}| < R$ on the amplitude, where R is related to the universal oscillator frequency ω by $R = c\omega^{-1}$. It turns out that the center-of-mass carries out a harmonic oscillation with the same frequency ω and the same limitations, and that conditions can be formulated for the set of particles to be in its "rest system." The method can be applied to hadrons considered as bags containing harmonically oscillating classical quarks.

PACS numbers: 03.20. + i,02.20. + b,03.30 + p,12.35.Kw

I. INTRODUCTION

Recently it has become clear from analyses of the hadron spectrum that the relativistic harmonic oscillator as it is defined by the anti-De Sitter group O(3,2) may play a role in its description.[1,2] The basic idea is that a hadron can be visualized, in zeroth-order approximation, as a spherical bag in which noninteracting quarks carry out harmonic oscillations of a universal frequency ω equal to c times the inverse radius of the bag. A Hamiltonian dynamics for classical particles can be formulated which not only leads to the ordinary equations of motion

$$\ddot{\mathbf{r}} = -\omega^2\mathbf{r}, \tag{1.1}$$

but has all limitations built in, namely

$$|\mathbf{r}| < R \quad \text{(confinement)}, \tag{1.2}$$

and

$$|\dot{\mathbf{r}}| < c \quad \text{(relativistic requirement)}. \tag{1.3}$$

(We shall henceforth take $c = 1$.) The dynamics can be seen as a consequence of the O(3,2) invariance of an effective metric, with respect to which quarks describe timelike geodesics.[3] This is the reason why the confinement is called geometrical. Since there apparently is a preferred time-like linear trajectory representing the uniform motion of the bag as a whole (in this case the line $\mathbf{r} = 0$), Poincaré invariance is broken down to a mere O(3) $\otimes T_0$, where O(3) represents space rotations and reflections and T_0 time translations and reflections with respect to the preferred trajectory.

Periodicity comes in if T_0 is considered as a covering of O(2). A connection with the anti-De Sitter group can be made by constructing this in a way as to make the above O(3) $\otimes$ O(2) a subgroup of O(3,2).

The Lorentz-group O(3,1) as a subgroup of O(3,2) is a stability subgroup, i.e., it leaves a particular space-time point inside the bag invariant, but otherwise transforms a set of confined harmonic oscillations into another set of confined harmonic oscillations. The point transformations corresponding to this O(3,1) are in general curvilinear in $\mathbf{r}$ and t since they must also leave the sphere $|\mathbf{r}| = R$ invariant. Ordinary Lorentz transformations, linear in $\mathbf{r}$ and t, are therefore quite different. When the point of stability is $(\mathbf{0}, t_0)$, then a Lorentz boost belonging to the above O(3,1) corresponds to an ordinary linear Lorentz boost only when $\mathbf{r} \approx \mathbf{0}$ and $t \approx t_0$. What happens is, that when a boost is given to a particle at rest in the origin, it starts out to oscillate harmonically.

Of practical importance is the question of how to define a rest system for harmonically oscillating particles with possibly different rest masses. If a meson is considered as a bag filled with a quark and an antiquark, then these particles carry six momentum degrees of freedom. In reality there should only be three, while the three remaining degrees of freedom should be associated with the momentum of the meson as a whole. The reduction to three internal degrees of freedom can be achieved by demanding the two quarks to be in their rest system, if such a system can be defined. For an arbitrary system of interacting particles, like the MIT-bag,[4] this is not satisfactorily possible. The harmonic oscillator bag is a favorable exception. Associated with it is the possibility of defining a "center-of-mass." In this article it will be shown that a center-of-mass can be uniquely and meaningfully defined in two ways, one of them O(3,2) invariant, the other not. Both definitions share a number of properties.

(a) Each center-of-mass (c.m.) depends on the individual orbits, rest masses, and spins of the participating point particles.

(b) Each c.m. is independent of how one groups particles together to form subunits and is also independent of the order of sequence of the particles.

(c) In the nonrelativistic limit of slow motions around the center of the sphere both correspond to the nonrelativistic definition of center-of-mass, in which case the individual spins are unimportant.

There are three cases where both definitions lead to the same result. One is the above mentioned nonrelativistic limit, the second is that of a set of particles with a net internal angular momentum equal to zero, and the third is that of no center-of-mass motion at all. Since one definition of c.m. is very similar to the nonrelativistic definition (particle masses replaced by particle energies) but depends on the coordinate system, I shall call this the "canonical center-of-mass" (c.c.m.). It can simply be constructed from components of a second-order antisymmetric tensor which plays a fundamental role in this article. It depends on the spin only implicitly via the energy.

The second definition of center-of-mass is independent of the coordinate system, but depends explicitly on the indi-

 0022-2488/84/092638-10$02.50

vidual spins. The functional relationships are intricate. I shall call this the "system independent center-of-mass" (s.i.c.m.). Both definitions have their advantages and disadvantages. It must be considered a lucky circumstance that the definition of a coordinate system with respect to which a system of particles is "at rest" is unique save for a space rotation around the center of the sphere, and a time translation. Thus there is a unique criterion for a set of oscillating classical "quarks" of different mass to be in its rest system. This is also the case in the quantum version of the model.

Besides centers-of-mass, a rest mass and spin of a conglomerate of particles can be defined. Under the condition that the magnitudes of the individual spins of the participating particles be smaller than, or equal to, ω^{-1} times the individual rest masses, the rest mass of the conglomerate is larger than or equal to the sum of the individual rest masses.

Additional O(3,2) breaking quark interactions can be required not to disturb the motion of the center-of-mass. Space and time reflections will not be considered and a discussion of SO(3,2) rather than O(3,2) will be given.

The article is built up as follows: In Sec. II a review is given of the relativistic harmonic oscillator and its relation to SO(3,2). A bivector T is introduced which characterizes an oscillating particle or set of particles. Section III is devoted to the parametrization of the restricted anti-De Sitter group as preparation for the study of the properties of T (Sec. IV). The physical interpretation of T and its connection with the center-of-mass is given in Sec. V. Two examples are worked out in Sec. VI.

II. THE RELATIVISTIC HARMONIC OSCILLATOR

For $r = |\mathbf{r}| < R$ consider the following point transformation $x^\mu \to x'^\mu$, with $\mathbf{r} = (x^1, x^2, x^3)$ and $t = x^4$:

$$\begin{aligned} &Z^N = \mathcal{T}(x^\mu) \quad N = 1,\ldots,5, \quad \mu = 1,\ldots,4, \\ &Z'^N = \Lambda^N{}_M Z^M \quad \text{(automatic summation)}, \\ &Z'^N = \mathcal{T}(x'^\mu). \end{aligned} \tag{2.1}$$

Here $Z^N = \mathcal{T}(x^\mu)$ stands for the transformation

$$\begin{aligned} &Z^k = \gamma x^k \quad (k = 1,2,3), \\ &Z^4 = \gamma R \sin(x^4/R), \\ &Z^5 = \gamma R \cos(x^4/R), \end{aligned} \tag{2.2}$$

with $\gamma = R/\sqrt{R^2 - r^2}$, while $\Lambda^N{}_M$ stands for a linear transformation with the properties

$$\det \Lambda^N{}_M = 1, \tag{2.3}$$

and

$$\Lambda^N{}_M \Lambda^Q{}_P H_{NQ} = H_{MP},$$

where

$$H_{NQ} = \text{diag}(-1,-1,-1,+1,+1)$$

is a metric in 5-dimensional flat space. (An extra restriction is needed to exclude "time inversions"; see Sec. III.)

The transformations $\Lambda^N{}_M$ form the group SO(3,2) of hyperbolic rotations in five-space. Thus, the transformations $x^\mu \to x'^\mu$ are isomorphic with SO(3,2), or the anti-De Sitter group. The Z^M as obtained from (2.2) satisfy

$$Z_M Z^M = Z^M Z^N H_{MN} = R^2, \tag{2.4}$$

i.e., they lie on a hyperboloid in Z space.

As is shown in Ref. 3, confined harmonic oscillations of frequency $\omega = R^{-1}$ in terms of x^μ transform into confined harmonic oscillations of the same frequency in terms of x'^μ. Moreover, any confined harmonic oscillation can be transformed to rest. It has also been shown in Ref. 3, that when the points x^μ belong to a particular oscillation, then the associated points Z^M not only satisfy (2.4), but also lie in a two-plane through the origin of Z-space. This observation enables one to represent the oscillatory motion by a bivector $T^{MN} (= -T^{NM})$ in Z-space. Since the plane is equally well represented by αT^{MN} with α an arbitrary scalar, in order to fix the magnitude of T^{MN} it should harbor one other characteristic of the motion. This turns out to be the rest mass of the oscillating particle. By doing so T^{MN} contains dynamical information. At this point one can make use of the fact that bivectors can be added. Let $T^{MN}_{(i)} (i = 1,\ldots,k)$ be the bivectors associated with k oscillating particles with rest masses m_i. Then

$$T^{MN} = \sum_{i=1}^{k} T^{MN}_{(i)} \tag{2.5}$$

is a bivector representing certain characteristics of the collective motion of the k particles. However, while the rank of each of the $T^{MN}_{(i)}$ is 2, the rank of T^{MN} is a general 4, which expresses the fact that the effective particle whose motion is represented by T^{MN} not only has a well-defined path and rest mass, but carries also intrinsic angular momentum. This "well-defined path" is to be interpreted as center-of-mass motion of either of the two types. The intrinsic angular momentum is the spin.

As a corollary to this, if the point particles themselves carry spin (as is the case with quarks), then this dynamical information must also be incorporated in the bivector $T^{MN}_{(i)}$ which now becomes of rank 4 instead of 2. Thus we see that the center-of-mass motion is influenced by the spins of the contributing point particles.

How do we obtain T^{MN} for a scalar particle? First find an SO(3,2) transformation which transforms the particle to rest in the center of the sphere. Except for special cases to be considered separately, this is always possible. Let this transformation be represented in Z-space by the 5×5 matrix Λ^P_Q. The particle at rest is represented by

$$T^{MN}_0 = m(\delta^{M5}\delta^{N4} - \delta^{M4}\delta^{N5}) \quad (m > 0), \tag{2.6}$$

where m is the rest mass of the particle and δ is the Kronecker symbol. Note that T^{MN}_0 is invariant under those transformations of SO(3,2) which are orthochronous and which leave the particle at rest. [We shall call a proper orthochronous O(3,2) transformation "restricted." Only these transformations are continuously connected with the identity.] The representing bivector for the oscillating particle now becomes

$$T^{MN} = (\Lambda^{-1})^M{}_P (\Lambda^{-1})^N{}_Q T^{PQ}_0. \tag{2.7}$$

As already remarked, expression (2.6) represents a spinless particle at rest. The world line of such a particle is presented by the four-vector

$$x^\mu = (\mathbf{0}, \tau), \tag{2.8}$$

which corresponds, with (2.2), to

$$Z_0^M = R(\delta^{M4}\sin\omega\tau + \delta^{M5}\cos\omega\tau). \tag{2.9}$$

Thus we find

$$T_0^{MN} = m\omega\left(Z_0^M\frac{dZ_0^N}{d\tau} - Z_0^N\frac{dZ_0^M}{d\tau}\right), \tag{2.10}$$

and, if

$$Z^M = (\Lambda^{-1})^M{}_P Z_0^P, \tag{2.11}$$

we have from (2.7)

$$T^{MN} = m\omega\left(Z^M\frac{dZ^N}{d\tau} - Z^N\frac{dZ^M}{d\tau}\right). \tag{2.12}$$

For small values of $Z^k (k = 1,2,3)$ we have $Z^k \approx x^k$ and $\tau \approx t$ and we find

$$T^{kl} \approx m\omega\left(x^k\frac{dx^l}{dt} - x^l\frac{dx^k}{dt}\right), \tag{2.13}$$

which is just ω times the angular momentum with respect to the origin. We also have

$$T^{k4} = -T^{4k} \approx m\omega x^k\cos\omega t - m\frac{dx^k}{dt}\sin\omega t, \tag{2.14}$$

and

$$T^{k5} = -T^{5k} \approx -m\omega x^k\sin\omega t - m\frac{dx^k}{dt}\cos\omega t. \tag{2.15}$$

From this, x^k can be retrieved:

$$x^k \approx (1/m\omega)(T^{k4}\cos\omega t - T^{k5}\sin\omega t). \tag{2.16}$$

Finally

$$T^{45} = -T^{54} = m\omega\left(Z^4\frac{dZ^5}{d\tau} - Z^5\frac{dZ^4}{d\tau}\right) \approx -m. \tag{2.17}$$

The transformations leaving the particle at rest form a subgroup SO(3) $\otimes$ SO(2).Clearly, the rank of T_0^{MN} is 2 and this rank is preserved under transformations.

A spinning particle at rest in the center of the sphere can be represented by

$$T_0^{MN} = m(\delta^{M5}\delta^{N4} - \delta^{M4}\delta^{N5}) + \omega\sum_{a,b,c=1}^{3} s_0^a\epsilon^{abc}\delta^{Mb}\delta^{NC} \quad (m>0), \tag{2.18}$$

where s_0^a represents the spin of the particle and ϵ^{abc} is the Levi–Civita symbol in three dimensions ($\epsilon^{123} = +1$). In Sec. IV it is proven that in order to obtain a consistent physical picture m must satisfy the inequality

$$m \geqslant \omega|\mathbf{s}_0|. \tag{2.19}$$

For $\mathbf{s}_0 \neq 0$ the rank of T_0^{MN} as given by (2.18) is 4 and this rank is preserved under transformation. The cases $m > \omega|\mathbf{s}_0|$ and $m = \omega|\mathbf{s}_0|$ are quite distinct.

There is a special class of particles (whose confined orbits touch or are embedded in the sphere $r = R$) which cannot be transformed to rest. They possess a well-defined T^{MN} which cannot be transformed into the form (2.18). This category must be treated separately.

In order to obtain insight in the requirements for consistency it is necessary to recall certain properties of the group SO(3,2) which is done in the next section.

III. PARAMETRIZATION OF THE RESTRICTED ANTI-DE SITTER GROUP[5]

Let Λ be a real 5×5 matrix satisfying

$$\Lambda H\Lambda^T = H, \tag{3.1}$$

where

$$H = \mathrm{diag}(-1,-1,-1,+1,+1) = \begin{Vmatrix} -I_3 & 0 \\ 0 & I_2 \end{Vmatrix}, \tag{3.2}$$

with I_n being the $n\times n$ unit matrix and T meaning transposition. The set of matrices Λ form the group O(3,2) or anti-De Sitter group. Since

$$H\Lambda H\Lambda^T = I_5 = H\Lambda^T H\Lambda, \tag{3.3}$$

it follows that if Λ is an element of the group, also Λ^T is an element of the group.

It is always possible to find an orthogonal matrix R and a symmetric positive definite matrix S, both real, for which

$$\Lambda = RS. \tag{3.4}$$

From this we find that

$$\Lambda^T\Lambda = S^2 \tag{3.5}$$

is an element of the group.

Let $\overline{S}$ be an arbitrary symmetric and positive definite element of O(3,2). Then it can be diagonalized by a real orthogonal matrix $\overline{R}$:

$$\overline{S} = \overline{R}\,\overline{S}_d\overline{R}^T, \tag{3.6}$$

where $\overline{S}_d$ satisfies

$$\overline{S}_d(\overline{R}^T H\overline{R})\overline{S}_d = \overline{R}^T H\overline{R}, \tag{3.7}$$

and has positive definite diagonal elements.

Since $\overline{R}^T H\overline{R}$ is nonsingular, $\overline{R}$ can be chosen such that

$$\overline{S}_d = \mathrm{diag}(1,\lambda_1,\lambda_2,\lambda_1^{-1},\lambda_2^{-1}), \tag{3.8}$$

with $\lambda_1, \lambda_2 > 0$. If now $\overline{R}$ is kept fixed, the parameters λ_1 and λ_2 can be varied arbitrarily to yield new matrices $\overline{S}$ as elements of O(3,2). In particular

$$\overline{S}_d^{\mu} = \mathrm{diag}(1,\lambda_1^{\mu},\lambda_2^{\mu},\lambda_1^{-\mu},\lambda_2^{-\mu}), \tag{3.9}$$

with arbitrarily real μ is of the form (3.8) and leads to $\overline{S}^{\mu}$ as element of O(3,2). From this we conclude the following.

(a) Any symmetric, positive definite element of O(3,2) is continuously connected with the identity.

(b) Any such element can be written in the form

$$\overline{S} = e^{A_s}, \tag{3.10}$$

where A_s is real and symmetric and satisfies

$$A_s H + HA_s = 0. \tag{3.11}$$

(c) From (3.5) one can solve for S as element of the group.

From (3.4) we now solve for R as element of the group. For R to satisfy

$$RHR^T = H \text{ and } RR^T = I_5, \tag{3.12}$$

it must have the general form

$$R = \begin{Vmatrix} R_3 & 0 \\ 0 & R_2 \end{Vmatrix}, \tag{3.13}$$

where R_n is an $n\times n$ orthogonal matrix. Both det R_3 and

$\det R_2$ are conserved under continuous change of Λ. If $\det R_3 = \det R_2 = 1$, then R is continuously connected with the identity and so is Λ. Then also R can be written as an exponential form

$$R = e^{A_a}, \tag{3.14}$$

where A_a is antisymmetric and real, and satisfies

$$A_a H - H A_a = 0. \tag{3.15}$$

From (3.11) and (3.15) we find the general forms for A_s and A_a:

$$A_s = \begin{Vmatrix} 0 & A_{32} \\ A_{32}^T & 0 \end{Vmatrix}, \tag{3.16}$$

$$A_a = \begin{Vmatrix} A_{33} & 0 \\ 0 & A_{22} \end{Vmatrix}, \quad A_{33} = -A_{33}^T, \quad A_{22} = -A_{22}^T, \tag{3.17}$$

where A_{mn} stands for an arbitrary real $m \times n$ matrix. We conclude that any O(3,2) matrix Λ which is continuously connected with the identity can be written in the form

$$\Lambda = e^{A_a} \cdot e^{A_s}. \tag{3.18}$$

We shall call the group of these matrices Λ the restricted O(3,2) group. The special group SO(3,2) contains besides the restricted group also those matrices Λ for which $\det R_3 = \det R_2 = -1$.

It is interesting to observe that the subset of matrices Λ for which

$$\Lambda = e^A, \tag{3.19}$$

with A a real matrix satisfying

$$AH + HA^T = 0, \tag{3.20}$$

do *not* form a group. As a counterexample we consider the matrix[6]

$$U = \begin{Vmatrix} -1 & \Theta & 0 & 0 & -\Theta \\ -\Theta & -1 & 0 & \Theta & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \Theta & 0 & -1 & -\Theta \\ -\Theta & 0 & 0 & \Theta & -1 \end{Vmatrix}, \tag{3.21}$$

with Θ arbitrary real. This matrix satisfies (3.1) and is continuously connected with the identity. For $\Theta \neq 0$ it cannot be written in the form (3.19) with A satisfying (3.20).

We shall call the elements of SO(3,2) "proper" and those elements of O(3,2) for which $\det R_2 = +1$ "orthochronous" in analogy with similar expressions used for elements of the Lorentz group O(3,1). The restricted group contains those elements which are both proper and orthochronous. In analogy with the Lorentz group we shall call an element of the form e^{A_s} a "generalized boost" and an element of the form e^{A_a} a "generalized rotation" which is in fact an ordinary three-rotation combined with a time translation. One needs six parameters to specify a generalized boost and four to specify a generalized rotation. When no confusion can arise, a generalized rotation will just be called a rotation and a generalized boost will be called a boost.

Note that the product of two boosts is in general not a boost, but a boost combined with a rotation. This rotation is the *generalized Thomas precession*.

Consider a boost S and its transform S' by a rotation R:

$$S' = RSR^T. \tag{3.22}$$

Then

$$A'_S = RA_s R^T, \tag{3.23}$$

and

$$A'_{32} = R_3 A_{32} R_2^T. \tag{3.24}$$

One can always choose R_3 and R_2 such that

$$A'_{32} = \begin{Vmatrix} a & 0 \\ 0 & b \\ 0 & 0 \end{Vmatrix}, \quad a \geqslant |b|. \tag{3.25}$$

Then S' is a special boost, comparable with a special Lorentz transformation. Apparently there are two parameters to specify a special boost, which we indicate with S^{sp}. Then

$$S^{\text{sp}} = e^{A_s^{\text{sp}}}, \tag{3.26}$$

where

$$A_s^{\text{sp}} = \begin{Vmatrix} 0 & A'_{32} \\ A'^T_{32} & 0 \end{Vmatrix}. \tag{3.27}$$

Then, with (3.25), we can find a closed expression for S^{sp}:

$$S^{\text{sp}} = \begin{Vmatrix} \cosh a & 0 & 0 & \sinh a & 0 \\ 0 & \cosh b & 0 & 0 & \sinh b \\ 0 & 0 & 1 & 0 & 0 \\ \sinh a & 0 & 0 & \cosh a & 0 \\ 0 & \sinh b & 0 & 0 & \cosh b \end{Vmatrix}, \tag{3.28}$$

with $a \geqslant |b|$. Every restricted Λ satisfying (3.1) can now be written as follows:

$$\Lambda = RS^{\text{sp}}R', \tag{3.29}$$

where R and R' are generalized rotations.

IV. PROPERTIES OF T^{MN}

With the information obtained in Sec. III we can study the properties of the "standard form" (2.18) of a bivector under restricted O(3,2) transformations. First we observe that the rotation R' occurring in (3.29) rotates the vector $\mathbf{s}_0$ but does not change the form of the T_0^{MN}. Also m is not changed, so R' can be left out of our discussion.

Next we apply S^{sp} as given by (3.28). We find, if

$$T = S^{\text{sp}} T_0 S^{\text{sp}}, \tag{4.1}$$

that

$$T^{54} = m \cosh a \cosh b - \omega s_{03} \sinh a \sinh b. \tag{4.2}$$

We shall postulate that T^{54} represents the *total energy* of the oscillating particle. Clearly, when $\mathbf{s}_0 = 0$, the energy of a moving particle is larger than that of a particle at rest.

Let $\omega|\mathbf{s}_0| < m$. Then

$$\begin{aligned} T^{54} &\geqslant m \cosh a \cosh b - \omega|\mathbf{s}_0| \sinh a \sinh|b| \\ &= m \cosh(a - |b|) + (m - \omega|\mathbf{s}_0|) \sinh a \sinh|b|. \end{aligned} \tag{4.3}$$

This increases indefinitely with increasing a. When $\omega|\mathbf{s}_0| = \omega s_{03} = m$ we have

$$T^{54} = m\cosh(a-b). \tag{4.4}$$

This is equal to m when $a-b=0$ and we see that a particle at rest can have the same energy as a moving particle, when an exceptional boost is carried out.

When $\omega|\mathbf{s}_0| = \omega s_{03} > m$ we find

$$T^{54} = m\cosh(a-|b|) - (\omega|\mathbf{s}_0| - m)\sinh a \sinh|b|. \tag{4.5}$$

If $a = |b|$ one can always find a such that $T^{54} < 0$, and indeed, it can be made arbitrarily small. In that case there is no lower limit on the energy and it follows that $\omega|\mathbf{s}_0| > m$ cannot represent a physical situation.

The transformation R as occurring in (3.29) does not change T^{54} and can also be left out of consideration. Suppose we add a number of $T_{(i)}$ together, each of which is obtained from a $T_{(i)0}$ for which $\omega|\mathbf{s}_{0(i)}| \leqslant m_{(i)}$, then for the sum $T = \Sigma_i T_{(i)}$ there exists a positive lower limit for the value $T^{54} = \Sigma_i T^{54}{}_{(i)}$ when an arbitrary restricted transformation is carried out. So, if a transformation exists which transforms T into the standard form (2.18), then it follows that $\omega|\mathbf{s}_0| \leqslant m$ for this sum. The equal sign will be valid if an exceptional boost exists which does not change the value of T^{54}. Such a boost may not cause any of the individual $T^{54}_{(i)}$ to grow indefinitely and it follows that every $T_{(i)}$ must satisfy for its standard form the relation $\omega|\mathbf{s}_{0(i)}| = m_{(i)}$. This is not a sufficient condition. In general, even if for all i this condition is satisfied, we will have for the sum $\omega|\mathbf{s}_0| < m$.

We shall first consider the case $\omega|\mathbf{s}_{0(i)}| < m_{(i)}$ for all i.

Let us transform a given standard form T_0 by means of a restricted transformation Λ into the general form T:

$$T = \Lambda T_0 \Lambda^{T}. \tag{4.6}$$

We write

$$T = \begin{Vmatrix} \omega J & V \\ -V^{T} & M \end{Vmatrix}, \tag{4.7}$$

where

$$J = \begin{Vmatrix} 0 & j_3 & -j_2 \\ -j_3 & 0 & j_1 \\ j_2 & -j_1 & 0 \end{Vmatrix}, \tag{4.8}$$

$$V = \begin{Vmatrix} q_1 & p_1 \\ q_2 & p_2 \\ q_3 & p_3 \end{Vmatrix}, \tag{4.9}$$

and

$$M = \begin{Vmatrix} 0 & -\mu \\ \mu & 0 \end{Vmatrix}, \quad \mu > 0. \tag{4.10}$$

Furthermore we write

$$\begin{aligned} \mathbf{q} &= (q_1,q_2,q_3), \\ \mathbf{p} &= (p_1,p_2,p_3), \\ \mathbf{j} &= (j_1,j_2,j_3). \end{aligned} \tag{4.11}$$

Let Λ be given by (3.29),

$$\Lambda = RS^{\mathrm{sp}}R',$$

and let us define $\overline{\Lambda}$ to be

$$\overline{\Lambda} = R(S^{\mathrm{sp}})^{-1}R'. \tag{4.12}$$

Moreover, let

$$\overline{T} = \overline{\Lambda} T_0 \overline{\Lambda}^{T}. \tag{4.13}$$

Then

$$\mathscr{T} = \tfrac{1}{2}(T + \overline{T}) \tag{4.14}$$

has the standard form

$$\mathscr{T} = \begin{Vmatrix} \omega J & 0 \\ 0 & M \end{Vmatrix}. \tag{4.15}$$

Clearly, since $\mathscr{T}$ is a sum of tensors for which $\omega|\mathbf{s}_0| < m$ we now find

$$\omega|\mathbf{j}| < \mu, \tag{4.16}$$

and this is a property of T as given by (4.7).

We now prove that an arbitrary sum of tensors $T_{(i)}$ for which $\omega|\mathbf{s}_{0(i)}| < m_{(i)}$ can be transformed into standard form. We may limit ourselves to a sum of two tensors, one of which is in the standard form (2.18), the other being infinitesimal. The general form of such a sum is

$$T = \begin{Vmatrix} \omega J & \Delta V \\ -\Delta V^{T} & M \end{Vmatrix}, \tag{4.17}$$

with J and M given by (4.8) and (4.10) and

$$\Delta V = \begin{Vmatrix} \Delta q_1 & \Delta p_1 \\ \Delta q_2 & \Delta p_2 \\ \Delta q_3 & \Delta p_3 \end{Vmatrix}. \tag{4.18}$$

Here we have also $\omega|\mathbf{j}| < \mu$. In analogy with (4.11) we define

$$\begin{aligned} \Delta\mathbf{q} &= (\Delta q_1,\Delta q_2,\Delta q_3), \\ \Delta\mathbf{p} &= (\Delta p_1,\Delta p_2,\Delta p_3). \end{aligned} \tag{4.19}$$

In order to transform T into the standard form we apply an infinitesimal transformation

$$\Lambda = \begin{Vmatrix} I_3 & \Delta\mathbf{f}^{T} & \Delta\mathbf{g}^{T} \\ \Delta\mathbf{f} & 1 & 0 \\ \Delta\mathbf{g} & 0 & 1 \end{Vmatrix}, \tag{4.20}$$

where

$$\begin{aligned} \Delta\mathbf{f} &= (\Delta f_1,\Delta f_2,\Delta f_3), \\ \Delta\mathbf{g} &= (\Delta g_1,\Delta g_2,\Delta g_3). \end{aligned} \tag{4.21}$$

Then

$$\Delta T = \Lambda T \Lambda - T = \begin{Vmatrix} 0 & \omega(\Delta\mathbf{f}\times\mathbf{j})^{T} + \mu\Delta\mathbf{g}^{T} & \omega(\Delta\mathbf{g}\times\mathbf{j})^{T} + \mu\Delta\mathbf{f}^{T} \\ -\omega(\Delta\mathbf{f}\times\mathbf{j}) - \mu\Delta\mathbf{g} & 0 & 0 \\ -\omega(\Delta\mathbf{g}\times\mathbf{j}) + \mu\Delta\mathbf{f} & 0 & 0 \end{Vmatrix}. \tag{4.22}$$

If now

$$\Delta \mathbf{q} = -\omega(\Delta \mathbf{f}\times\mathbf{j}) - \mu\Delta\mathbf{g},$$

and (4.23)

$$\Delta \mathbf{p} = -\omega(\Delta\mathbf{g}\times\mathbf{j}) + \mu\Delta\mathbf{f},$$

then Δ transforms T into the standard form. By eliminating $\Delta\mathbf{f}$ we find

$$\omega^2\mathbf{j}(\mathbf{j}\cdot\Delta\mathbf{g}) + (\mu^2 - \omega^2 j^2)\Delta\mathbf{g} = -\mu\Delta\mathbf{q} + \omega\mathbf{j}\times\Delta\mathbf{p}. \quad (4.24)$$

This equation must be solved for $\Delta\mathbf{g}$. If we write Eq. (4.24) in the form

$$\sum_j A_{ij}\Delta g_j = \Delta b_i, \quad (4.25)$$

then

$$A_{ij} = \omega^2 j_i j_j + (\mu^2 - \omega^2 j^2)\delta_{ij}. \quad (4.26)$$

Since $\mu^2 > \omega^2 j^2$ we find $\det A \neq 0$ and the equations can be solved. This proves the statement that if $T_{(i)}$ can be transformed into the standard form $T_{(i)0}$ with $\omega|\mathbf{s}_{0(i)}| < m_{(i)}$ for all i, then $T = \Sigma_i T_{(i)}$ can be transformed into the standard form T_0 with $\omega|\mathbf{s}_0| < m$. Since $T^{54}_{(i)} \geqslant m_{(i)}$, and $T^{54}_{(i)} = m_i$ only when T has the standard form, we find that

$$m \geqslant \sum_i m_{(i)}, \quad (4.27)$$

where the equal sign is valid only when all $T_{(i)}$ can be put into the standard form simultaneously.

When $\mathbf{j}$ is pointing in three-direction and $\mu = \omega|\mathbf{j}|$, then for T^{MN} to be transformable into the standard form (2.18) with $m = \omega|\mathbf{s}_0|$ it should have the form

$$T^{MN} = \left|\left| \begin{array}{ccccc} 0 & \omega j & 0 & \phi & \zeta \\ -\omega j & 0 & 0 & \zeta & -\phi \\ 0 & 0 & 0 & 0 & 0 \\ -\phi & -\zeta & 0 & 0 & -\mu \\ -\zeta & \phi & 0 & \mu & 0 \end{array} \right|\right|,$$

$$\phi^2 + \zeta^2 < \mu^2, \quad j = |\mathbf{j}|, \quad \mu = \omega|\mathbf{j}|, \quad (4.28)$$

This can be proven straightforwardly by demanding that no transformation can be found which makes T^{54} negative.

Let us next turn to the massless case. Starting from a given standard form (2.18) with $\omega|\mathbf{s}_0| \leqslant m$ we apply a boost and multiply the result with a positive number. Then the limit of an infinitely large boost combined with an infinitely small number is taken. The result is a finite tensor T which can always be transformed by a *finite* restricted O(3,2) transformation into either

$$\tilde{T}^{MN}_0 = \delta^{N1}\delta^{M5} + \delta^{N4}\delta^{M5} - \delta^{M1}\delta^{N5} - \delta^{M4}\delta^{N5}, \quad (4.29)$$

or

$$\tilde{T}^{MN}_0 = (\delta^{M1} + \delta^{M5})(\delta^{N2} + \delta^{N4}) - (\delta^{N1} + \delta^{N5})(\delta^{M2} + \delta^{M4}), \quad (4.30)$$

both forms being not transferable into each other, nor into the standard form (2.18). There exist finite transformations which result in the multiplication of $\tilde{T}^{MN}_0$ by any positive number. An extention of the theorem on sums of tensors will be given without proof.

Let $T_{(i)}$ be either transformable into (2.18) for $m_{(i)} \geqslant \omega|\mathbf{s}_{0(i)}|$, or into (4.29) or (4.30), then $\Sigma_i T_{(i)}$ is always transformable into one of these three forms, and almost always into the form (2.18) with $m > \omega|\mathbf{s}_0|$ [i.e., the forms (4.29) and (4.30) occur exceptionally]. There is a clear analogy with the effective mass of two photons, which is almost always nonzero.

If $m_{(i)}$ is put equal to zero for all $T_{(i)}$ which can be transformed into (4.29) or (4.30), then the statement (4.27) is also valid in this case.

Whether a given T can be transformed into one of the standard forms (2.18), (4.29), or (4.30) is not immediately clear. There are a number of necessary conditions to be satisfied. First of all, the quantity

$$T^2 = \tfrac{1}{2}T_{MN}T^{MN} = \mu^2 + \omega^2\mathbf{j}^2 - \mathbf{q}^2 - \mathbf{p}^2, \quad (4.31)$$

which is equal to $m^2 + \omega^2\mathbf{s}_0^2$ when T can be transformed into the form (2.18), is an invariant. We then have

$$T^2 > 0. \quad (4.32)$$

When either (4.29) or (4.30) is the standard form we have

$$T^2 = 0, \quad (4.33)$$

so $T^2 \geqslant 0$ is one of the necessary conditions. A second, independent invariant can be obtained from the five-vector S_N defined as follows:

$$S_N = (1/8\omega)T^{AB}T^{CD}\epsilon_{ABCDN} = (-\mathbf{S}, S^4, S^5), \quad (4.34)$$

with

$$\mathbf{S} = \mu\mathbf{j} + (\mathbf{q}\times\mathbf{p})_\omega, \quad S^4 = -\mathbf{p}\cdot\mathbf{j}, \quad S^5 = \mathbf{q}\cdot\mathbf{j}, \quad (4.35)$$

and ϵ_{ABCDN} being the five-dimensional Levi–Civita symbol with $\epsilon_{12345} = 1$. This invariant is

$$S^2 = -S_N S^N = \mu^2(\mathbf{j} + \mathbf{q}\times\mathbf{p}/\omega\mu)^2 - [(\mathbf{q}\cdot\mathbf{j})^2 + (\mathbf{p}\cdot\mathbf{j})^2]. \quad (4.36)$$

If T has the standard form (2.18) we have

$$S^2 = m^2\mathbf{s}_0^2 \geqslant 0, \quad (4.37)$$

while for the standard forms (4.29) and (4.30) we find

$$S^2 = 0. \quad (4.38)$$

Sharper inequalities can be obtained by considering two tensors T_1^{MN} and T_2^{MN}, both of which being transformable into the form (2.18) with $\omega|\mathbf{s}_0| < m$. Suppose we have

$$T_1^{MN} = \left|\left| \begin{array}{cc} \omega J & V \\ -V^T & M \end{array} \right|\right|, \quad (4.39)$$

with J, V, and M defined by (4.8)–(4.10), and T_2^{MN} having the standard form (2.18). Then

$$T_1\cdot T_2 = \tfrac{1}{2}T_{1MN}T_2^{MN} = \mu m + \omega^2\mathbf{j}\cdot\mathbf{s}_0. \quad (4.40)$$

Because $\omega|\mathbf{s}_0| < m$ and $\omega|\mathbf{j}| < \mu$ we now have

$$T_1\cdot T_2 > 0. \quad (4.41)$$

Since $T_1\cdot T_2$ is an invariant, this is a general property valid also when T_2 is chosen otherwise. Now, let T_1 be given by (4.39) and T_2 by

$$T_2^{MN} = \left|\left| \begin{array}{cc} 0 & V' \\ -V'^T & M \end{array} \right|\right|, \quad (4.42)$$

with either $V' = (\mathbf{q}'^T, \mathbf{0}^T)$ or $V' = (\mathbf{0}^T, \mathbf{p}'^T)$. This tensor T_2 can

be obtained by applying a boost of the form (3.28) with $b = 0$ to expression (2.6) and then applying a rotation. From this construction it follows that there are no restrictions on the directions of $\mathbf{q}'$ and $\mathbf{p}'$, and that it is always possible to find a boost such that

$$\mu - \epsilon < |\mathbf{q}'| \quad \text{or} \quad \mu - \epsilon < |\mathbf{p}'|, \tag{4.43}$$

for any positive ϵ. With (4.41) we now obtain

$$\mathrm{Tr}(-VV'^{T}) + \mathrm{Tr}(-V^{T}V' + M^{2}) > 0. \tag{4.44}$$

This leads to either

$$\mu^2 > \mathbf{q}\cdot\mathbf{q}' \quad \text{or} \quad \mu^2 > \mathbf{p}\cdot\mathbf{p}'. \tag{4.45}$$

With (4.43), choosing $\mathbf{q}' = q'\hat{\mathbf{q}}$ or $\mathbf{p}' = p'\hat{\mathbf{p}}$ we conclude that

$$|\mathbf{q}| \leqslant \mu \quad \text{and} \quad |\mathbf{p}| \leqslant \mu. \tag{4.46}$$

A closer examination [by using the expression (4.29)] shows that when T can be put in the standard form (2.18) we always have

$$|\mathbf{q}| < \mu \quad \text{and} \quad |\mathbf{p}| < \mu. \tag{4.47}$$

When (4.29) is the standard form, then $|\mathbf{q}| = \mu$ and $|\mathbf{p}| = \mu$ are exceptional, however, when the standard form is (4.30) we always have

$$|\mathbf{q}| = |\mathbf{p}| = \mu. \tag{4.48}$$

V. PHYSICAL INTERPRETATION OF T^{MN}

The antisymmetric tensor T^{MN} contains ten pieces of information which are to a large extent independent. A spinning object contains the same amount of information about its state of motion at any given time: position (three parameters), momentum (three parameters), energy (one parameter), and spin (three parameters). Let us now see which identifications can be made with regard to the components of T^{MN}. First of all, let us find out which quantities are conserved. For that we apply a "time translation" Λ of the form

$$\Lambda = \begin{Vmatrix} I_3 & 0 \\ 0 & R_2 \end{Vmatrix}, \tag{5.1}$$

where R_2 is given by

$$R_2 = \begin{Vmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{Vmatrix}, \tag{5.2}$$

to T as given by (4.7):

$$T(t) = \Lambda T \Lambda^{T} = \begin{Vmatrix} \omega J & VR_2^{T} \\ -R_2 V^{T} & M \end{Vmatrix}. \tag{5.3}$$

Apparently M and ωJ are conserved. We have

$$M = \begin{Vmatrix} 0 & -\mu \\ \mu & 0 \end{Vmatrix}, \tag{5.4}$$

where $\mu = T^{54}$ has already been identified with the total energy. The conservation of ωJ is related to the conservation of $\mathbf{j}$ defined by (4.8) and (4.11). We have

$$\mathbf{j} \to j^{l} = \frac{1}{2\omega} \sum_{k,l=1}^{3} \epsilon^{ikl} T^{kl}. \tag{5.5}$$

Because of the explicit form (5.5), together with the properties that $\mathbf{j}$ is conserved and additive, it is justified to identify this quantity with the *total angular momentum* of the system of oscillating particles.

From (4.35) we have

$$\mathbf{j} = \frac{\mathbf{S}}{\mu} - \frac{\mathbf{q}\times\mathbf{p}}{\omega\mu}. \tag{5.6}$$

Here $\mathbf{q}$ and $\mathbf{p}$ are not conserved. Indeed, from (5.2) and (5.3) we find

$$\mathbf{q}(t) = \mathbf{q}\cos\omega t - \mathbf{p}\sin\omega t,$$

and

$$\mathbf{p}(t) = \mathbf{q}\sin\omega t + \mathbf{p}\cos\omega t = -\frac{1}{\omega}\frac{d\mathbf{q}(t)}{dt}, \tag{5.7}$$

so that $\mathbf{q}$ and $\mathbf{p}$ carry out harmonic oscillations. The quantity

$$\mathbf{q}(t)\times\mathbf{p}(t) = \mathbf{q}\times\mathbf{p} \tag{5.8}$$

is conserved. From (5.6) it then follows that $\mathbf{S}/\mu$ is conserved. Apparently, $\mathbf{j}$ can be split up into two separately conserved, but not additive quantities.

When T^{MN} has the standard form (2.18) it is clear from (4.35) that $S_n = 0$ if and only if the spin $\mathbf{s}_0 = 0$. We may take $S_N = 0$ as a criterion for spinlessness in general, also when the standard forms are (4.29) or (4.30).

From (4.37) and (4.38) it follows that $\mathbf{S}$ can never be zero without S_N being zero, so the condition

$$\mathbf{S} = 0 \tag{5.9}$$

can be taken as a criterion for spinlessness. Thus for a spinless system we find from (5.6)

$$\mathbf{j} = -(\mathbf{q}\times\mathbf{p})/\omega\mu. \tag{5.10}$$

Since in that case the total angular momentum is equal to the orbital angular momentum we can make the following identification, valid also when the spin is nonzero:

$$\mathbf{l} = -(\mathbf{q}\times\mathbf{p})/\omega\mu \tag{5.11}$$

is the orbital angular momentum of a system of harmonically oscillating particles. If we now write

$$\mathbf{j} = \mathbf{s} + \mathbf{l}, \tag{5.12}$$

we find for the spin angular momentum

$$\mathbf{s} = \mathbf{S}/\mu. \tag{5.13}$$

Note that $\mathbf{s}$ becomes equal to $\mathbf{s}_0$ when T^{MN} has the standard form (2.18).

Let us now turn to the interpretation of $\mathbf{q}(t)$ and $\mathbf{p}(t)$. Suppose we start from Z_0^{M} as given by (2.9) and apply a boost (3.28). Then

$$\begin{aligned} Z^{M} &= S^{\mathrm{sp}M}{}_{N} Z_0^{N} \\ &= R\,[(\delta^{M1}\sinh a + \delta^{M4}\cosh a)\sin\omega\tau \\ &\quad + (\delta^{M2}\sinh b + \delta^{M5}\cosh b)\cos\omega\tau]. \end{aligned} \tag{5.14}$$

With the help of the transformations (2.2) one would then obtain

$$x^{k} = R\,(\delta^{k1}\tanh a\,\sin\omega t + \delta^{k2}\tanh b\,\cos\omega t). \tag{5.15}$$

The T^{MN} tensor associated with Z^{M} according to (2.12) has the form

$$\begin{aligned} T^{MN} = m[&(\delta^{N1}\delta^{M2} - \delta^{M1}\delta^{N2})\sinh a\,\sinh b \\ &+ (\delta^{N1}\delta^{M5} - \delta^{M1}\delta^{N5})\sinh a\,\cosh b \\ &+ (\delta^{N4}\delta^{M2} - \delta^{M4}\delta^{N2})\cosh a\,\sinh b \\ &+ (\delta^{N4}\delta^{M5} - \delta^{M4}\delta^{N5})\cosh a\,\cosh b\,], \end{aligned} \tag{5.16}$$

which gives

$$q^k = T^{k4} = m\delta^{k2}\cosh a \sinh b,$$
$$p^k = T^{k5} = -m\delta^{k1}\sinh a \cosh b, \tag{5.17}$$
$$T^{54} = \mu = m\cosh a \cosh b.$$

Apparently, from (5.15) and (5.17) we obtain

$$\mathbf{r}(t) = (1/\omega\mu)(\mathbf{q}\cos\omega t - \mathbf{p}\sin\omega t) = \mathbf{q}(t)/\omega\mu, \tag{5.18}$$

which is the position as function of time of a single spinless particle with finite restmass. For a set of oscillating particles which is represented by some T^{MN} we can now define its "position" at time t by

$$\mathbf{r}_c(t) = \mathbf{q}(t)/\omega\mu. \tag{5.19}$$

This is oscillatory in nature but does not necessarily correspond to the position of any individual particle in the set. From the additivity of $\mathbf{q}(t)$ it follows that

$$\mu\mathbf{r}_c(t) = \sum_i \mu_{(i)}\mathbf{r}_{c(i)}(t). \tag{5.20}$$

From (4.47), (4.48), and (5.19) it follows that

$$|\mathbf{r}_c(t)| \leqslant R, \tag{5.21}$$

the equal sign being exceptional.

The equation (5.20) shows that $\mathbf{r}_c(t)$ is analogous to the nonrelativistic center-of-mass. The masses are replaced by energies. According to Eq. (5.21) the center-of-mass is confined to a sphere.

From (5.7) and (5.19) we find

$$\mathbf{p}(t) = -\mu\frac{d\mathbf{r}_c(t)}{dt}. \tag{5.22}$$

Apparently $\mathbf{p}(t)$ is just the opposite of what one would like to call the kinetic momentum of the set of particles, namely the total energy times the velocity of the center-of-mass. Therefore we shall define

$$\mathbf{p}_c(t) = -\mathbf{p}(t) = \mu\frac{d\mathbf{r}_c(t)}{dt} \tag{5.23}$$

as the total kinetic momentum. At any given time it is the sum of the kinetic momenta of the individual oscillating particles wherever they happen to be. The orbital angular momentum $\mathbf{l}$ as given by (5.11) can now be reexpressed in the form

$$\mathbf{l} = \mathbf{r}_c \times \mathbf{p}_c. \tag{5.24}$$

A further remark is in order. What is called here the kinetic momentum is not part of an energy-momentum four-vector, nor is $\mathbf{r}_c(t)$ a system-independent orbit. It is, however, most easily expressed in terms of the tensor T which displays the most important properties of a system of particles. For this reason I shall call this the "canonical center-of-mass" (c.c.m.).

It is possible to define a system-independent orbit, but this can only be done at the cost of the simple additivity rule as given by (5.20). Suppose that the transformation Λ turns T^{MN} into the standard form (2.18) with $\omega|\mathbf{s}_0| < m$. Then we can apply the inverse transformation to the vector $(0,0,0,R\sin\omega\tau, R\cos\omega\tau)$ to produce Z^M. Finally, the inverse of (2.2) produces the system-independent orbit $x^\mu(\tau)$ which is also confined. This I will call the "system-independent center-of-mass" (s.i.c.m.) $\mathbf{r}_{c'}(t)$. We have

$$\mathbf{r}_c(t) = \mathbf{r}_{c'}(t), \quad \text{when} \quad \mathbf{s} = 0, \tag{5.25}$$

and

$$\text{if} \quad \mathbf{r}_c(t) = 0, \quad \text{then also} \quad \mathbf{r}_{c'}(t) = 0. \tag{5.26}$$

For small oscillations, $\mu_{(i)} = m_{(i)}$ and $\mu = m$ and (5.20) turns into the nonrelativistic definition of center-of-mass. It is easily verified that in that case $\mathbf{r}_c(t) = \mathbf{r}_{c'}(t)$. The s.i.c.m. is probably of limited value. It cannot be defined for systems which have (4.29) or (4.30) as standard form, while for such systems the c.c.m. is well defined and unique.

Let me finally comment on the fact that when $\omega|\mathbf{s}_0| = m$, there exists an infinite number of special boosts which, when applied to the standard form (2.18), do not change the energy. In a quantized version of the model this would mean infinite degeneracy of the ground state. In order to avoid this, the case $\omega|\mathbf{s}_0| = m$ must be considered not physical. Whether the forms (4.29) and/or (4.30) may be called physical or not is a delicate question since none of these represent states of lowest energy. Most likely the form (4.30) must be considered unphysical.

VI. EXAMPLES

In order to illustrate the ideas outlined in the previous sections we consider two cases.

(1) The $\mathbf{r}_c(t)$ and $\mathbf{r}_{c'}(t)$ of a system of two spinless particles, one of which being at rest and the other carrying out a linear oscillating motion, is to be found. We have

$$\mathbf{r}_1(t) = 0,$$
$$\mathbf{r}_2(t) = \mathbf{r}_0\sin\omega t, \quad |\mathbf{r}_0| = r_0 < R. \tag{6.1}$$

The rest masses are m_1 and m_2. The form of T_1 is given by (2.6):

$$T_1^{MN} = m_1(\delta^{M5}\delta^{N4} - \delta^{M4}\delta^{N5}). \tag{6.2}$$

In order to construct T_2 we must find the transformation putting $\mathbf{r}_2$ to zero. With (2.2) we find Z^N:

$$Z^k = (Rr_0^k\sin\omega t)/\sqrt{R^2 - r_0^2\sin^2\omega t},$$
$$Z^4 = (R^2\sin\omega t)/\sqrt{R^2 - r_0^2\sin^2\omega t}, \tag{6.3}$$
$$Z^5 = (R^2\cos\omega t)/\sqrt{R^2 - r_0^2\sin^2\omega t}.$$

Let the boost Λ be defined by

$$Z^M = \Lambda^M{}_N Z'^N, \tag{6.4}$$

where

$$Z'^N = R\delta^{N4}\sin\omega\tau + R\delta^{N5}\cos\omega\tau. \tag{6.5}$$

Then the explicit form of Λ is

$$\Lambda^M{}_N = \begin{vmatrix} \alpha\mathbf{r}_0^T\mathbf{r}_0 + I_3 & \beta\mathbf{r}_0^T & 0 \\ \beta\mathbf{r}_0 & R\beta & 0 \\ 0 & 0 & 1 \end{vmatrix}, \tag{6.6}$$

where

$$\alpha = (R - \sqrt{R^2 - r_0^2})/(r_0^2\sqrt{R^2 - r_0^2}), \tag{6.7}$$

and

$$\beta = 1/\sqrt{R^2 - r_0^2}. \tag{6.8}$$

We apply (6.6) to

$$T_{20}^{MN} = m_2(\delta^{M5}\delta^{N4} - \delta^{M4}\delta^{N5}), \tag{6.9}$$

and obtain

$$T_2^{MN} = \left|\left| \begin{matrix} 0 & 0 & -m_2\beta\mathbf{r}_0^T \\ 0 & 0 & -R\beta m_2 \\ m_2\beta\mathbf{r}_0 & R\beta m_2 & 0 \end{matrix} \right|\right|. \tag{6.10}$$

The energy of particle 2 is

$$\mu_2 = R\beta m_2 = Rm_2/\sqrt{R^2 - r_0^2}.$$

Now we add T_1 and T_2 together:

$$T^{MN} = T_1^{MN} + T_2^{MN}$$

$$= \left|\left| \begin{matrix} 0 & 0 & -m_2\beta\mathbf{r}_0^T \\ 0 & 0 & -m_1 - R\beta m_2 \\ m_2\beta\mathbf{r}_0 & m_1 + R\beta m_2 & 0 \end{matrix} \right|\right|. \tag{6.11}$$

With (5.7) we find $\mathbf{q}(t)$ and $\mathbf{p}(t)$:

$$\mathbf{q}(t) = m_2\beta\mathbf{r}_0 \sin \omega t, \tag{6.12}$$

$$\mathbf{p}(t) = -m_2\beta\mathbf{r}_0 \cos \omega t. \tag{6.13}$$

The energy μ is equal to $m_1 + R\beta m_2$. With (5.19) we find for $\mathbf{r}_c(t)$,

$$\mathbf{r}_c(t) = \mathrm{Rm}_2\mathbf{r}_0 \sin \omega t/(Rm_2 + m_1\sqrt{R^2 - r_0^2}). \tag{6.14}$$

while

$$\mathbf{p}_c(t) = -\mathbf{p}(t) = (m_2\mathbf{r}_0/\sqrt{R^2 - r_0^2}) \cos \omega t. \tag{6.15}$$

In order to find $\mathbf{r}_{c'}(t)$ we must transform (6.11) into the standard form. For that we apply the boost

$$\Lambda'^{M}{}_{N} = \left|\left| \begin{matrix} \alpha'\mathbf{r}_0^T\mathbf{r}_0 + I_3 & -\beta'\rho\mathbf{r}_0^T & 0 \\ -\beta'\rho\mathbf{r}_0 & R\beta' & 0 \\ 0 & 0 & 1 \end{matrix} \right|\right|, \tag{6.16}$$

with

$$\alpha' = \left(R - \sqrt{R^2 - \rho^2 r_0^2}\right)/r_0^2\sqrt{R^2 - \rho^2 r_0^2}, \tag{6.17}$$

and

$$\beta' = 1/\sqrt{R^2 - \rho^2 r_0^2}, \tag{6.18}$$

while ρ must be determined. This gives

$$\Lambda' T \Lambda' = \left|\left| \begin{matrix} 0 & 0 & -\mathbf{a}^T \\ 0 & 0 & -b \\ \mathbf{a} & b & 0 \end{matrix} \right|\right|, \tag{6.19}$$

with

$$a^k = [(\alpha' r_0^2 + 1)m_2\beta - (m_1 + R\beta m_2)\beta'\rho]r_0^k, \tag{6.20}$$

and

$$b = -m_2\beta\beta'\rho r_0^2 + (m_1 + R\beta m_2)R\beta'. \tag{6.21}$$

For $\Lambda' T \Lambda'$ to be in the standard form we require $a^k = 0$, which equation must be solved for ρ. The result is

$$\rho = Rm_2/(Rm_2 + m_1\sqrt{R^2 - r_0^2}). \tag{6.22}$$

If this is inserted into (6.21) we find

$$b = m = \left[(m_1^2 + m_2^2) + 2Rm_1m_2/\sqrt{R^2 - r_0^2}\right]^{1/2}. \tag{6.23}$$

Next we determine $\mathbf{r}_{c'}(t)$. We apply Λ'^{-1} to Z'^{N} as given by (6.5), where

$$\Lambda'^{-1} = \Lambda(\rho \to -\rho) = \left|\left| \begin{matrix} \alpha'\mathbf{r}_0^T\mathbf{r}_0 + I_3 & \beta'\rho\mathbf{r}_0^T & 0 \\ \beta'\rho\mathbf{r}_0 & R\beta' & 0 \\ 0 & 0 & 1 \end{matrix} \right|\right|. \tag{6.24}$$

Then we have

$$Z^M = (\Lambda'^{-1})^M{}_N Z'^N, \tag{6.25}$$

so that

$$Z^k = \beta'\rho R r_0^k \sin \omega\tau,$$

$$Z^4 = R^2\beta' \sin \omega\tau, \tag{6.26}$$

$$Z^5 = R \cos \omega\tau.$$

We determine x^k and x^4 with the help of the inverse of (2.2) and eliminate τ. The result is identical with (6.14), so $\mathbf{r}_c(t) \equiv \mathbf{r}_{c'}(t)$. From the form of (6.19) we see also that the total spin is zero.

(2) The $\mathbf{r}_c(t)$ and $\mathbf{r}_{c'}(t)$ of a single spinning particle is to be found.

We start from the standard form (2.18) with $\omega|\mathbf{s}_0| < m$ and apply the boost (3.28). We find

$$q_1 = 0,$$

$$q_2 = m \cosh a \sinh b - \omega s_{03} \sinh a \cosh b, \tag{6.27}$$

$$q_3 = \omega s_{02} \sinh a,$$

and

$$p_1 = -m \sinh a \cosh b + \omega s_{03} \cosh a \sinh b,$$

$$p_2 = 0, \tag{6.28}$$

$$p_3 = -\omega s_{01} \sinh b,$$

while [see (4.2)]

$$\mu = m \cosh a \cosh b - \omega s_{03} \sinh a \sinh b. \tag{6.29}$$

From (5.7) we obtain

$$q_1(t) = (m \sinh a \cosh b - \omega s_{03} \cosh a \sinh b) \sin \omega t,$$

$$q_2(t) = (m \cosh a \sinh b - \omega s_{03} \sinh a \cosh b) \cos \omega t,$$

$$q_3(t) = \omega s_{02} \sinh a \cos \omega t + \omega s_{01} \sinh b \sin \omega t. \tag{6.30}$$

From (5.19), with the help of (6.29) and (6.30) we obtain

$$r_{c_1}(t) = R\frac{m \sinh a \cosh b - \omega s_{03} \cosh a \sinh b}{m \cosh a \cosh b - \omega s_{03} \sinh a \sinh b} \sin \omega t,$$

$$r_{c_2}(t) = R\frac{m \cosh a \sinh b - \omega s_{03} \sinh a \cosh b}{m \cosh a \cosh b - \omega s_{03} \sinh a \sinh b} \cos \omega t, \tag{6.31}$$

$$r_{c_3}(t) = \frac{s_{02} \sinh a \cos \omega t + s_{01} \sinh b \sin \omega t}{m \cosh a \cosh b - \omega s_{03} \sinh a \sinh b}.$$

This is the c.c.m. For the s.i.c.m. we find

$$r_{c_1'}(t) = R \tanh a \sin \omega t,$$

$$r_{c_2'}(t) = R \tanh b \cos \omega t, \tag{6.32}$$

$$r_{c_3'}(t) = 0,$$

and is simply obtained from (6.31) by putting $\mathbf{s}_0$ equal to zero.

VII. CONCLUSIONS

To define a center-of-mass or even a center-of-mass system for relativistic particles moving in an external field is a

notoriously hard problem. As demonstrated in this article, the relativistic harmonic oscillator is an exception. Because of the high symmetry, quantities like energy, momentum, spin, and orbital angular momentum can be satisfactorily defined, although they do not transform in a conventional way. The possibility to define a center-of-mass in a meaningful way may be of use in the consideration of systems of quarks which are permanently confined inside a "bag" with SO(3,2) symmetry.[2] Here of course, quantized fields should be considered. On this subject much work has been done already, but so far the results have been only of academic interest and the problems have only been formulated mathematically.[7] It is the purpose of future research to apply this to the geometric quark–gluon model of which the contents of this paper are only a classical simplification.

ACKNOWLEDGMENTS

I wish to thank Dr. E. van Beveren for pointing out to me the relevance of understanding the nature of center-of-mass motions, and the members of the Institute for Theoretical Physics at Nijmegen for discussions.

[1]E. van Beveren, G. Rupp, T. A. Rijken, and C. Dullemond, Phys. Rev. D **27**, 1527 (1983).

[2]C. Dullemond, T. A. Rijken, and E. van Beveren, Nijmegen Report THEF-NYM-83.02, 1983.

[3]C. Dullemond and E. van Beveren, Phys. Rev. D **28**, 1028 (1983).

[4]A. Chodos, R. L. Jaffe, K. Johnson, C. B. Thorn, and V. F. Weisskopf, Phys. Rev. D **9**, 3471 (1974).

[5]See, e.g., Hans Freudenthal and H. de Vries, *Linear Lie Groups* (Academic, New York, 1969).

[6]I thank Prof. H. de Vries (University of Nijmegen) for pointing this out to me.

[7]N. T. Evans, J. Math. Phys. **8**, 170 (1967); C. Fronsdal, Phys. Rev. D **12**, 3819 (1975); W. Heidenreich, J. Math. Phys. **22**, 1566 (1981).

Homogeneous canonical formulation of the nonrelativistic hydrogen atom

Horacio Grinberg,[a] Julio Marañon,[a] and Hector Vucetich[a]
Laboratorio de Física Teórica, Departamento de Física, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, C.C. No. 67 (1900) La Plata, Argentina

(Received 20 July 1982; accepted for publication 13 January 1984)

The homogeneous canonical formulation is applied to the Hamiltonian of the nonrelativistic hydrogen atom. Its connection with the isotropic harmonic oscillator in a four-dimensional Riemann space leads to the quantum analog of the Kepler problem.

PACS numbers: 03.20. + i, 03.65. − w

I. INTRODUCTION

The study of the motion in a Coulomb field, as well as its connection with the isotropic harmonic oscillator, has been a subject of increasing interest over the past few years. For conservative systems, it was classically shown that there exists a "local diffeomorphism" which maps a given Hamiltonian into another one with a new potential function[1]: that is, solutions for a given potential can be transformed into solutions for another different one. These relationships have also been investigated, in connection with the application of quantum mechanics to the quarkonium, through the study of the expression governing the "power law potentials." In fact, it was shown that the power n of the starting potential is related to the power n' of the potential to be obtained through the constraint[2]

$$2[(n'/n)+1]+n=0 \quad (-2<n<\infty). \tag{1}$$

Simultaneously, using the Schrödinger representation and the path integral formalism,[3] together with the Kustaanheimo–Stiefel transformation,[4] the equivalence between the nonrelativistic hydrogen atom and a four-dimensional isotropic harmonic oscillator was established.[5] This method is essentially based on the study of the system in a four-dimensional symmetric space (isotropic and homogeneous space), where it behaves as a free particle.[6,7] As is well known, as the dimension of the space increases it becomes important to take care of the hidden symmetry. In the particular case of the hydrogen atom the new quantity that is conserved (Noether theorem) in addition to the angular momentum is the Runge–Lenz vector.

Section II of the present paper is devoted to a short review of the homogeneous canonical formulation of classical mechanics following the lines of Dirac.[8] This formulation will enable us to justify in a natural way the method used in previous works.[4–7] In Sec. III the homogeneous canonical formulation is applied to the nonrelativistic hydrogen atom in order to show its equivalence to a four-dimensional isotropic harmonic oscillator.

II. HOMOGENEOUS CANONICAL FORMULATION

In $\mathbf{R}^3$ the Lagrangian of a conservative mechanical system is

$$L=p_i\dot{q}_i-H(p,q), \tag{2}$$

where the q's and the p's are the dynamical variables of the Hamiltonian theory. Since the variation of the quantity $(p_i\dot{q}_i - L)$ does not involve the variation of the velocities, the Hamiltonian H is given by $(m=1)$

$$H(p,q)=\tfrac{1}{2}p_i^2+V(q_i), \tag{3}$$

where $V(q)$ is the potential energy. If the time t is considered as an independent variable we are led to the inhomogeneous canonical formalism. Let us imagine a new independent parameter s. When a dynamical problem is solved, the functions

$$p_i=p_i(s), \quad q_i=q_i(s), \quad t=t(s) \tag{4}$$

are known, and the disymmetry in the set (4) is evident. The formalism would be more elegant if one could count the time t along the q_i's. This requires the introduction of a momentum p_t canonically conjugate to t. This can only be done if we move from the inhomogeneous Hamiltonian H to a homogeneous Hamiltonian H^H. Of course then

$$\frac{\partial H^H}{\partial s}=0. \tag{5}$$

In fact, through the integral transform of time scaling

$$t=\int_{s_1}^{s_2} a(s)ds \tag{6}$$

one is naturally led to the homogeneous canonical formulation of a mechanical system in a four-dimensional curved space (three spatial coordinates and one temporal). In the Minkowski space of special relativity, for instance, time is one of the coordinates describing the world line of a particle. But the use of a formalism based on a homogeneous treatment is in no way restricted to special relativity, because it must remain connected on one "time-coordinate" only and can, on this ground, only describe reasonably the behavior of a one-particle system and not a system of several particles each requiring a different proper time. Of course the use of the homogeneous formalism is connected with one difficulty, i.e., the definition and interpretation of the momentum p_t canonically conjugated to the coordinate t. For in the inhomogeneous formalism, either the Lagrangian $L(q,\dot{q},t)$ or the Hamiltonian $H(p,q,t)$ are functions of t but not of a magnitude t' or p_t. Therefore no definition of the kind

$$\dot{p}_t=\frac{\partial L}{\partial t'} \quad \left(' = \frac{d}{ds}\right) \tag{7}$$

[a] MCIC CONICET República Argentina.

 0022-2488/84/092648-03$02.50

can be agreed upon, unless the functions L or H are replaced by more comprehensive functions L^H and H^H.

Let us therefore consider the more general Lagrangian L^H from which the possible behavior of the system is described either by a variational principle

$$\delta\int_{t_1}^{t_2}\left[\left(p_i\dot{q}_i\frac{dt}{ds}-H\frac{dt}{ds}\right)\right]ds=\delta\int_{s_1}^{s_2}L^H\,ds=0, \tag{8}$$

that is,

$$\delta\int_{s_1}^{s_2}\left(p_i\frac{dq_i}{ds}-H\frac{dt}{ds}\right)ds=0, \tag{9}$$

or by corresponding canonical equations

$$q_i'=[q_i,H]_-,\quad p_i'=[p_i,H]_-. \tag{10}$$

Here, the brackets [,] are to be defined by the help of derivations with regard to the q's, p's:

$$[\Theta,\Omega]_-=\sum_i\left(\frac{\partial\Theta}{\partial q_i}\frac{\partial\Omega}{\partial p_i}-\frac{\partial\Theta}{\partial p_i}\frac{\partial\Omega}{\partial q_i}\right). \tag{11}$$

The transformation (6) cannot change the fact that H is constant, i.e., from Eq. (9) it can be seen that the canonical momentum p_t of the time coordinate is

$$p_t=-H=\frac{\partial L^H}{\partial t'}. \tag{12}$$

Thus, the Lagrangian L^H can be written as

$$L^H=p_iq_i'+p_tt'. \tag{13}$$

The new homogeneous Lagrangian L^H fulfills Euler's theorem, as it is a first-order homogeneous function in q' and t'. This theorem leads to an accessory condition. In fact, the homogeneous Hamiltonian H^H corresponding to L^H is by definition

$$H^H=p_iq_i'+p_tt'-L^H=0. \tag{14}$$

It can be seen that the introduction of the new independent variable s, through the constraint (6), transforms the homogeneous Lagrangian L^H into a singular Lagrangian. In fact, differentiating Eq. (13) with respect to t', yields

$$\frac{\partial^2L^H}{\partial q_i'\,\partial t'}=0, \tag{15}$$

that is,

$$\text{Det}\left\|\frac{\partial^2L^H}{\partial q_i'\,\partial t'}\right\|=0, \tag{16}$$

which is precisely the condition for L^H to be singular.

In the nomenclature of Dirac[8] Eq. (13) is a strong condition and Eq. (14) is a weak condition. The next step of the formalism is to find the form of the homogeneous Hamiltonian H^H. To this end we write the Hamilton equations as

$$t'\frac{\partial H}{\partial p_i}=\frac{\partial H^H}{\partial p_i},\quad t'\frac{\partial H}{\partial q_i}=\frac{\partial H^H}{\partial q_i}, \tag{17a}$$

$$t'=\frac{\partial H^H}{\partial p_t},\quad p_t'=\frac{-\partial H^H}{\partial t}. \tag{17b}$$

Equations (17) can be regarded as differential equations for H^H. Their solution can easily be given if H^H is assumed to depend upon the variables through the intermediary of

$$\zeta=H+p_t. \tag{18}$$

Then

$$H^H=F(\zeta)=F(H+p_t). \tag{19}$$

Therefore, in order to satisfy Eq. (17) we must require that

$$\frac{\partial H^H}{\partial\zeta}=\frac{\partial F(\zeta)}{\partial\zeta}=t'. \tag{20}$$

Hence, applying condition (6), i.e., $t'=a(s)$, one gets

$$\frac{\partial H^H}{\partial\zeta}=a(s), \tag{21}$$

that is,

$$H^H=(H+p_t)a(s). \tag{22}$$

III. HYDROGEN ATOM

We now apply the homogeneous canonical formalism described in Sec. II to the Coulomb problem. The Hamiltonian involving the Coulomb potential reads

$$H=\tfrac{1}{2}p^2+\mu/r, \tag{23}$$

where μ is the strength of the potential. The homogeneous Hamiltonian H^H [Eq. (22)] is

$$H^H=(\tfrac{1}{2}p^2+\mu/r+p_t)a(s). \tag{24}$$

The change of time scaling $a(s)$ is chosen so as to regularize the homogeneous Hamiltonian H^H and the new auxiliary variable s plays the same role as that of the "proper" time in the relativistic problem; the choice is

$$a(s)=r(s). \tag{25}$$

With this substitution, Eq. (24) becomes

$$H^H=\tfrac{1}{2}rp^2+rp_t+\mu. \tag{26}$$

Since we are interested in bound states, $H(q_i,p_i)=E$ (constant of the motion) and the homogeneous Hamiltonian can be written as

$$H^H=\tfrac{1}{2}rp^2-rE+\mu. \tag{27}$$

It is easily shown that H^H in Eq. (27) satisfies the subsidiary condition (14). In fact, from Dirac's theory we have

$$\tfrac{1}{2}p^2+p_t+\mu/r=0, \tag{28}$$

since as stated, it is a weak condition.[8]

Let us now write the Hamiltonian (27) in $\mathbb{R}^4$ and apply the Kustaanheimo–Stiefel[4] mapping that realizes a point canonical transformation from the Euclidean space $\mathbb{R}^4$ onto a three-dimensional curved Riemann space of the canonical variables, that is, $(q_i,p_i)\rightarrow(u,p_u)$ so that $r=u^2$:

$$q_i=\sum_{j=1}^{4}A_{ij}(u)u_j\quad(i=1,2,3), \tag{29a}$$

$$dq_4=2\sum_{j=1}^{4}A_{4j}(u)du_j, \tag{29b}$$

$$p_i=\frac{1}{2r}\sum_{j=1}^{4}A_{ij}(u)(p_u)_j\quad(i=1,2,3,4), \tag{29c}$$

where the matrix $A(u)$ is given by

$$A(u)=\begin{pmatrix} u_3 & u_4 & u_1 & u_2 \\ -u_2 & -u_1 & u_4 & u_3 \\ -u_1 & u_2 & u_3 & -u_4 \\ u_4 & -u_3 & u_2 & -u_1 \end{pmatrix}. \tag{30}$$

With this matrix transformation the homogeneous canonical Hamiltonian $\overline{H}^H$ becomes

$$\overline{H}^H = \tfrac{1}{8}p_u^2 - u^2E + \mu, \tag{31}$$

with the annihilation condition

$$dq_4 = 2(u_4\,du_1 - u_3\,du_2 + u_2\,du_3 - u_1\,du_4) = 0. \tag{32}$$

Equation (29b) satisfies the consistency condition[8]

$$q_4' = \frac{dq_4}{ds} = \{q_4, \overline{H}^H\} = 0, \tag{33}$$

where the bracket is given by

$$\{q_4, \overline{H}^H\} = 2p_4 = 2[u_4(p_u)_1 - u_3(p_u)_2 + u_2(p_u)_3 - u_1(p_u)_4] = 0, \tag{34}$$

which is a weak equation.

The physical motion in $\mathbb{R}^3$ can be studied through the homogeneous Hamiltonian (31) as the motion of a particle in a four-dimensional Riemann space subjected to the constraint (34). Finally, the accessory condition (14) is transformed in the equation

$$\overline{H}^H = 0, \tag{35}$$

which is again a weak condition.

Introducing the relationships

$$-E = \tfrac{1}{2}\omega^2, \quad E_0 = -\mu = -p_s, \tag{36}$$

Eq. (35) can be written as

$$\overline{H}^H = \tfrac{1}{8}p_u^2 + \tfrac{1}{2}u^2\omega^2 + p_s = 0. \tag{37}$$

This expresses a conservation law

$$\overline{H} + p_s = 0, \tag{38}$$

where

$$\overline{H} = \tfrac{1}{8}p_u^2 + \tfrac{1}{2}u^2\omega^2 \tag{39}$$

is the Hamiltonian of a four-dimensional isotropic harmonic oscillator. This result is the quantum analog of the Kepler problem whose connection with the isotropic harmonic oscillator was analyzed by Stiefel *et al.* (the new variable s played the role of excentric anomaly) using the homogeneous canonical formulation,[9] but in a way unrelated to the Dirac formulation.[8]

The homogeneous Hamiltonians H^H fulfill the Dirac condition (i.e., they lead to weak equations) and identifying the p_u and p_s through the connections

$$(p_u)_j \to -ih\frac{\partial}{\partial u_j} \quad (j = 1, 2, 3, 4), \tag{40a}$$

$$p_s \to -ih\frac{\partial}{\partial s}, \tag{40b}$$

one obtains the Schrödinger equation for a four-dimensional harmonic oscillator; well-known results for the hydrogen wave function can be recovered from it.[7] In this way, we have shown the equivalence of the homogeneous formulation of the three-dimensional hydrogen atom with the inhomogeneous formulation of a four-dimensional harmonic oscillator.

It should also be mentioned that the bilinear annihilation condition (32) together with the integral transform of time scaling (6) lead to the singularity of the homogeneous Lagrangian $L^H(u', t', u, t)$. These do not represent primary constraints in the sense of Dirac. In fact, both conditions commute with the Dirac Hamiltonian H^D,

$$\overline{H}^D = \overline{H}^H + \lambda_1(t' - r) + \lambda_2(u_4\,du_1 - u_3\,du_2 + u_2\,du_3 - u_1\,du_4), \tag{41}$$

i.e., $\lambda_1 = \lambda_2 = 0$. They also commute with each of the 15 generators of the corresponding Lie algebra SO(4, 2).[10] Consequently, both equations express the cyclic character of the coordinates q_4 and t.

ACKNOWLEDGMENT

We are grateful to CONICET for a research grant in support of this work.

[1] P. Collas, J. Math. Phys. **22**, 2512 (1981).
[2] C. Quigg and J. L. Rosner, Phys. Rep. **56**(4), 167 (1979).
[3] A. C. Chen, Phys. Rev. A **22**, 333 (1980).
[4] H. Duru and H. Kleinert, Phys. Lett. B **84**, 185 (1979).
[5] P. Kustaanheimo and E. Stiefel, J. Reine Angew. Math. **218**, 204 (1965).
[6] G. A. Ringwood and J. T. Devreese, J. Math. Phys. **21**, 1390 (1980).
[7] H. Grinberg, J. Marañón, and H. Vucetich, J. Chem. Phys. **78**, 839 (1983); H. Grinberg, J. Marañón, and H. Vucetich, Int. J. Quantum Chem. **23**, 379 (1983); H. Grinberg, J. Marañón, and H. Vucetich, Z. Phys. C **20**, 147 (1983).
[8] P. A. M. Dirac, *Lectures on Quantum Mechanics* (Belfer Graduate School of Science, Yeshiva University, New York, 1964). See also A. Mercier, *Analytical and Canonical Formalism in Physics* (North-Holland, Amsterdam, 1959).
[9] E. L. Stiefel and G. Scheifele, *Linear and Regular Celestial Mechanics* (Springer, Berlin, 1971).
[10] J. Baumgarte, J. Reine Angew. Math. **301**, 59 (1978).

Double phase-integral approximations: A systematic simplification technique for wave equations with cutoffs and resonances

Andrzej A. Skorupski
Institute of Nuclear Research, Hoża 69, 00-681 Warsaw, Poland

(Received 23 August 1983; accepted for publication 20 January 1984)

The time-independent wave equation, $d^2\psi/dz^2 + Q^2(z)\psi = 0$, where $Q^2(z)$ may have arbitrary order zeros and poles on or close to the real axis, is transformed to a simpler wave equation of similar properties (model). Approximate transformations leading from the original wave equation to the model are simply related to Fröman's higher-order phase integrals, but are nevertheless well defined at the pertinent zeros and poles of $Q^2(z)$.

PACS numbers: 03.40.Kf, 02.30.Mv, 02.30.Qy

I. INTRODUCTION

We consider the time-independent wave equation in one dimension,

$$\frac{d^2\psi}{dz^2} + Q^2(z)\psi = 0, \quad z = x + iy, \; Q^2 \propto \lambda^{-2}, \tag{1}$$

where $Q^2(z)$ may have zeros and poles in some vicinity of the real axis. These zeros will be called cutoffs (they are called turning points in quantum mechanics); the poles will be referred to as resonances. We assume that $Q^2(z)$ contains a small parameter λ. It will only be used to define ordering, and will never appear explicitly in the final results. Equation (1) can be solved systematically in λ^2 by using the higher-order phase-integral approximations (PIA) introduced by Fröman.[1] They have certain advantages over the standard higher-order JWKB approximations.[2] However, all these approximations break down in the vicinity of zeros and simple poles of $Q^2(z)$, and in general they are also poor in the vicinity of the second-order poles.[3] Presence of cutoffs and resonances may thus require going to the modified PIA,[4–5] or using the connection formulas for zeros of $Q^2(z)$,[6–8] but all these concepts have their limitations. A more general approach would be to transform Eq. (1) to a simpler wave equation with the same type of cutoffs and resonances,

$$\frac{d^2\tilde{\psi}}{d\tilde{z}^2} + \tilde{Q}^2(\tilde{z})\tilde{\psi} = 0, \quad \tilde{z} = \tilde{x} + i\tilde{y}, \; \tilde{Q}^2 \propto \lambda^{-2}, \tag{2}$$

which can be solved analytically. Such an equation will be called a model, or comparison equation; model quantities will be distinguished from those referring to the original wave equation by a tilde. Approximate transformation leading from Eq. (1) to Eq. (2) in lowest order in λ^2 was found by Miller and Good,[9] and by Dingle.[10] Here we generalize this result to higher orders in λ^2 for arbitrary type of cutoffs and resonances, both real and complex conjugate. Approximate transformations from Eq. (1) to Eq. (2) are shown to be simply related to ordinary higher-order PIA but for both Eq. (1) and Eq. (2). Therefore we call them double phase-integral approximations (DPIA). In contrast to ordinary PIA they are well defined at the cutoffs and resonances. Earlier attempts to determine this type of higher-order corrections required tedious calculations,[11–13] and separate analysis for the neighboring physical situations, such as, e.g., the transmission through a potential barrier with the energy slightly below or slightly above the top of the barrier. These results now follow immediately from our theory as special cases.

This paper deals with general properties of DPIA. Typical applications are described in Ref. 14.

II. THEORY OF DPIA

The simultaneous transformation of the independent variable, $z \to \tilde{z}$, and the unknown function, $\psi \to \tilde{\psi}$, leading from Eq. (1) to Eq. (2), is generated by one function $q(z)$ [$= d\tilde{z}/dz$, see, e.g., Eqs. (3.1)–(3.5) in Ref. 7]:

$$\tilde{z} = \int q(z)\, dz, \tag{3a}$$

$$\tilde{\psi} = q^{1/2}\psi, \tag{3b}$$

$$\tilde{Q}^2 = q^{-2}\left(Q^2 + q^{1/2}\frac{d^2}{dz^2}q^{-1/2}\right). \tag{3c}$$

We assume that $Q^2(z)$ in Eq. (1) is analytic, except for isolated singularities, and is also real on the x axis. In a given segment (a,b) of the x axis (or close to it) $Q^2(z)$ is assumed to have a certain number of zeros and/or poles z_i, $i = 1,2,...,M_x$, but no essential singularities. Thus $Q^2(z)$ can be written

$$Q^2(z) = f(z)\prod_{i=1}^{M_x}(z - z_i)^{m_i}, \tag{4}$$

where m_i is a positive or negative integer, and $f(z)$ has no zeros or singularities in the vicinity of (a,b); obviously $f(x)$ must have a constant sign in (a,b)

$$f(x) > 0 \text{ or } f(x) < 0 \quad \text{for } a < x < b. \tag{5}$$

The function $\tilde{Q}^2(\tilde{z})$ in Eq. (2) is assumed to have the same type of cutoffs and resonances as $Q^2(z)$, i.e.,

$$\tilde{Q}^2(\tilde{z}) = \tilde{f}(\tilde{z})\prod_{i=1}^{M_x}(\tilde{z} - \tilde{z}_i)^{m_i}, \tag{6}$$

where $\tilde{f}(\tilde{z})$ has the same general properties as $f(z)$ in Eq. (4), and $\tilde{z}_i = \tilde{z}(z = z_i)$. Usually $\tilde{f}(\tilde{z})$ will be simpler than $f(z)$, and the simplest choice is $\tilde{f}(\tilde{z}) = \text{const}$. The actual problem, however, is to simplify the differential equation. Thus we assume that Eq. (2) is solvable analytically.

Equation (3c) is a nonlinear differential equation for the function $\tilde{z}(z)$ if $Q^2(z)$ and $\tilde{Q}^2(\tilde{z})$ are given. Assuming that $q(z) \neq 0, \infty$, a solution of Eq. (3c) can immediately be found in lowest order in λ^2 by deleting the differential term,[9,10] which yields

 0022-2488/84/092651-04$02.50

$$q(z) \equiv \frac{d\tilde{z}}{dz} = \frac{Q(z)}{\tilde{Q}(\tilde{z})} . \tag{7a}$$

Integrating Eq. (7a) one finds

$$\int Q(z)\, dz = \int \tilde{Q}(\tilde{z})\, d\tilde{z} . \tag{7b}$$

To check the assumption of $q(z) \neq 0, \infty$ at $z = z_j$ we expand $Q^2(z)$ given by Eq. (4) in powers of $(z - z_j)$,

$$Q^2(z) = c_j (z - z_j)^{m_j} \left[1 + \sum_{k=1}^{\infty} g_k (z - z_j)^k \right] , \tag{8a}$$

where

$$c_j = f(z_j) \prod_{\substack{i=1 \\ (i \neq j)}}^{M_x} (z_j - z_i)^{m_i} , \tag{8b}$$

and similarly for $\tilde{Q}^2(\tilde{z})$. Inserting these expansions into Eq. (7a) we obtain, in the limit $z \to z_j$,

$$[q(z_j)]^{m_j + 2} = c_j / \tilde{c}_j , \tag{9}$$

which defines $q(z_j) \neq 0, \infty$, as required, if $m_j \neq -2$. For $m_j = -2$ Eq. (9) gives a constraint upon the unknown parameters contained in Eq. (6), $\tilde{z}_i$, etc.

Finding the mapping $z \to \tilde{z}$ in higher orders by a straightforward iteration starting from Eqs. (7a),(7b) can be inconvenient (see Sec. I). One might think of the nonstandard approach to solving Eq. (3c) (both in zeroth and higher orders) proposed in Ref. 15, but applicability of these results is too limited (Ref. 15 treats one turning point only, and offers no straightforward generalizations). Our approach is to relate the solution in question to the higher-order phase-integral approximations[1] (PIA), but for both the original wave equation (1) and the comparison equation (2). Using PIA of order $2N + 1$ ($N \geqslant 0$) we find, for Eq. (1),

$$\psi(z) \approx [q_{2N+1}(z)]^{-1/2} \exp[\pm i\, w_{2N+1}(z)] , \tag{10a}$$

$$q_{2N+1}(z) = Q(z)[1 + Y_2(z) + \cdots + Y_{2N}(z)] , \tag{10b}$$

$$w_{2N+1}(z) = \int q_{2N+1}(z)\, dz , \tag{10c}$$

where the higher-order corrections $Y_{2n}(z)$ ($\propto \lambda^{2n}$) are uniquely defined polynomials in the quantities ϵ_p, defined as

$$\epsilon_p = \frac{d^p \epsilon_0}{d\zeta^p} , \quad p \geqslant 0 , \tag{11a}$$

$$\epsilon_0 = \frac{1}{Q^2} \left(Q^{1/2} \frac{d^2}{dz^2} Q^{-1/2} \right) = \frac{1}{16 Q^6} \left[5 \left(\frac{dQ^2}{dz} \right)^2 - 4 Q^2 \frac{d^2 Q^2}{dz^2} \right] , \tag{11b}$$

$$\zeta = \int Q(z)\, dz , \tag{11c}$$

[e.g., $Y_4 = \epsilon_0/2$, $Y_4 = -(\epsilon_0^2 + \epsilon_2)/8$, etc., see Ref. 16 for Y_{2n} up to Y_{20}].

The transformation $z \to w_{2N+1}$ generated by q_{2N+1} given by Eq. (10b), transforms Eq. (1) into the wave equation in which the pertinent coefficient, $K^2(w_{2N+1})$, is close to unity (see Ref. 3, Chap. 2B):

$$K^2(w_{2N+1}) = 1 + 2 Y_{2N+2} [1 + O(\lambda^2)] = 1 + O(\lambda^{2N+2}) . \tag{12}$$

Including only terms up to λ^{2N} [to be consistent with Eq. (10b)] we thus obtain $K^2 = 1$. Similarly we obtain $\tilde{K}^2 = 1$ after the transformation $\tilde{z} \to \tilde{w}_{2N+1}$, generated by

$$\tilde{q}_{2N+1} = \tilde{Q} [1 + \tilde{Y}_2 + \cdots + \tilde{Y}_{2N}] , \tag{13}$$

where again for consistency only terms up to λ^{2N} are included ($\tilde{Y}_{2n} \propto \lambda^{2n}$). Thus at the considered level of accuracy the wave equations after the transformations $z \to w_{2N+1}$, and $\tilde{z} \to \tilde{w}_{2N+1}$ coincide, which implies

$$\frac{d\tilde{w}_{2N+1}}{dw_{2N+1}} \equiv \frac{\tilde{q}_{2N+1}\, d\tilde{z}}{q_{2N+1}\, dz} = 1 . \tag{14}$$

[This is a consequence of the equation analogous to Eq. (3c), for the transformation $w_{2N+1} \to \tilde{w}_{2N+1}$.] With this approximation the generating function for the mapping $z \to \tilde{z}$ thus becomes

$$q(z) \equiv \frac{d\tilde{z}}{dz} = \frac{q_{2N+1}}{\tilde{q}_{2N+1}} \equiv \frac{Q[1 + Y_2 + \cdots + Y_{2N}]}{\tilde{Q}[1 + \tilde{Y}_2 + \cdots + \tilde{Y}_{2N}]} . \tag{15a}$$

Integrating Eq. (15a) we find

$$\int q_{2N+1}(z)\, dz = \int \tilde{q}_{2N+1}(\tilde{z})\, d\tilde{z} . \tag{15b}$$

The higher-order approximation to the function $\tilde{z}(z)$, defined by Eq. (15b), will be called double phase-integral approximation of order $2N + 1$.

Note that our derivation requires $|Y_{2n}| \ll 1$ (and $|\tilde{Y}_{2n}| \ll 1$), which is only true for $|z - z_{jc}| \gg R_{jc}$, where z_{jc} is either a zero or a simple pole of $Q^2(z)$, and R_{jc} is a critical radius which can be associated with any such critical point z_{jc} (see Ref. 3). In favorable situations (i.e., λ small enough) R_{jc} are small so that the critical circles, $|z - z_{jc}| < R_{jc}$, do not overlap. The DPIA should then be accurate, and improve in higher orders, along any path in the complex plane passing by the critical circles. This can lead to very accurate results concerning the connection problems (e.g., transmission coefficients etc.), as illustrated in Ref. 14. At the same time the accuracy of the DPIA in the immediate vicinity of z_{jc} (where $|Y_{2n}| \gg 1$, and $|\tilde{Y}_{2n}| \gg 1$) requires separate treatment. Thus the question of how good the DPIA could be for relating $\psi(z)$ to $\tilde{\psi}(\tilde{z})$ in the immediate vicinity of cutoffs and resonances remains open. The only point we would like to make is that Eq. (15a) is bounded in the limit $z \to z_j$, i.e., we do not spoil the lowest-order behavior of $q(\neq 0, \infty; Y_{2n} \to \infty$ at z_{jc}, but $Y_{2n}/\tilde{Y}_{2n} \to \text{const} \neq 0, \infty)$. This can be seen from general expressions for $Y_{2n}(z)$ corresponding to $Q^2(z)$ given by Eq. (8a) [see Eqs. (57a), (57b) in Ref. 3], leading to

$$q_{2N+1}(z) = c_j^{1/2} (z - z_j)^{m_j/2} \left[1 + \sum_{k=1}^{\infty} a_k (z - z_j)^k \right] \times \left\{ 1 - \sum_{n=1}^{N} \frac{1}{c_j^n} \left[1 + \sum_{k=1}^{\infty} a_{nk} (z - z_j)^k \right] \times \frac{1}{2} m_j (m_j + 4)(m_j + 2)^{2n-2} \times (-16)^{-n} \beta_n (2n - 1)! (z - z_j)^{-(m_j + 2)n} \right\} , \tag{16a}$$

where β_n is a slowly varing function of n, of order of magni-

tude one. Equation (16a), being valid for $m_j \neq -2, -4$, requires the replacements

$$\tfrac{1}{2} m_j(m_j+4)\cdots \to \delta_n = \frac{(2n)!}{(n!)^2(2n-1)4^{2n}}, \quad \text{for } m_j = -2, \tag{16b}$$

$$[1 + \Sigma a_{nk}(z-z_j)^k]\cdots \to (z-z_j)^{2n+1} \sum_{k=0}^{\infty} a_{nk}(z-z_j)^k, \tag{16c}$$

for $m_j = -4$; a_k and a_{nk} in Eqs. (16a) and (16c) depend on g_k in Eq. (8a), and vanish for $g_k \equiv 0$.

Inserting Eqs. (16a)–(16c), and similar expressions for $\tilde{q}_{2N+1}$, into Eq. (15a) and taking the limit $z \to z_j$, we again arrive at Eq. (9) derived in the lowest order for $m_j \neq -2$, and at a constraint on $\tilde{z}_i, \ldots$, for $m_j = -2$, which now reads

$$c_j \left(1 - \sum_{n=1}^{N} \frac{\delta_n}{c_j{}^n}\right)^2 = \tilde{c}_j \left(1 - \sum_{n=1}^{N} \frac{\delta_n}{\tilde{c}_j{}^n}\right)^2, \quad \text{for } m_j = -2. \tag{17}$$

The unknown parameters in Eq. (6), $\tilde{z}_i$ etc. [which are also contained in $\tilde{c}_j$ in Eqs. (9) and (17)], should be determined from Eq. (15b). They will, in general, be different in different orders. Note in this connection that we have a freedom in choice of the integration constant in Eq. (3a). Therefore, for any given point z_0, we can prescribe $\tilde{z}_0 = \tilde{z}(z_0)$. In particular we can always make a convenient choice for one of the points $\tilde{z}_j$. Also the unit length in the $\tilde{z}$ plane can be prescribed, for example by fixing the multiplication constant in $\tilde{f}(\tilde{z})$. However, the sign of $\tilde{f}(\tilde{x})$ must be the same as that of $f(x)$, in order that the signs of $Q^2(x)$ and $\tilde{Q}^2(\tilde{x})$ be the same. If the signs of the square roots $Q(x)$ and $\tilde{Q}(\tilde{x})$ are chosen in the same way, positive sense of the x axis will be invariant under the mapping, i.e.,

$$q(x) \equiv \frac{d\tilde{x}}{dx} > 0. \tag{18}$$

In lowest order ($N = 0$) the integrals in Eq. (15b) are convergent at zeros and simple poles of $Q^2(z)$, and some equations for $\tilde{z}_i$, etc., can be obtained by integrating between these points. In other situations, one should instead introduce appropriate contour integrals. For a pair of odd order zeros and/or poles of $Q^2(z)$, z_j, and z_l, we define

$$\Gamma_{2N+1}(j,l) = g \oint_C q_{2N+1}(z)\, dz, \quad m_j, m_l = 2m-1, \tag{19}$$

(and similar expression for $\tilde{\Gamma}_{2N+1}$ in the $\tilde{z}$ plane, $\tilde{g} = g$), where the integration contour C encloses z_j and z_l, but no other zero or pole of $Q^2(z)$, and g is a normalization constant; $q_{2N+1}(z)$ is assumed to be single valued along C, i.e., a cut connecting z_j and z_l is implied. The Γ-integral (19) is independent of C. In particular, integrating along the two lips of the cut, z' and z'', we obtain

$$\Gamma_{2N+1}(j,l) = 2g \int_{z_j}^{z_l} q_{2N+1}(z')\, dz', \tag{19'}$$

if the integral in Eq. (19') is finite. Therefore, $g = \frac{1}{2}$ (or $g = 1/2i$) is often a convenient choice. Equation (19) remains meaningful when z_j and z_l merge into an even-order zero or pole z_j. In that case $q_{2N+1}(z)$ becomes single valued around z_j, and we can write

$$\Gamma_{2N+1}(j) \equiv \Gamma_{2N+1}(j,j) = \pm 2\pi i g \operatorname{Res} q_{2N+1}(z_j), \quad m_j = 2m. \tag{20}$$

From Eq. (15b) we obtain

$$\Gamma_{2N+1}(j,l) = \tilde{\Gamma}_{2N+1}(j,l), \tag{21}$$

where the contours C and $\tilde{C}$ should be circled in the same direction, if condition (18) is fulfilled.

Our initial assumption that the zero or pole z_j is mapped into the same type of zero or pole $\tilde{z}_j$ was crucial in the derivation of Eq. (9), both in first and higher orders. However, for simple models this requirement may be inconsistent with Eq. (21), if the order of z_j is even. This simply means that the mergence of z_j and z_l into an even-order zero or pole z_j (with $\Gamma_{2N+1} = \tilde{\Gamma}_{2N+1}$) is not necessarily accompanied by the similar mergence in the $\tilde{z}$ plane. For example, if $\tilde{Q}^2(\tilde{z}) = \text{const}(\tilde{z} - \tilde{z}_j)^{m_j} (m_j = 2m \neq -2)$, one obtains $\operatorname{Res} \tilde{q}_{2N+1}(\tilde{z}_j) = 0$, whereas in general $\operatorname{Res} q_{2N+1}(z_j) \neq 0$, in contradiction to Eq. (21). For that reason, to be consistent with Eq. (21), we now admit for the even-order zero or pole z_j being represented, if necessary, by a pair of the neighboring odd-order zeros and/or poles, $\tilde{z}_{j'}$ and $\tilde{z}_{j''}$, such that $m_{j'} + m_{j''} = m_j$ (and similarly for the inverse mapping $\tilde{z} \to z$). The usefulness of this extension, in spite of the fact that $q(z)$ may now be singular at z_j, is demonstrated in Ref. 14.

In applications $Q^2(z)$ [or $\tilde{Q}^2(\tilde{z})$] is often a rational function

$$Q^2(z) = c \prod_{i=1}^{M} (z - z_j)^{m_i} = c z^{m_\infty}[1 + O(z^{-1})],$$
$$m_\infty = \sum_{i=1}^{M} m_i, \tag{22}$$

where $c = \text{const}$, and $m_i = \text{integer} \neq 0$. In that case it can be shown using the results of Ref. 3 (Secs. 3 and 4) that $Y_{2n}(z)$ is also a rational function, of the following form:

$$Y_{2n}(z) = \frac{1}{c^n} P_{l_n}(z) \prod_{i=1}^{M} (z - z_i)^{\mu_{in}}, \tag{23a}$$

where

$$l_n = 2(M-1)n, \quad \mu_{in} = -(m_i + 2)n, \tag{23b}$$

and $P_{l_n}(z)$ is a polynomial of l_n th degree, independent of c. Equations (23b) assume all $m_i \neq -4$, and $m_\infty \neq 0, -4$; for $m_i = -4$ $\mu_{in} = 2n + 1$, and l_n should be decreased by one; l_n should also be decreased if $m_\infty = 0$ or -4, in general by one, or by two if $Q^2(z)$ is even. [Functions $Y_{2n}(z)$ are even if $Q^2(z)$ is even.[3]]

III. SUMMARY OF RESULTS AND CONCLUSIONS

The wave equation with cutoffs and resonances can be transformed to a simpler equation with similar properties by using the mapping $z \to \tilde{z}$ defined by Eq. (15b). If the model equation (2) is analytically solvable, all interesting quantities such as reflection and absorption coefficients etc., can be expressed in terms of parameters entering the model, $\tilde{z}_i$ etc., see Eq. (6). This is illustrated in Ref. 14. One of the points $\tilde{z}_i$,

and the multiplication constant in $\tilde{f}(\tilde{z})$ can always be prescribed, and the remaining parameters should be determined from Eqs. (21). However, if their number, n_p, is greater than the number of equations (21), n_e, $(n_p - n_e)$ parameters can also be chosen for the best fit.

For a given model the rhs of Eq. (21) is a uniquely defined function of the parameters $\tilde{z}_i$ etc., which in simpler cases can effectively be determined in arbitrary order. In more complicated situations one can use a computer to perform necessary algebraic manipulations, and tabulate the results for $N = 0,1,2,...$, for subsequent applications. The integrals on the lhs of Eq. (21), pertinent to the given wave equation (1), are the same quantities which are dealt with in ordinary higher-order phase-integral approximations.

[1]N. Fröman, Ark. Fys. **32**, 541 (1966).
[2]Ö. Dammert and P. O. Fröman, J. Math. Phys. **21**, 1683 (1980).
[3]A. A. Skorupski, Rep. Math. Phys. **17**, 161 (1980).
[4]N. Fröman and P. O. Fröman, Ann. Phys. (N.Y.) **83**, 103 (1974).
[5]N. Fröman and P. O. Fröman, Nuovo Cimento B **20**, 121 (1974).
[6]W. H. Furry, Phys. Rev. **71**, 360 (1947).
[7]N. Fröman and P. O. Fröman, *JWKB-Approximation, Contributions to the Theory* (North-Holland, Amsterdam, 1965).
[8]N. Fröman, Ann. Phys. (N.Y.) **61**, 451 (1970).
[9]S. G. Miller, Jr., and R. H. Good, Jr., Phys. Rev. **91**, 174 (1953).
[10]R. B. Dingle, Appl. Sci. Res. Sect. B **5**, 345 (1956).
[11]P. Lu and E. M. Measure, Phys. Rev. D **5**, 2514 (1972).
[12]S. S. Wald and P. Lu, Phys. Rev. D **9**, 895 (1974).
[13]S. S. Wald and P. Lu, Phys. Rev. D **9**, 2254 (1974).
[14]A. A. Skorupski, J. Math. Phys. **25**, 2655 (1984).
[15]C. E. Hecht and J. E. Mayer, Phys. Rev. **106**, 1156 (1957).
[16]J. A. Campbell, J. Comput. Phys. **10**, 308 (1972).

Transmission through cutoffs and resonances in the double phase-integral approximation

Andrzej A. Skorupski
Institute of Nuclear Research, Hoża 69, 00-681 Warsaw, Poland

(Received 23 August 1983; accepted for publication 20 January 1984)

Using the double phase-integral approximation technique developed earlier for the wave equation $d^2\psi/dz^2 + Q^2(z)\,\psi = 0$, we derive analytical formulas for the reflection (R), transmission (T), and absorption (A) coefficients. They are valid to arbitrary order in the expansion parameter, for functions $Q^2(z)$ having either two cutoffs or one cutoff and one resonance. For two examples of this type the formulas for R, T, and A are checked against numerical results, using approximations up to fifth order.

PACS numbers: 03.40.Kf, 02.30.Mv, 02.30.Qy

I. INTRODUCTION

The time-independent wave equation in one dimension,

$$\frac{d^2\psi}{dz^2} + Q^2(z)\,\psi = 0, \quad z = x + iy, \tag{1}$$

can be reduced to simpler equations of the same form (models) by using the double phase-integral approximations[1] (DPIA). In this paper we illustrate the general theory of DPIA developed in Ref. 1 by considering three models solvable analytically:

(i) Parabolic model[2] (Sec. II),
(ii) Epstein's model[3] (Sec. III),
(iii) Budden's model[4] (Sec. IV).

Quantities referring to a model (such as ψ, z, etc.) will be distinguished from those for other functions $Q^2(z)$ by a tilde (just as in Ref. 1). Models (i) to (iii) have some common features. Thus $\tilde{Q}^2(\tilde{x})$ is real, and becomes positive when $|\tilde{x}|$ is large enough; $\tilde{\psi}(\tilde{z})$ involves special functions, but one can always identify solutions which asymptotically represent propagating waves (for $\tilde{x} \to \pm\infty$); the reflection and transmission coefficients, R and T, can be determined analytically, and are simple elementary functions of the model parameters. Models (i) and (ii) conserve the wave energy (or the current in the quantum mechanical language), i.e., $R + T = 1$, which is a consequence of $\tilde{Q}^2(\tilde{x})$ being both real and regular. In Budden's model $\tilde{Q}^2(\tilde{x})$ has a simple pole at $\tilde{x} = \tilde{p}$ (resonance), where part of the wave energy is absorbed ($R + T < 1$). In such cases one defines the absorption coefficient A, so as to satisfy $R + T + A = 1$. [A pole on the real axis is obviously an idealization; it should be properly bypassed in the complex plane when tracing $\psi(x)$ between $x = \pm\infty$; for $Q^2(x)$ resembling that of Budden's model, see, e.g., Eq. (27), $A > 0$ corresponds to tracing in the lower half-plane, if the time dependence is $\exp(-i\omega t)$.] A basic concept in the DPIA is the Γ-integral [see Eq. (19) in Ref. 1] which can be associated with any pair of odd-order zeros and/or poles of $Q^2(z)$, located either on or close to the x axis (cutoffs or resonances). In the simpler situations dealt with in this paper, there is only one such pair, z_1 and z_2. The corresponding Γ-integral can be written (with convenient normalization)

$$\Gamma_{2N+1} = \frac{1}{i}\oint_C Q(z)\,[1 + Y_2(z) + \cdots + Y_{2N}(z)]\,dz, \tag{2}$$

where C is assumed to encircle z_1 and z_2 in a positive sense; $z_{1,2}$ are either real ($z_1 < z_2$) or complex conjugate ($\mathrm{Im}\, z_2 > 0$); the sign of Q will always be chosen so that $Q(x) > 0$ for $x > \mathrm{Re}\, z_2$. According to the general theory of the DPIA[1] we make the model, $\tilde{Q}^2(\tilde{z})$, pertinent to other functions $Q^2(z)$ [which should resemble $\tilde{Q}^2(\tilde{z})$ on the real axis, see Sec. V for more details] by requiring $\tilde{\Gamma}_{2N+1} = \Gamma_{2N+1}$. In view of this basic relation we will use the same symbol Γ_{2N+1} in connection with both the given function $Q^2(z)$ and the model. This will make the formulas for R, etc., derived for the model, automatically applicable (approximately) to $Q^2(z)$ provided the model parameters in these formulas are expressed in terms of Γ_{2N+1}. Such formulas as following from models (i) to (iii) are given in Secs. II to IV. They are then checked in Sec. VI against numerical results in two situations for which the Γ-integrals can be calculated analytically. In Sec. V, we formulate the applicability conditions for our formulas for R, etc.

II. R and T FROM THE PARABOLIC MODEL

The essential results of Ref. 2 can be summarized as follows:

$$\tilde{Q}^2(\tilde{z}) = \tilde{c}(\tilde{z}^2 - \tilde{z}_1^2), \quad \tilde{c} > 0 \quad \text{(cutoffs at } \pm\tilde{z}_1\text{)}, \tag{3a}$$

$$\tilde{\psi}(\tilde{z}) = \text{parabolic cylinder (Weber's) function}, \tag{3b}$$

$$T = [1 + \exp(\pi\tilde{c}^{1/2}\tilde{z}_1^2)]^{-1}. \tag{3c}$$

To calculate Γ_{2N+1} we first use Eqs. (23a), (23b) of Ref. 1 ($M = m_\infty = 2$) to find the general form of $\tilde{Y}_{2n}(\tilde{z})$, and then expand $\tilde{Y}_{2n}(\tilde{z})$ and $\tilde{Q}(\tilde{z})$ in a Laurent series convergent for $|\tilde{z}| > |\tilde{z}_1|$:

$$\tilde{Y}_{2n}(\tilde{z}) = \tilde{c}^{-n} P_n(\tilde{z}^2)(\tilde{z}^2 - \tilde{z}_1^2)^{-3n} = \tilde{z}^{-4n}\sum_{k=0}^{\infty} a_{nk}\,\tilde{z}^{-k}, \tag{4a}$$

$$\tilde{Q}(\tilde{z}) = \tilde{c}^{1/2}\left(\tilde{z} - \frac{\tilde{z}_1^2}{2\tilde{z}} - \frac{\tilde{z}_1^4}{8\tilde{z}^3} - \cdots\right). \tag{4b}$$

Inserting Eqs. (4a), (4b) into Eq. (2) we obtain

$$\Gamma_{2N+1} = \frac{1}{i}\oint_C \tilde{Q}(\tilde{z})\,d\tilde{z} = -\pi\tilde{c}^{1/2}\tilde{z}_1^2, \tag{5}$$

and we can finally write Eq. (3c) as ($R = 1 - T$)

$$T = [1 + \exp(-\Gamma_{2N+1})]^{-1}. \tag{6}$$

 0022-2488/84/092655-07$02.50

In view of the DPIA arguments given in Sec. I, Eq. (6) being exact for the parabolic model, should be approximately valid also for other functions $Q^2(z)$ satisfying the applicability conditions formulated in Sec. V. For $N=0$, Eq. (6) reduces to one of the well-known Kemble expressions for T [see, e.g., Eqs. (9.2) and (9.3) in Ref. 5]. For $N>0$, Eq. (6) is not new either, but its direct derivation without referring to the parabolic model is more complicated.[6] First numerical check of Eq. (6) in higher orders ($N\leqslant 5$) was given by Karlsson,[7] who used for this purpose the analytically solvable Epstein model discussed in the following section.

III. R AND T FROM EPSTEIN'S MODEL

In Epstein's model,[3] $\tilde{Q}^2(\tilde{z})$ can be written

$$\tilde{Q}^2(\tilde{z}) = P_2(u) \equiv au^2 + bu + c, \tag{7a}$$

$$u = \tanh(\tilde{z}/L), \quad L>0, \tag{7b}$$

and $\tilde{\psi}(\tilde{z})$ is expressible in terms of the hypergeometric function. Thus $\tilde{Q}^2(\tilde{x})$ has finite limits at $\tilde{x} = \pm\infty$, $\tilde{Q}^2_{\pm}$, which we assume positive, and is additionally characterized by $\tilde{Q}^2_0 = \tilde{Q}^2(\tilde{x}=0)$:

$$a = \tfrac{1}{2}(\tilde{Q}^2_+ + \tilde{Q}^2_-) - \tilde{Q}^2_0,$$

$$b = \tfrac{1}{2}(\tilde{Q}^2_+ - \tilde{Q}^2_-), \quad c = \tilde{Q}^2_0. \tag{7c}$$

The constant L in Eq. (7b) defines the unit length in the $\tilde{z}$ plane and can be fixed in a convenient way. Thus we put $L=1$, which simplifies the algebra. With this choice one obtains[3] ($R = 1 - T$)

$$T = (S_+ - S_-)/(C + S_+), \tag{7d}$$

where

$$S_{\pm} = \sinh^2[(\pi/2)(\tilde{Q}_+ \pm \tilde{Q}_-)], \tag{7e}$$

$$C = \cosh^2[\pi(a - \tfrac{1}{4})^{1/2}]. \tag{7f}$$

Equation (7f) is valid for any a, including $a<\frac{1}{4}$ or $a<0$. However, in our analysis we assume $a>0$. In that case the two roots of the quadratic $P_2(u)=0$, u_1 and u_2, lie either on or close to the segment $-1<u<1$, which is the image of the $\tilde{x}$ axis [$P_2(u=\pm 1) = \tilde{Q}^2_{\pm} > 0$]. The corresponding zeros of $\tilde{Q}^2(\tilde{z})$ located either on or close to the $\tilde{x}$ axis (cutoffs) will be denoted by $\tilde{z}_1$ and $\tilde{z}_2$. [There are more zeros of $\tilde{Q}^2(\tilde{z})$, which is periodic with period $i\pi L$.] The Γ-integral associated with $\tilde{z}_{1,2}$ can be written

$$\frac{\Gamma_{2N+1}}{2\pi} = \frac{\tilde{Q}_+ + \tilde{Q}_-}{2} - a^{1/2}\left(1 - \frac{1}{8a} - \frac{1}{128a^2} - \frac{1}{1024a^3} - \cdots - \frac{\delta_N}{a^N}\right), \tag{8}$$

where δ_N is given by Eq. (A3) in Appendix A. This result can be derived by using Eqs. (8) and (9) of Ref. 7. It also follows easily if one knows the general form of $\tilde{Y}_{2n}(u)$, see Appendix A. Inserting in Eq. (8) Γ_{2N+1} corresponding to a given $Q^2(z)$ we obtain one equation for three model parameters. Two constraints can thus be imposed for uniqueness. For example, one can define

$$\beta_i = (z_2 - z_1)^3 \left.\frac{dQ^2}{dz}\right|_{z_i} \tag{9}$$

and require

$$\tilde{\beta}_i = \beta_i, \quad i = 1,2. \tag{10}$$

This will guarantee $d\tilde{Q}^2/d\tilde{z}$ at $\tilde{z}_i$ to be equal to dQ^2/dz at z_i, if L in Eq. (7b) is chosen so that $\tilde{z}_2 - \tilde{z}_1 = z_2 - z_1$. [$\beta_i$ defined by Eq. (9) is the simplest dimensionless combination of $z_2 - z_1$ and dQ^2/dz at z_i.]

For symmetric propagation [$Q^2(-z) \equiv Q^2(z)$] it is natural to impose the same symmetry upon the model. The constraints in that case might be (cutoffs at $\pm z_1$)

$$\tilde{\beta}_1 = \beta_1 \equiv z_1^3 \left.\frac{dQ^2}{dz}\right|_{z_1}, \tag{11a}$$

or

$$\tilde{\beta}_0 = \beta_0 \equiv -z_1^2\, Q_0^2, \tag{11b}$$

where

$$\tilde{\beta}_1 = [f(|u_1|)]^3\, 2\tilde{Q}^2_{\infty}\, |u_1|, \tag{12a}$$

$$\tilde{\beta}_0 = [f(|u_1|)]^2\, |\tilde{Q}^2_{\infty} - a|, \tag{12b}$$

$\tilde{Q}^2_{\infty} = \tilde{Q}^2_{\pm}$, $f = \tanh^{-1}$ for $\tilde{Q}^2_0 \leqslant 0$, or $f = \tan^{-1}$ for $\tilde{Q}^2_0 > 0$, $a = \tilde{Q}^2_{\infty} - \tilde{Q}^2_0$, and $|u_1| = |1 - \tilde{Q}^2_{\infty}/a|^{1/2}$. Equation (11a) is equivalent to Eq. (10), whereas (11b) implies $\tilde{Q}^2_0 = Q^2_0$, if L in Eq. (7b) is chosen so that $\tilde{z}_1 = z_1$.

Thus to determine R and T from Epstein's model the model parameters $\tilde{Q}_{\pm}$ and a should first be found from Eq. (8) plus constraints, to be used afterwards in Eqs. (7d) to (7f).

IV. R, T, AND A FROM BUDDEN'S MODEL

Summary of results of Ref. 4:

$$\tilde{Q}^2(\tilde{z}) = \tilde{K}^2\tilde{z}/(\tilde{z} - \tilde{p}), \quad \tilde{K}^2 > 0, \ \tilde{p} > 0, \tag{13a}$$

$\tilde{\psi}(\tilde{z})$ is expressible in terms of the confluent hypergeometric function, and the reflection, transmission, and absorption coefficients are

$$R_0 = (1-T)^2, \quad T_0 = T, \quad A_0 = T(1-T), \tag{13b}$$

$$R_{\infty} = 0, \quad T_{\infty} = T, \quad A_{\infty} = 1 - T, \tag{13c}$$

where

$$T = \exp(-\pi\tilde{K}\tilde{p}), \tag{13d}$$

and the subscript 0 refers to propagation towards the zero of $\tilde{Q}^2(\tilde{x})$ (i.e., from $x=-\infty$ to $x=\infty$), and the subscript ∞ to propagation towards the pole (from ∞ to $-\infty$).

Using Eqs. (23a), (23b) of Ref. 1 ($M=2$, $m_{\infty}=0$) to find a general form of $\tilde{Y}_{2n}(\tilde{z})$, and expanding $\tilde{Y}_{2n}(\tilde{z})$ and $\tilde{Q}(\tilde{z})$ in a Laurent series convergent for $|\tilde{z}| > \tilde{p}$ we easily find, as in the case of Eq. (5),

$$\Gamma_{2N+1} = \pi\tilde{K}\tilde{p}. \tag{14}$$

Using this result in Eq. (13d) we get

$$T = \exp(-\Gamma_{2N+1}), \tag{15}$$

which, if inserted into Eqs. (13b), (13c), gives explicit expressions for R_0 etc.

V. APPLICABILITY CONDITIONS

The approximate formulas for R, T, and A, derived in Secs. II to IV, are applicable to functions $Q^2(z)$ which resemble one of the models (i) to (iii), in the following sense:

(a) $Q^2(z)$ should have the same type of zeros and poles on

or close to the real axis, i.e., two cutoffs for models (i) and (ii), or one real cutoff and one real resonance for model (iii).

(b) $Q^2(z)$ should admit propagating waves at $x \to \pm\infty$. The meaning of (b) and the definition of R and T are obvious if $Q^2(x) \to \text{const} > 0$ as $x \to \pm\infty$. A more general situation is described in Ref. 5, Chap. 9, in connection with the potential barrier penetration. Thus one can define R and T uniquely by using the JWKB approximation (or higher-order phase-integral approximations) at large distances from the origin, $|x| > x_0$, if $Q^2(x) > 0$ there, and these approximations tend to exact solutions of the wave equation as $x \to \pm\infty$. This in turn requires the quantity called the μ-integral to remain finite with the integration intervals extending to infinity ($-\infty < x < -x_0$, and $x_0 < x < \infty$). Approximate formulas for such μ-integrals are derived in Ref. 8 under the assumption that $Q^2(z)$ has an isolated singular point at $z = \infty$, but not an essential singularity; this allows for the Laurent expansion about infinity, of the form

$$Q^2(z) = cz^{m_\infty}\left[1 + \sum_{k=1}^{\infty} g_k\, z^{-k}\right]. \tag{16}$$

In that case the μ-integrals in question are only finite for $m_\infty \geqslant -1$ [see Eqs. (64) and (65) in Ref. 8]. Thus to comply with $Q^2(x) > 0$ for $|x| > x_0$ we must require

$$c > 0, \quad m_\infty = 0,2,4,\cdots. \tag{17}$$

This includes $Q^2(x) \to \text{const} > 0$ ($m_\infty = 0$), but additionally also $Q^2(x) \to +\infty$, as $x \to \pm\infty$.

VI. EXAMPLES

Formulas for R, T, and A, derived in Secs. II to IV, are particularly useful if the Γ-integral (2) can be calculated analytically. For example, if $Q^2(z)$ is a rational function, Γ_{2N+1} is often an elliptic or even elementary integral. These two possibilities will be illustrated by two examples given in this section. However, the amount of algebra needed in higher orders is often prohibitively large even if the final results are not too complicated. Such cases can only be treated by performing part of the algebra automatically by computer. In our examples we searched for an effective compromise between manipulations done by hand, and symbolic computations. This methodology, described in Appendices B and C, should be useful also in other, more physical, applications.

Our first example,

$$Q^2(z) = Q_\infty^2 - (Q_\infty^2 - Q_0^2)/(z^2+1),$$
$$Q_\infty^2 > 0, \quad Q_0^2 < Q_\infty^2, \tag{18}$$

describes symmetric propagation with two cutoffs (or transmission through a symmetric potential barrier). The applicability condition (17) is obviously fulfilled, and one can use Eq. (6) or (7d) to calculate T and R. We assume $Q_0^2 < Q_\infty^2$ to prevent the zeros of $Q^2(z)$, $\pm z_1$, from getting too far from the x axis (i.e., beyond the poles at $\pm i$). In any order the Γ-integral (2) associated with $\pm z_1$ can be expressed in terms of the complete elliptic integrals $K(k)$ and $E(k)$ (see Appendix B), e.g.,

$$\Gamma_1 = 4Q_\infty\,[E(k) - k'^2 K(k)]$$
$$= Q_\infty \pi k^2(1 + k^2/8 + \cdots), \tag{19a}$$
$$\Gamma_3 - \Gamma_1 = (6Q_\infty k^2 k'^2)^{-1}$$
$$\times[(1+7k^2)E(k) - (1+3k^2)k'^2 K(k)]$$
$$= (3\pi/8Q_\infty)(1 + \tfrac{9}{8}k^2 + \cdots), \tag{19b}$$
$$\Gamma_5 - \Gamma_3 = (1440Q_\infty^3 k^6 k'^6)^{-1}$$
$$\times[(56 - 265k^2 + 459k^4 - 1243k^6 - 31k^8)$$
$$\times E(k) - (56 - 237k^2 + 351k^4$$
$$- 727k^6 + 45k^8)K(k)]$$
$$= (135\pi/1024Q_\infty^3)(1 + \tfrac{545}{144}k^2 + \cdots), \tag{19c}$$

where

$$k^2 = -z_1^2 = Q_0^2/Q_\infty^2, \quad k'^2 = 1 - k^2. \tag{20}$$

The power series in k^2 in Eqs. (19a)–(19c) gives the behavior of the Γ-integrals in the limit $|Q_0^2| \ll Q_\infty^2$. For $Q_0^2 < 0$ (i.e., $k^2 < 0$) we can use the imaginary modulus transformations (see Eqs. 160.2 in Ref. 9)

$$K(k) = k_1' K(k_1), \quad E(k) = E(k_1)/k_1',$$
$$k_1^2 = -k^2/k'^2, \quad k_1'^2 = 1/k'^2. \tag{21}$$

For $Q_0^2 = 0$ (double zero, $z_1^2 = k^2 = 0$) we obtain $\Gamma_1 = 0$, and $\Gamma_{2N+1} > 0$ for $N > 0$, if Q_∞^2 is not too small, e.g.,

$$\Gamma_3 = 3\pi/8Q_\infty > 0,$$
$$\Gamma_5 = \Gamma_3(1 - 45/128Q_\infty^2) > 0, \quad \text{if } Q_\infty^2 > \tfrac{45}{128}, \text{ etc.} \tag{22}$$

This means, in view of Eq. (5), that in the parabolic model the double zero, $z_1^2 = 0$, is represented in first order also by the double zero, $\tilde{z}_1^2 = 0$, but in higher orders by two simple pure imaginary zeros ($\tilde{z}_1^2 < 0$). In Epstein's model with the constraint (11a) or (11b) the double zero in question is represented, in any order, by the double zero. This is a consequence of the constraints which now read

$$[f(|u_1|)]^3 |u_1|\, \tilde{Q}_\infty^2 = Q_0^4/(Q_\infty^2 - Q_0^2), \tag{23a}$$
$$[f(|u_1|)]^2 |\tilde{Q}_\infty^2 - a| = Q_0^4/Q_\infty^2. \tag{23b}$$

[Inserting in Eq. (23a) or (23b) $Q_0^2 = 0$ we obtain $u_1 = 0$, which is equivalent to $\tilde{z}_1^2 = 0$.] In general, calculating $\tilde{Q}_\infty$ [$= (\tilde{Q}_+ + \tilde{Q}_-)/2$] from Eq. (8) and inserting it into Eq. (23a) or (23b), a nonlinear equation for a is obtained, which can easily be solved numerically. For $Q_0^2 = 0$ the constraints (23a), (23b) reduce to

$$a = \tilde{Q}_\infty^2. \tag{24}$$

Inserting Eq. (24) into Eq. (8) we arrive at an identity for $N = 0$, but in higher orders we obtain ($N > 0$)

$$\frac{1}{8\tilde{Q}_\infty} + \frac{1}{128\tilde{Q}_\infty^3} + \frac{1}{1024\tilde{Q}_\infty^5} + \cdots + \frac{\delta_N}{\tilde{Q}_\infty^{2N-1}} = \frac{\Gamma_{2N+1}}{2\pi}. \tag{25}$$

Equation (25) requires $\Gamma_{2N+1} > 0$, see Eqs. (22). For $N = 0$ we can only go to the limit $Q_0^2 \to 0$. Eliminating $a (= \tilde{Q}_\infty^2 - \tilde{Q}_0^2)$ from Eqs. (8) and (23a), (23b), and using the

TABLE I. Transmission coefficients for $Q^2(z)$ given by Eq. (18), for $Q_0^2 = 0$; numerical (T) versus the DPIA results up to fifth order (T_1, T_3, T_5), first line corresponding to the parabolic model, Eq. (6), and second line to Epstein's model, Eq. (7d).

Q_∞^2	T	T_1	T_3	T_5
1.0	0.731 938 180 3	0.5	0.76	0.68
		0.78	0.78	0.68
3.0	0.652 579 142 5	0.5	0.664	0.646
		0.671	0.671	0.646
10.0	0.589 493 844 5	0.50	0.592	0.588 9
		0.593	0.593	0.588 9
30.0	0.552 992 821 0	0.50	0.553 6	0.552 94
		0.553 8	0.553 8	0.552 95
100.0	0.529 318 032 6	0.50	0.529 4	0.529 315
		0.529 5	0.529 5	0.529 315
300.0	0.516 978 101 1	0.50	0.517 00	0.516 977 9
		0.517 01	0.517 01	0.516 977 9
1000.0	0.509 309 335 8	0.50	0.509 313	0.509 309 326
		0.509 314	0.509 314	0.509 309 327
3000.0	0.505 376 417 1	0.50	0.505 377 0	0.505 376 416 3
		0.505 377 3	0.505 377 3	0.505 376 416 3
10 000.0	0.502 945 105 8	0.500	0.502 945 21	0.502 945 105 5
		0.502 945 25	0.502 945 25	0.502 945 105 5

power expansions in $\tilde{Q}_0^2/\tilde{Q}_\infty^2$ and Q_0^2/Q_∞^2 we easily find for either of the constraints (23a), (23b)

$$\tilde{Q}_\infty(Q_0^2 \to 0, N=0) = \tilde{Q}_\infty(Q_0^2 = 0, N=1) = \tfrac{2}{3} Q_\infty . \tag{26}$$

This indicates that the results obtained from Epstein's model with the adopted constraints, in first order, tend to those in third order as $Q_0^2 \to 0$. One should therefore expect Epstein's model to be particularly good in first order.

In Tables I and II, we tabulate typical results for T from the parabolic and Epstein models versus results obtained numerically. In first order, as expected, the Epstein model is significantly better than the parabolic one, while in higher orders the difference between them is insignificant. Thus in this application the extra freedom (but also extra complication) in the Epstein model are not paid for in higher orders; this, perhaps, would be the case if the propagation, unlike the

TABLE II. Transmission coefficients as in Table I but for $Q_0^2 \neq 0$; $Q_\infty^2 = 10^3$. First line corresponds to the parabolic model, Eq. (6), second line to Epstein's model, Eq. (7d), with the constraint (23a), and third line to Epstein's model with the constraint (23b).

Q_0^2	T	T_1	T_3	T_5
100.0	0.999 959 151 0	0.999 957	0.999 959 151 5	0.999 959 150 7
		0.999 959 10	0.999 959 151 8	0.999 959 150 7
		0.999 959 08	0.999 959 151 8	0.999 959 150 7
10.0	0.737 375 081 4	0.730	0.737 378	0.737 375 073
		0.737 357	0.737 379	0.737 375 074
		0.737 350	0.737 379	0.737 375 074
1.0	0.534 107 412 5	0.525	0.534 111	0.534 107 403
		0.534 109	0.534 112	0.534 107 403
		0.534 108 3	0.534 112	0.534 107 403
− 1.0	0.484 471 584 8	0.475	0.484 475	0.484 471 575
		0.484 479	0.484 476	0.484 471 576
		0.484 480	0.484 476	0.484 471 576
− 10.0	0.277 801 249 0	0.270	0.277 804	0.277 801 242
		0.277 83	0.277 805	0.277 801 242
		0.277 83	0.277 805	0.277 801 242
− 100.0	$0.564\,896\,364\,1\times10^{-4}$	0.546×10^{-4}	$0.564\,901\times10^{-4}$	$0.564\,896\,352\times10^{-4}$
		$0.565\,5\times10^{-4}$	$0.564\,904\times10^{-4}$	$0.564\,896\,353\times10^{-4}$
		$0.565\,7\times10^{-4}$	$0.564\,904\times10^{-4}$	$0.564\,896\,353\times10^{-4}$
− 1000.0	$0.787\,227\,218\times10^{-39}$	0.77×10^{-39}	$0.787\,227\,4\times10^{-39}$	$0.787\,227\,212\times10^{-39}$
		0.791×10^{-39}	$0.787\,228\,2\times10^{-39}$	$0.787\,227\,212\times10^{-39}$
		0.792×10^{-39}	$0.787\,228\,3\times10^{-39}$	$0.787\,227\,212\times10^{-39}$

parabolic model, was asymmetric. Tables I and II indicate that in fifth order both models give excellent accuracy if Q^2_∞ is large enough, the accuracy only weakly depending on Q^2_0. Note that for $Q^2_\infty = 1$ (Table I) the parabolic model gives the best accuracy in the third rather than fifth order. The existence of such an optimum order is typical of the PIA, but the optimum is usually far beyond the practically interesting orders. Different behavior is only possible if the first order accuracy is rather poor,[8] which in our example corresponds to $Q^2_\infty \lesssim 1$.

TABLE III. Reflection, transmission, and absorption coefficients (R, T, and A) for $Q^2(z)$ given by Eq. (27); numerical results followed by the DPIA results in order 1, 3, and 5, for propagation towards the cutoff (first four lines), and towards the resonance.

K	p	R	T	A
1.0	0.3	0.550 309 14	0.272 873 88	0.176 816 98
		0.40	0.37	0.23
		0.58	0.24	0.181
		0.9	0.04	0.04
		0.032 264 53	0.272 873 88	0.694 861 60
		0	0.37	0.63
		0	0.24	0.76
		0	0.04	0.9
3.0	0.1	0.406 699 20	0.362 300 65	0.231 000 14
		0.38	0.39	0.237
		0.400	0.368	0.233
		0.410	0.360	0.230 4
		0.000 235 11	0.362 300 65	0.637 464 24
		0	0.39	0.61
		0	0.368	0.632
		0	0.360	0.640
10.0	0.03	0.375 099 24	0.387 546 38	0.237 354 38
		0.373	0.389	0.237 8
		0.375 00	0.387 62	0.237 37
		0.375 087	0.387 556	0.237 356
		0.2×10^{-11}	0.387 546 38	0.612 453 62
		0	0.389	0.610 5
		0	0.387 62	0.612 37
		0	0.387 556	0.612 443
30.0	0.01	0.372 791 596 6	0.389 433 381 4	0.237 775 022 0
		0.372 5	0.389 6	0.237 82
		0.372 790 5	0.389 434 2	0.237 775 2
		0.372 791 58	0.389 433 391	0.237 775 025
		$\lesssim 10^{-17}$	0.389 433 381 4	0.610 566 618 7
		0	0.389 6	0.610 36
		0	0.389 434 2	0.610 565 8
		0	0.389 433 391	0.610 566 609
100.0	0.003	0.372 538 471 4	0.389 640 703 6	0.237 820 825 0
		0.372 52	0.389 66	0.237 825
		0.372 538 463	0.389 640 710	0.237 820 827
		0.372 538 471 4	0.389 640 703 0	0.237 820 825 6
		$\lesssim 10^{-17}$	0.389 640 703 6	0.610 359 296 4
		0	0.389 66	0.610 34
		0	0.389 640 710	0.610 359 290
		0	0.389 640 703 0	0.610 359 297 0
30.0	0.002	0.029 543 868 58	0.828 116 701 7	0.142 339 429 7
		0.029 51	0.828 20	0.142 28
		0.029 543 7	0.828 117 1	0.142 339 2
		0.029 543 867	0.828 116 705	0.142 339 428
		$\lesssim 10^{-17}$	0.828 116 701 7	0.171 883 298 3
		0	0.828 20	0.171 80
		0	0.828 117 1	0.171 882 9
		0	0.828 116 705	0.171 883 295
30.0	0.05	0.982 291 057 9	0.008 894 022 888	0.008 814 919 215
		0.982 24	0.008 92	0.008 84
		0.982 290 9	0.008 894 12	0.008 815 02
		0.982 291 056	0.008 894 024 1	0.008 814 920 4
		$\lesssim 10^{-17}$	0.008 894 022 888	0.991 105 977 1
		0	0.008 92	0.991 08
		0	0.008 894 12	0.991 105 88
		0	0.008 894 024 1	0.991 105 975 9

Our second example

$$Q^2(z) = K^2(z^2+1)^2 z/(z-p), \quad K^2>0,\ p>0, \tag{27}$$

describes propagation through one cutoff (at $z=0$) and one resonance (at $z=p$). It differs from the Budden model, Eq. (13a), only through the factor $(z^2+1)^2$ which was chosen so as to satisfy Eq. (17) ($m_\infty = 4$), and lead to elementary integrals. Thus Γ_{2N+1} associated with $z=0$ and $z=p$ is an algebraic function of K and p for arbitrary N (see Appendix C), e.g.,

$$\Gamma_1/\pi Kp = 1 + \tfrac{5}{8}p^2, \tag{28a}$$

$$\frac{\Gamma_3-\Gamma_1}{\pi Kp} = \frac{S(18+13p^2)-(4+9p^2)}{32K^2D(1+p^2)} = \frac{1}{2K^2}\left(1-\frac{13}{16}p^2+\cdots\right), \tag{28b}$$

$$\frac{\Gamma_5-\Gamma_3}{\pi Kp} = \frac{1}{262144K^4D(1+p^2)^4} \times [2S(248560+928704p^2 + 1362813p^4+897564p^6+222595p^8) - 11200+255504p^2+1325064p^4 + 2159781p^6+1502466p^8+385645p^{10}] = \frac{15}{128K^4}\left(1-\frac{1}{3}p^2+\cdots\right), \tag{28c}$$

where

$$S = 1+(1+p^2)^{1/2}, \quad D = [2S(1+p^2)]^{1/2}. \tag{29}$$

Inserting these Γ-integrals into Eq. (15), and the result into Eqs. (13b), (13c), we obtain approximate formulas for R_0 etc., in successive orders. Typical results are tabulated in Table III versus results obtained numerically. In the first five cases we increase K while keeping $Kp = \text{const}$, which leads to a systematic increase in the accuracy of the DPIA. The last two cases demonstrate only weak dependence of the accuracy on p for $K = \text{const}$.

When integrating the wave equation numerically we were using third rather than first-order phase-integral approximation to decompose $\psi(x)$ into incoming and outgoing waves at large distances (see Sec V). This reduced the computing time by several orders of magnitude for $Q^2_\infty \gg 1$, or $K^2 \gg 1$. Finally in the first example, Eq. (18), the computing times ranged from 20 seconds for $Q^2_\infty = 1$ to 95 seconds for $Q^2_\infty = 10^4$, versus 0.07 second for a typical DPIA calculation (both models included), on the CDC CYBER 73 computer. The second example, Eq. (27), required from 136 seconds for $K=1$ to 690 seconds for $K=100$, versus 0.02 second for the DPIA calculation. The computational details will be described elsewhere, and the programs are obtainable from the author on request.

ACKNOWLEDGMENTS

Part of this work, including symbolic computations, was performed at the Institute of Theoretical Physics, University of Uppsala, where the author spent September and October in 1982. He would like to thank Professor M. Hison Holmdahl, the Uppsala University Rector, for financial support during this visit. Interesting discussions with Professors N. Fröman, P. O. Fröman, and A. Kuszell, are also acknowledged.

APPENDIX A: Γ-INTEGRAL FOR EPSTEIN'S MODEL

Using Eqs. (7a), (7b) ($L=1$) in Eq. (2) we obtain

$$\Gamma_{2N+1} = \frac{1}{i}\oint_{C_u} \frac{[P_2(u)]^{1/2}}{1-u^2}\left[1+\sum_{n=1}^{N}\tilde{Y}_{2n}(u)\right]du, \tag{A1}$$

where C_u encircles $u_{1,2}$ in a positive sense. Functions $\tilde{Y}_{2n}(u)$ have the following general form ($c_{njl} = \text{const}$):

$$\tilde{Y}_{2n}(u) = \frac{1-u^2}{P_2(u)} \times\left[\frac{\delta_n}{a^{n-1}} + \sum_{l=1}^{3n-1}\sum_{j=1,2}\frac{c_{njl}}{(u-u_j)^l}\right], \tag{A2}$$

$$\delta_n = (2n)!/(n!)^2(2n-1)\,4^{2n}. \tag{A3}$$

This can be shown by first checking that Eq. (A2) holds for $n=1$, and then using the recurrence relations for Y_{2n} [see, e.g., Eq. (22) in Ref. 8], along with Eq. (31) in Ref. 8. In view of Eqs. (A1) and (A2) Γ_{2N+1} involves the following integrals ($l = 0,1,2,\ldots$):

$$\oint_{C_u}(u-u_j)^{-l}[P_2(u)]^{-1/2}\,du = \oint_{C_u} u^{-(l+1)}\left(a^{-1/2}+\sum_{k=1}^{\infty}a_{jlk}\,u^{-k}\right)du = 2\pi i a^{-1/2}\delta_{0l}, \tag{A4}$$

$$\oint_{C_u}(u\mp 1)^{-1}[P_2(u)]^{-1/2}\,du = \mp\frac{2\pi i}{Q_\pm}, \tag{A5}$$

which easily leads to Eq. (8); derivation of Eq. (A5) is similar to that for Eq. (C4).

APPENDIX B: Γ-INTEGRAL FOR EQUATION (18)

Assume first that $Q_0^2 > 0$, i.e., the zeros $\pm z_1$ are pure imaginary. Equation (19a) follows easily by integrating along

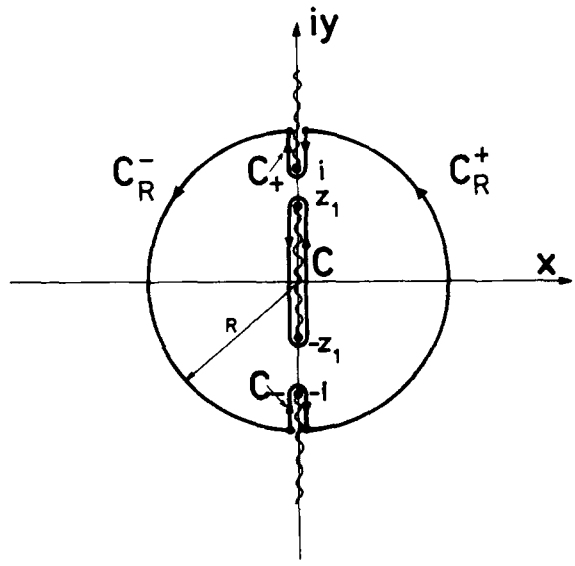

FIG. 1. Integration contours for $Q^2(z)$ given by Eq. (18).

the two lips of the cut connecting $\pm z_1$, see Fig. 1, and using Eqs. 220.04 and 318.02 of Ref. 9.

Using Eqs. (23a), (23b) of Ref. 1 we can write

$$Y_{2n}(z) = \tfrac{1}{4} Q_\infty^{-2n} P_{3n-1}(z^2)(z^2 - z_1^2)^{-3n}(z^2+1)^{-n}, \tag{B1}$$

where the polynomial $P_{3n-1}(z^2)$ is independent of Q_∞^2. This allows for the following decomposition into irreducible fractions:

$$Y_{2n}(z)\, 4Q_\infty^{2n}(z^2 - z_1^2) = \sum_{j=1}^{n} \frac{r_{nj}}{(z^2+1)^j} + \sum_{l=1}^{3n-1} \frac{s_{nl}}{(z^2 - z_1^2)^l}. \tag{B2}$$

Using Eqs. (18) and (B2), the Γ-integral (2) can be written as

$$\Gamma_{2N+1} = \Gamma_1 + \sum_{n=1}^{N} \frac{1}{Q_\infty^{2n-1}} \times \left(\sum_{j=1}^{n} r_{nj} R_j + \sum_{l=1}^{3n-1} s_{nl} S_l \right), \tag{B3}$$

$$R_j = \frac{1}{4i} \oint_C (z^2+1)^{-j} \times [(z^2+1)(z^2 - z_1^2)]^{-1/2}\, dz, \tag{B4}$$

$$S_l = \frac{(-1)^l}{4i} \oint_C (z_1^2 - z^2)^{-l} \times [(z^2+1)(z^2 - z_1^2)]^{-1/2}\, dz. \tag{B5}$$

To determine R_j ("regular" contribution to Γ) we integrate along the cut connecting $\pm z_1$, which yields [see Eq. (20), and Eqs. 219.07, 315.00, 315.02, and 315.05 in Ref. 9]

$$R_0 = K(k), \quad R_1 = E(k)/k'^2,$$
$$R_{j+1} = [(2j+1)k'^2]^{-1} \times [2j(2-k^2) R_j - (2j-1) R_{j-1}], \quad j = 1,2,\ldots. \tag{B6}$$

To determine S_l (the "singular" contribution) we deform the contour C as shown in Fig. 1, and take the limit $R \to \infty$. Contributions coming from $C_R^{\pm}$ vanish as $R \to \infty$, and those corresponding to integration along $C_\pm$ lead to (see Eqs. 215.06, 318.00, 318.02, and 318.05 in Ref. 9)

$$S_0 = K(k), \quad S_1 = (k^2 k'^2)^{-1}[k'^2 K(k) - E(k)],$$
$$S_{l+1} = [(2l+1)k^2 k'^2]^{-1} \times [2l(1-2k^2) S_l + (2l-1) S_{l-1}], \quad l = 1,2,\ldots. \tag{B7}$$

For $Q_0^2 < 0$ (real zeros $\pm z_1$) calculations are along similar lines. They lead to the results which can also be obtained from those for $Q_0^2 > 0$ by using Eqs. (21).

Finding r_{nj} and s_{nl} from Eq. (B2), R_j and S_l from the recurrence relations (B6) and (B7), and finally Γ_{2N+1} from Eq. (B3) is an algorithmic procedure which can be programmed on a computer. Our Eqs. (19b), (19c) were obtained by using the REDUCE system for algebraic manipulations implemented at the QZ computer in Stockholm.

APPENDIX C: Γ-INTEGRAL FOR EQUATION (27)

Expanding $Q(z)$ in a Laurent series convergent for $|z| > p$ we easily calculate Γ_1 given by Eq. (28a). Using Eqs. (23a), (23b) of Ref. 1 we find

$$Y_{2n}(z) = K^{-2n} P_{6n}(z)\, z^{-3n}(z^2+1)^{-4n}(z-p)^{-n}, \tag{C1}$$

and so we can write

$$Y_{2n}(z)\, K^{2n} z(z^2+1) = \sum_{j=1}^{3n-1} \frac{a_{nj}}{z^j} + \sum_{l=1}^{n} \frac{b_{nl}}{(z-p)^l} + \sum_{m=1}^{4n-1} \left[\frac{c_{nm}}{(z-i)^m} + \frac{c_{nm}^*}{(z+i)^m} \right]. \tag{C2}$$

Equations (C2) and (27) indicate that $\Gamma_{2N+1} - \Gamma_1$ involves three types of integrals ($a = 0, p$, and $\pm i$):

$$I_m(a) = \frac{1}{i} \oint_C (z-a)^{-m} [z(z-p)]^{-1/2}\, dz, \quad m = 1,2,\ldots. \tag{C3}$$

The first two integrals are zero (Laurent expansion), and $I_m(i)$ can be calculated by deforming the contour C so as to include apart from $z = 0$ and $z = p$ also the mth-order pole at $z = i$. The integral along the deformed contour is zero, i.e.,

$$I_m(i) + \frac{2\pi}{(m-1)!} \left. \frac{d^{m-1}[z(z-p)]^{-1/2}}{dz^{m-1}} \right|_{z=i} = 0, \tag{C4}$$

which easily leads to $[I_m(-i) = I_m(i)^*]$

$$\Gamma_{2N+1} = \Gamma_1 + 4\pi \times \sum_{n=1}^{N} \frac{1}{K^{2n-1}} \sum_{m=1}^{4n-1} \frac{1}{(m-1)!} \times \mathrm{Re}\left\{ i c_{nm} \left. \frac{d^{m-1}[z(p-z)]^{-1/2}}{dz^{m-1}} \right|_{z=i} \right\}, \tag{C5}$$

where the main branch of the square root must be taken.

Equations (C2) and (C5) define $\Gamma_{2N+1} - \Gamma_1$ in an algorithmic way for arbitrary N, and can be the basis for symbolic computations. Our Eqs. (28b), (28c) were obtained by using the REDUCE system mentioned in Appendix B.

[1] A. A. Skorupski, J. Math. Phys. **25**, 2651 (1984).
[2] O. Rydbeck, Philos. Mag. **30**, 282 (1940); **34**, 342 (1943).
[3] P. S. Epstein, Proc. Nat. Acad. Sci. USA **16**, 627 (1930).
[4] K. G. Budden, "The Non-existence of Fourth-reflection Condition for Radio Waves in the Ionosphere," in *Physics of the Ionosphere: Report of the Physical Society Conference at Cavendish Laboratory* (Physical Society, London, 1955), p. 320.
[5] N. Fröman and P. O. Fröman, *JWKB-Approximation, Contributions to the Theory* (North-Holland, Amsterdam, 1965).
[6] N. Fröman and P. O. Fröman, Nucl. Phys. A **147**, 606 (1970).
[7] F. Karlsson, Phys. Rev. D **11**, 2120 (1975).
[8] A. A. Skorupski, Rep. Math. Phys. **17**, 161 (1980).
[9] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists*, second ed. (Springer, New York, 1971).

Coherent state theory of the noncompact symplectic group

D. J. Rowe
Department of Physics, University of Toronto, Toronto, Ontario, Canada M5S 1A7

(Received 7 February 1984; accepted for publication 6 April 1984)

An extended coherent state theory is presented for the noncompact sp(3,R) group which reveals a simple relationship between the sp(3,R) algebra and its contracted u(3)-boson limit. The relationship is used to derive a remarkably accurate analytic expression for sp(3,R) matrix elements for the generic lowest-weight representations. The expression is shown to be exact whenever the states involved are multiplicity free with respect to the u(3) subalgebra. It is further shown how exact matrix elements are easily calculated in general. Dyson and Holstein–Primakoff type u(3)-boson expansions are given.

PACS numbers: 03.65.Fd, 02.20. + b, 21.60.Fw

I. INTRODUCTION

Coherent states are important for several reasons. Often a set of coherent states is isomorphic to some classical or semiclassical phase space; e.g., the classical phase space of Bargmann coherent states[1,2] or the semiclassical phase space of time-dependent Hartree–Fock theory.[3] Thus coherent states facilitate classical or semiclassical descriptions of quantal systems.

Coherent states also serve a useful function as generator states for the basis vectors of a representation space. In particular, techniques for using generator states to facilitate microscopic sp(3,R) model[4] calculations of collective states have been developed by Filippov *et al.*[5] The analysis of this paper is very relevant to that program.

Another important application of coherent states is to boson expansion theory (see Refs. 6 and 7 for reviews and lists of references). Boson expansions are useful because of the simplicity of the Weyl algebras and the fact that, in many important physical applications, it is sufficient to retain only the leading terms. In this way, one can formalize, and provide the corrections to, the familiar approximation of treating composite fermion systems, like a Cooper pair or a superposition of particle–hole pairs, as bosons.

Coherent state representations are also directly useful in their own right. For example, many problems are more simply executed in Bargmann (coherent state) space than in the more traditional Schrödinger Hilbert space (cf. Ref. 8, for example).

In this paper, we construct a coherent state realization of the real symplectic group Sp(3,R) [sometimes called Sp(6,R)] and use it to obtain expressions for the matrix elements of the sp(3,R) Lie algebra for all the lowest-weight representations, i.e., the discrete series with lower bound.

Our coherent theory differs from that of Perelomov[9] and Onofri[10] as follows: If $|0\rangle$ is a lowest weight state for a group G and $H \subset G$ is the little group, a Perelomov–Onofri coherent state Hilbert space is a space of holomorphic functions on G/H. Thus, in essence, their coherent state representations are induced from one-dimensional representations of a subgroup. In our approach, representations of Sp(n,R) are induced from arbitrary representations of the maximal compact subgroup U(n). Thus our coherent state Hilbert space is isomorphic to a space of holomorphic vector-valued functions on Sp(n,R)/U(n) which take values in the given U(n) representation space. Our approach is therefore closely related to the theory of induced representations.[11] It also has much in common with the Sp(n,R) representation theory of Ref. 12.

Whereas standard coherent state realizations naturally generate boson expansions,[7] our realization generates U(n)-boson expansions which are more useful for calculating matrix elements in an Sp(n,R)$\supset$U(n) basis. In this paper, we restrict consideration, for simplicity, to $n = 3$. However, the extension to $n > 3$ is straightforward.

The action of the sp(3,R) algebra on N-particle Hilbert space is given by a standard realization[4] of an sp(3,R) basis

$$\begin{aligned} A_{ij} &= \sum_{n=1}^{N} b^{\dagger}_{ni}\, b^{\dagger}_{nj}, \quad i,j = 1,2,3, \\ B_{ij} &= \sum_{n} b_{ni} b_{nj}, \\ C_{ij} &= \frac{1}{2}\sum_{n} (b^{\dagger}_{ni} b_{nj} + b_{nj} b^{\dagger}_{ni}) \\ &= \sum_{n} b^{\dagger}_{ni} b_{nj} + \delta_{ij}\frac{N}{2}, \end{aligned} \tag{1.1}$$

where

$$\begin{aligned} b_{in} &= \frac{1}{\sqrt{2}}\left(x_{in} + \frac{\partial}{\partial x_{in}}\right), \\ b^{\dagger}_{in} &= \frac{1}{\sqrt{2}}\left(x_{in} - \frac{\partial}{\partial x_{in}}\right) \end{aligned} \tag{1.2}$$

are Weyl boson operators. Thus it is apparent that one already has, in Eq. (1.1), a boson expansion for the sp(3,R) algebra. However, it is an uneconomical expansion involving $3N$ distinct bosons. One can of course put $N = 1$. However, the resulting realization then admits only the relatively trivial one-particle $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$ and $(\frac{3}{2},\frac{1}{2},\frac{1}{2})$ representations. Furthermore, we are particularly interested in representations for large N. What is wanted, therefore, is a realization that is both economical in terms of the number of bosons it invokes and that, at the same time, admits the generic $(\sigma_1,\sigma_2,\sigma_3)$ representations.

In the large N limit ($N \gtrsim 50$), it has been shown that the sp(3,R) algebra contracts to a simple u(3)-boson algebra[13] in which the A_{ij} and B_{ij} raising and lowering operators behave like single boson creation and annihilation operators, respec-

 0022-2488/84/092662-10$02.50

tively. It was pointed out by Smirnov[14] that these boson limits must surely be the first terms of a Dyson[15] or a Holstein–Primakoff[16] expansion, corresponding to embeddings of sp(3,R) in the u(3)-boson enveloping algebra. We show that such expansions, in which pairs of bosons behave to leading order like single bosons, are obtained directly from the coherent state realization. We recall that coherent states for a symplectic algebra were derived originally by Barut and Girardello[17] for the sp(1,R) algebra. They were constructed for the sp(2,R) algebra by Mlodinow and Papanicolau[18] and used to obtain an $N\rightarrow\infty$ contraction of sp(2,R) with applications to the s-wave states of the helium molecule in mind. Deenen and Quesne[19] similarly analyzed the sp(3,R) algebra and obtained a Barut type coherent state realization applicable to $O(N)$ invariant representations. They furthermore obtained Dyson and Holstein–Primakoff boson expansions. Although too restrictive for our purpose, these papers provide a valuable background for the more ambitious objective of finding realizations that admit all the discrete series representations of sp(3,R) with lower bound [i.e., all the lowest-(or highest-)weight representations], which are precisely the representations that are relevant in applications of the sp(3,R) model[4] to nuclear collective states.

Our analysis was motivated, in part, by a recent preprint of Castaños *et al.*[20] which gave analytic expressions for the matrix elements of sp(3,R) for the $(\sigma_1=\sigma_2=\sigma_3)$ representations. The similarity between their expressions and recently obtained matrix elements for the u(3)-boson algebra[21] suggested a closer relationship between the sp(3,R) and u(3)-boson algebras than was hitherto envisaged. This turned out to be the case and in a recent letter,[22] it was shown that the results of Ref. 20 are a special case of a general analytic expression which gives exact matrix elements whenever the states involved are multiplicity free in an sp(3,R)$\supset$u(3) classification. Furthermore, it was shown that, even when not exact, the analytic approximation to matrix elements is remarkably accurate. The coherent state origin of the analytic expression is presented here and it is further shown how to calculate exact matrix elements in general.

Note added in proof: Following submission of this paper, a preprint was received from Deenen and Quesne on a partially coherent state representation of the sp(N,R) algebras which has substantial overlap with this paper. They have also reproduced the analytic matrix elements of Castãnos *et al.*[20] using coherent state methods in a letter article [J. Deenen and C. Quesne, J. Phys. A: Math. Gen. **17**, L405 (1984)].

II. NOTATIONS AND MAJOR RESULTS

In this section we summarize the results and in subsequent sections derive and enlarge on them. The basic result is a coherent state realization of the Sp(3,R) algebra:

$$\begin{aligned}&\Gamma(A_{ij})=(\mathbb{C}z)_{ji}+(\mathbb{C}z)_{ij}-4z_{ij}+(z\nabla z)_{ij},\\&\Gamma(B_{ij})=\nabla_{ij},\\&\Gamma(C_{ij})=\mathbb{C}_{ij}+(z\nabla)_{ij},\end{aligned}\tag{2.1}$$

where (z_{ij}) is a 3×3 array of complex variables, ∇_{ij} is the differential operator

$$\nabla_{ij}=(1+\delta_{ij})\frac{\partial}{\partial z_{ij}},\tag{2.2}$$

and the $(\mathbb{C}_{ij})$ are a basis for an "intrinsic" u(3) algebra. Note that we use the matrix notation of Deenen and Quesne,[19] e.g., $(\mathbb{C}z)_{ij}=\Sigma_k\mathbb{C}_{ik}z_{kj}$, etc.

The elementary building blocks of this realization have a simple agebraic structure, viz.,

$$\begin{aligned}&[\nabla_{ij},z_{lk}]=\delta_{il}\delta_{jk}+\delta_{ik}\delta_{jl},\\&[\mathbb{C}_{ij},\mathbb{C}_{lk}]=\delta_{jl}\mathbb{C}_{ik}-\delta_{ik}\mathbb{C}_{lj},\\&[\mathbb{C}_{ij},z_{lk}]=[\mathbb{C}_{ij},\nabla_{lk}]=0.\end{aligned}\tag{2.3}$$

Indeed, they are also components of a coherent state realization of a u(3)-boson algebra, viz.,

$$\begin{aligned}&\gamma(a^\dagger_{ij})=z_{ij},\\&\gamma(a_{ij})=\nabla_{ij},\\&\gamma(C_{ij})=\mathbb{C}_{ij}+(z\nabla)_{ij},\end{aligned}\tag{2.4}$$

where $a^\dagger_{ij}=a^\dagger_{ji}$ and $a_{ij}=a_{ji}$ are the boson operators of a six-dimensional Weyl algebra

$$[a_{ij},a^\dagger_{lk}]=\delta_{il}\delta_{jk}+\delta_{ik}\delta_{jl},\tag{2.5}$$

and

$$[\mathbb{C}_{ij},a^\dagger_{lk}]=[\mathbb{C}_{ij},a_{lk}]=0.\tag{2.6}$$

An sp(3,R) lowest weight unirrep[12] is characterized by the u(3) quantum numbers $\sigma=(\sigma_1,\sigma_2,\sigma_3)$ of its lowest weight state $|\sigma\rangle$;i.e.,$|\sigma\rangle$ satisfies

$$\begin{aligned}&C_{ii}|\sigma\rangle=\sigma_i|\sigma\rangle,\\&C_{ij}|\sigma\rangle=0,\quad i<j,\\&B_{ij}|\sigma\rangle=0,\end{aligned}\tag{2.7}$$

for $i,j=1,2,3$. The carrier space $\mathbb{H}^\sigma_{\rm sp}$ for this representation is constructed by operating on the lowest weight state with all polynomials in the raising operators.

A u(3)-boson unirrep[13] is likewise characterized by the u(3) quantum numbers $\sigma=(\sigma_1,\sigma_2,\sigma_3)$ of its lowest weight state $|\sigma)$, which satisfies

$$\begin{aligned}&C_{ii}|\sigma)=\sigma_i|\sigma),\\&C_{ij}|\sigma)=0,\quad i<j,\\&a_{ij}|\sigma)=0,\end{aligned}\tag{2.8}$$

for all $i,j=1,2,3$. Note that we now use rounded kets, as opposed to angular kets, to distinguish u(3)-boson and sp(3,R) states.

A convenient U(3) coupled basis for the carrier space $\mathbb{H}^\sigma_{\rm ub}$ of this representation is obtained by first combining the boson raising operators $(a^\dagger_{ij})$ into U(3) tensors $X^{(n)}(a^\dagger)$ of rank $n=(n_1,n_2,n_3)$. These tensors are then U(3)-coupled to the lowest weight state $|\sigma)$ to give an orthonormal basis of states of total U(3) symmetry $\omega=(\omega_1,\omega_2,\omega_3)$, with multiplicity ρ,

$$|\sigma n\rho\omega\alpha)=N_{\sigma n\rho\omega}(X^{(n)}(a^\dagger)\times|\sigma))^{\rho\omega}_\alpha,\tag{2.9}$$

where $N_{\sigma n\rho\omega}$ is a normalization constant and α indexes a basis for the U(3) unirrep ω. Since the $(a^\dagger_{ij})$ are themselves components of a (2,0,0) U(3) tensor, the construction of $X^{(n)}(a^\dagger)$ is standard. Its lowest-weight component, for example, is given explicitly by

$$X^{(n)}_0(a^\dagger)=(a^\dagger_{11})^{(n_1-n_2)/2}\begin{vmatrix}a^\dagger_{11}&a^\dagger_{12}\\a^\dagger_{21}&a^\dagger_{22}\end{vmatrix}^{(n_2-n_3)/2}\times(\det a^\dagger)^{n_3/2}.\tag{2.10}$$

An orthonormal basis for $\mathbb{H}^{\sigma}_{sp}$ will later be defined implicitly by a mapping $\mathbb{H}^{\sigma}_{ub} \to \mathbb{H}^{\sigma}_{sp}$. From the two isomorphisms (2.1) and (2.4), one immediately infers the Dyson type realization of sp(3,R)

$$D(A_{ij}) = (\mathbb{C}a^{\dagger})_{ij} + (\mathbb{C}a^{\dagger})_{ji} - 4a^{\dagger}_{ij} + (a^{\dagger}aa^{\dagger})_{ij},$$
$$D(B_{ij}) = a_{ij},$$
$$D(C_{ij}) = \mathbb{C}_{ij} + (a^{\dagger}a)_{ij}, \tag{2.11}$$

which is manifestly nonunitary. The nonunitarity arises because, whereas the coherent state realization Γ is unitary with respect to the Sp(3,R) measure, it is not unitary with respect to the U(3)-boson measure. Thus, to obtain a unitary action of Sp(3,R) on the U(3)-boson space, one must take account of the change of measure.

It will be shown that there exists a Hermitian positive definite U(3)-scalar operator κ that transforms Γ into a unitary realization γ, with respect to the U(3)-boson measure, with

$$\gamma(A_{ij}) = \kappa^{-1}\Gamma(A_{ij})\kappa = \kappa z_{ij}\kappa^{-1},$$
$$\gamma(B_{ij}) = \kappa^{-1}\Gamma(B_{ij})\kappa = \kappa^{-1}\nabla_{ij}\kappa,$$
$$\gamma(C_{ij}) = \kappa^{-1}\Gamma(C_{ij})\kappa = \mathbb{C}_{ij} + (z\nabla)_{ij}. \tag{2.12}$$

The slight abuse of notation of using the same symbol γ to denote both a realization of Sp(3,R) and of the U(3)-boson algebra [Eq. (2.4)] will be convenient and unambiguous.

The determination of κ is greatly facilitated by the observation of the identity

$$\Gamma(A_{ij}) = [\Lambda, z_{ij}], \tag{2.13}$$

where Λ is the U(3) invariant

$$\Lambda = \tfrac{1}{2}[(\mathbb{C} + z\nabla)(\mathbb{C} + 2\nabla)] - \tfrac{1}{4}\mathrm{tr}(z\nabla z\nabla) - \mathrm{tr}(z\nabla). \tag{2.14}$$

Together, Eqs. (2.12) and (2.13) imply

$$[\Lambda, z_{ij}] = \kappa^2 z_{ij}\kappa^{-2}. \tag{2.15}$$

This expression is particularly useful because Λ is diagonal in the basis (2.9) with eigenvalues

$$\Omega(\sigma n\omega) = \frac{1}{4}\sum_i [2\omega_i^2 - n_i^2 + 8(\omega_i - n_i) - 2i(2\omega_i - n_i)]. \tag{2.16}$$

Equation (2.15) implies

$$\kappa^2\,\mathrm{tr}(zz) = \mathrm{tr}([\Lambda, z]\kappa^2 z), \tag{2.17}$$

which generates recursion relations for the matrix elements of κ which are easily solved.

Since κ is a U(3) scalar, it has nonvanishing matrix elements only between basis states, (2.9), of the same $\omega\alpha$. Thus, states which are unambiguously identified by $\omega\alpha$ are necessarily eigenstates of κ. We shall refer to such states as "simple." One infers that matrix elements of $\gamma(A_{ij})$ between simple states are given by

$$\langle\omega'\alpha'|\gamma(A_{ij})|\omega\alpha\rangle = (\Omega(\omega') - \Omega(\omega))^{1/2}\langle\omega'\alpha'|z_{ij}|\omega\alpha\rangle, \tag{2.18}$$

where we now suppress all but the U(3) labels $\omega\alpha$ of the states to emphasize the restriction of this equation to simple states. From Eq. (2.18), one immediately infers a corresponding relationship between sp(3,R) matrix elements in $\mathbb{H}^{\sigma}_{sp}$ and U(3)-boson matrix elements in $\mathbb{H}^{\sigma}_{ub}$

$$\langle\omega'\|A\|\omega\rangle = (\Omega(\omega') - \Omega(\omega))^{1/2}\langle\omega'\|a^{\dagger}\|\omega\rangle. \tag{2.19}$$

Matrix elements for the sp(3,R) lowering operators are obtained from (2.19) by the general relationship

$$\langle\sigma n\rho\omega\|B\|\sigma n'\rho'\omega'\rangle = (-1)^{\lambda+\mu+\lambda'+\mu'}[\dim(\lambda'\mu')/\dim(\lambda\mu)]^{1/2} \times\langle\sigma n'\rho'\omega'\|A\|\sigma n\rho\omega\rangle^{*}, \tag{2.20}$$

where

$$\dim(\lambda\mu) = \tfrac{1}{2}(\lambda+1)(\mu+1)(\lambda+\mu+2) \tag{2.21}$$

is the dimension of the SU(3) representation $(\lambda\mu)$ and

$$\lambda = \omega_1 - \omega_2, \quad \mu = \omega_2 - \omega_3, \tag{2.22}$$

etc. The matrix elements of the u(3) operators (C_{ij}) are well-known.

We recall that the matrix elements of the u(3)-boson algebra are known and are given, in Ref. 21, by

$$\langle\sigma n'\rho'\omega'\|a^{\dagger}\|\sigma n\rho\omega\rangle = (-1)^{\lambda'+\mu'-\lambda-\mu} \times U((\lambda_\sigma\mu_\sigma)(\lambda_n\mu_n)(\lambda'\mu')(20);(\lambda\mu)\rho(\lambda'_n\mu'_n)\rho') \times\langle n'\|a^{\dagger}\|n\rangle, \tag{2.23}$$

where U is an SU(3) Racah coefficient,

$$\lambda_\sigma = \sigma_1 - \sigma_2, \quad \mu_\sigma = \sigma_2 - \sigma_3,$$
$$\lambda_n = n_1 - n_2, \quad \mu_n = n_2 - n_3, \tag{2.24}$$

etc., and

$$\langle n'\|a^{\dagger}\|n\rangle = \left[\frac{(n_1+4)(n_1-n_2+2)(n_1-n_3+3)}{(n_1-n_2+3)(n_1-n_3+4)}\right]^{1/2} \times\delta_{n_1',n_1+2}\delta_{n_2',n_2}\delta_{n_3',n_3} + \left[\frac{(n_2+3)(n_1-n_2)(n_2-n_3+2)}{(n_1-n_2-1)(n_2-n_3+3)}\right]^{1/2} \times\delta_{n_1',n_1}\delta_{n_2',n_2+2}\delta_{n_3',n_3} + \left[\frac{(n_3+2)(n_2-n_3)(n_1-n_3+1)}{(n_1-n_3)(n_2-n_3-1)}\right]^{1/2} \times\delta_{n_1',n_1}\delta_{n_2',n_2}\delta_{n_3',n_3+2}. \tag{2.25}$$

Thus Eq. (2.19) gives analytic expressions for all the matrix elements of the sp(3,R) algebra between simple states for any sp(3,R) unirrep $\{\sigma_1,\sigma_2,\sigma_3\}$. For representations with $\sigma_1 = \sigma_2 = \sigma_3$, all basis states are simple and the analytic expressions obtained are then identical to those given by Castaños *et al.*[20] for this special case.

For nonsimple states, the matrix elements are obtained from Eq. (2.12) in the form

$$\langle i\omega'\|A\|l\omega\rangle = \sum_{j,k}\kappa_{ij}(\omega')\langle j\omega'\|a^{\dagger}\|k\omega\rangle K^{-1}_{kl}(\omega), \tag{2.26}$$

and can be evaluated following the solution of some simple recursion equations to obtain the elements of $\kappa(\omega)$ for each ω.

A neglect of the off-diagonal elements of κ, with respect to the basis (2.9), gives the approximate analytic expression for sp(3,R) matrix elements

$$\langle \sigma n'\rho'\omega' \| A \| \sigma n\rho\omega\rangle \cong (\Omega(\sigma n'\omega') - \Omega(\sigma n\omega))^{1/2} \langle \sigma n'\rho'\omega' \| a^\dagger \| \sigma n\rho\omega\rangle. \tag{2.27}$$

It has been shown[22] that this expression gives remarkably accurate results particularly for large σ. The reason is explained in Sec. VII, where it is shown that the approximate matrix elements (2.27) correspond to the approximation

$$\begin{aligned} \gamma(A_{ij}) &\cong \Lambda^{1/2}\cdot z_{ij}, \\ \gamma(B_{ij}) &\cong \nabla_{ij}\cdot\Lambda^{1/2}, \\ \gamma(C_{ij}) &= \mathbb{C}_{ij} + (z\nabla)_{ij}, \end{aligned} \tag{2.28}$$

where $\Lambda^n\cdot z$ and $\nabla\cdot\Lambda^n$ are defined for integer n by

$$\begin{aligned} \Lambda^n\cdot z &= [\Lambda, \Lambda^{n-1}\cdot z], \quad \Lambda\cdot z = [\Lambda, z], \\ \nabla\cdot\Lambda^n &= [\nabla\cdot\Lambda^{n-1}, \Lambda], \quad \nabla\cdot\Lambda = [\nabla, \Lambda], \end{aligned} \tag{2.29}$$

and $\Lambda^{1/2}\cdot z$ and $\nabla\cdot\Lambda^{1/2}$ are extensions of these definitions to $n = \frac{1}{2}$. It is shown that the approximation (2.28) is accurate up to terms of order $[(\lambda_\sigma + \mu_\sigma)/2\sigma]^4$. In particular, it is exact for $\lambda_\sigma = \mu_\sigma = 0$.

Finally, note that the replacement $z_{ij} \to a^\dagger_{ij}, \nabla_{ij} \to a_{ij}$ in Eq. (2.12) induces a unitary Holstein–Primakoff type realization of sp(3,R) in terms of the u(3)-boson algebra. It is given explicitly in Sec. VII for representations containing only simple states, e.g., $\sigma_1 = \sigma_2 = \sigma_3$. In general, the leading terms of the Holstein–Primakoff expansion are given by

$$\begin{aligned} P(A_{ij}) &= \sqrt{2\sigma}a^\dagger_{ij} + \tfrac{1}{2}\sqrt{2\sigma}\{(\mathbb{C}a^\dagger)_{ij} + (\mathbb{C}a^\dagger)_{ji} \\ &\quad - (2\sigma+4)a^\dagger_{ij} + (a^\dagger a a^\dagger)_{ij}\} + \cdots, \\ P(B_{ij}) &= \sqrt{2\sigma}a_{ij} + \tfrac{1}{2}\sqrt{2\sigma}\{(a\mathbb{C})_{ij} + (a\mathbb{C})_{ji} \\ &\quad - (2\sigma+4)a_{ij} + (aa^\dagger a)_{ij}\} + \cdots, \\ P(C_{ij}) &= \mathbb{C}_{ij} + (a^\dagger a)_{ij}, \end{aligned} \tag{2.30}$$

where $\sigma = (\sigma_1 + \sigma_2 + \sigma_3)/3$. Dropping the second- and higher-order terms of $P(A_{ij})$ and $P(B_{ij})$ gives precisely the result obtained previously by a contraction of sp(3,R) to the semidirect sum u(3)-boson algebra. Thus the Holstein–Primakoff expansion gives the higher-order corrections to that contraction limit.

III. COHERENT STATE REALIZATIONS OF U(3) AND GL(3,$\mathbb{C}$)

Consider a unirrep U of U(3) on N-particle state space, for example. This representation naturally extends to a nonunitary representation T of GL(3,C) by the standard process of complexification.

Let $|\sigma\rangle$ be any fixed state in the representation space, e.g., the lowest weight state. Then any state $|\Psi\rangle$ in the space can be represented as a coherent state wave function on GL(3,C) by

$$\psi(g) = \langle g|\Psi\rangle \equiv \langle \sigma|T(g)|\psi\rangle, \tag{3.1}$$

where

$$|g\rangle = T(g^\dagger)|\sigma\rangle, \quad g \in \mathrm{GL}(3,C). \tag{3.2}$$

A representation Γ of GL(3,C) is now defined by

$$[\Gamma(\alpha)\psi] = \psi(g\alpha). \tag{3.3}$$

Evidently Γ is isomorphic to T. However, Γ replaces what is generally a complicated realization in turns of functions of many-particle coordinates by much simpler and known functions. More important, it is a vital first step in proceeding to sp(3,R) coherent states.

To obtain the explicit realization $\Gamma(X)$ of an element X in the gl(3,C) algebra, first expand

$$X = \sum_{ij} X_{ij} E^{ij}, \tag{3.4}$$

where E^{ij} is the 3×3 matrix with lk element

$$E^{ij}_{lk} = \delta^i_l \delta^j_k. \tag{3.5}$$

Then, putting $\alpha = \exp(X)$ and replacing $g\alpha$ in Eq. (3.3) by

$$g(X) = g\exp(X), \tag{3.6}$$

we obtain

$$\begin{aligned} \Gamma(E_{ij})\Psi(g) &= \frac{\partial}{\partial X_{ij}}\Psi(g(X))\Big|_{X=0} \\ &= \sum_{lk} \frac{\partial g_{lk}(X)}{\partial X_{ij}}\Big|_{X=0} \frac{\partial \Psi(g)}{\partial g_{lk}}, \end{aligned} \tag{3.7}$$

giving

$$\Gamma(E_{ij}) = \sum_l g_{li}\frac{\partial}{\partial g_{lj}}. \tag{3.8}$$

To simplify subsequent expressions, we define $\mathbb{C}_{ij}$ by

$$\mathbb{C}_{ij}\langle g|\Psi\rangle = \langle g|C_{ij}|\Psi\rangle. \tag{3.9}$$

With this definition

$$\mathbb{C}_{ij} = \sum_l g_{li}\frac{\partial}{\partial g_{lj}} + \delta_{ij}\frac{N}{2}. \tag{3.10}$$

If $|\sigma\rangle$ is now chosen to be the lowest weight state for a U(3) representation $\sigma = (\sigma_1,\sigma_2,\sigma_3)$, then its coherent state wave function

$$X^{(\sigma)}(g) = \langle g|\sigma\rangle \tag{3.11}$$

must, by definition, satisfy

$$\begin{aligned} &\mathbb{C}_{ij}X^{(\sigma)}(g) = 0, \quad i<j, \\ &\mathbb{C}_{ii}X^{(\sigma)}(g) = \sigma_i X^{(\sigma)}(g). \end{aligned} \tag{3.12}$$

These equations have the well-known solution

$$X^{(\sigma)}(g) = (g_{11})^{\sigma_1-\sigma_2}\begin{vmatrix} g_{11} & g_{12} \\ g_{21} & g_{22}\end{vmatrix}^{\sigma_2-\sigma_3}(\det g)^{\sigma_3}. \tag{3.13}$$

Note that the normalization is correct because

$$X^\sigma(g = I) = \langle\sigma|\sigma\rangle = 1. \tag{3.14}$$

The overlap integral for different coherent states is also obtained immediately. From the definition, it follows that

$$\langle g'|g\rangle = \langle\sigma|T(g')T(g^\dagger)|\sigma\rangle = \langle\sigma|T(g'g^\dagger)|\sigma\rangle, \tag{3.15}$$

giving

$$\langle g'|g\rangle = X^{(\sigma)}(g'g^\dagger). \tag{3.16}$$

The inner product (ψ_1,ψ_2) can be defined in several ways to satisfy the fundamental requirement

$$(\psi_1,\psi_2) = \langle\Psi_1|\Psi_2\rangle. \tag{3.17}$$

One way is to identify the wave functions with Bargmann wave functions[2] and hence employ the Bargmann measure to give

$$(\psi_1,\psi_2)=\frac{1}{\pi^9}\int \psi_1^*(g)\psi_2(g)\exp\Big(-\sum_{ij}|g_{ij}|^2\Big)$$

$$\times\prod_{i,j} dg_{ij}\, dg_{ij}^*. \tag{3.18}$$

Another possibility is to follow Perelomov[9] and note that the operator

$$K=\int_{U(3)} U(g^\dagger)|\sigma\rangle\langle\sigma|U(g)dv(g), \tag{3.19}$$

with the integral restricted to $g\in U(3)$ and with $dv(g)$ the U(3) invariant measure, satisfies the equation

$$U(\alpha)K=K\,U(\alpha),\quad \alpha\in U(3). \tag{3.20}$$

Hence, by Schur's lemma, it is a multiple of the identity. Inserting K between states, $\langle\Psi_1|K|\Psi_2\rangle$, gives the inner product

$$(\psi_1,\psi_2)=k\int_{U(3)}\psi_1^*(g)\psi_2(g)dv(g) \tag{3.21}$$

with

$$k^{-1}=\int_{U(3)}|X^{(\sigma)}(g)|^2\,dv(g). \tag{3.22}$$

Both of the above inner products, by construction, must give the same overlaps for all states in the representation space. Furthermore, since they are defined independently of a particular representation (apart from the constant k) they must be valid and equivalent (up to the explicit value of k) for all $(\sigma_1,\sigma_2,\sigma_3)$. Other inner products presumably exist that are only applicable to a particular representation or to some subset of representations. However, in what follows, the explicit form of the inner product is not needed.

The above construction evidently extends to U(n) and GL(n,C) for any integer n.

IV. COHERENT STATE REALIZATION OF sp(3,R)

The sp(3,R) lowest weight state $|\sigma\rangle$ defined by Eq. (2.7), is evidently also a U(3) lowest weight state. We can therefore extend the U(3) [also GL(3,C)] coherent states of Eq. (3.2) to sp(3,R) by

$$|z,g\rangle=\exp\Big(\frac{1}{2}\sum_{ij}z_{ij}^*A_{ij}\Big)|g\rangle, \tag{4.1}$$

where $z_{ij}=z_{ji}$ is a 3×3 symmetric array of complex numbers. Thus we obtain the coherent state wave function

$$\psi(z,g)=\langle z,g|\Psi\rangle \tag{4.2}$$

corresponding to any state $|\Psi\rangle$ in the sp(3,R) representation space.

The coherent state realization of a step-down operator B_{ij} is inferred immediately from the identity

$$\langle z,g|B_{ij}|\Psi\rangle=\nabla_{ij}\langle z,g|\Psi\rangle. \tag{4.3}$$

Writing

$$\langle z,g|C_{ij}|\Psi\rangle=\langle g|e^X C_{ij}e^{-X}e^X|\Psi\rangle, \tag{4.4}$$

with

$$X=\frac{1}{2}\sum_{ij}z_{ij}B_{ij}, \tag{4.5}$$

and using the expansion

$$e^X C_{ij}e^{-X}=C_{ij}+[X,C_{ij}]$$
$$=C_{ij}+(zB)_{ij}, \tag{4.6}$$

we obtain, with Eqs. (3.10) and (4.3),

$$\langle z,g|C_{ij}|\Psi\rangle=(\mathbb{C}_{ij}+(z\nabla)_{ij})\langle z,g|\Psi\rangle. \tag{4.7}$$

Similarly, one finds that

$$\langle z,g|A_{ij}|\Psi\rangle$$
$$=\Big((\mathbb{C}z)_{ij}+(\mathbb{C}z)_{ji}+\sum_{lk}z_{il}z_{jk}\nabla_{lk}\Big)\langle z,g|\Psi\rangle. \tag{4.8}$$

Thus we obtain the coherent state realization Γ given in Eq. (2.1).

The volume element can be found for the special case $(\sigma_1=\sigma_2=\sigma_3=\sigma)$, using the techniques of Dobaczewski.[7] The lowest weight state, in this case, being SU(3) invariant, we can suppress the dependence of the wave functions on $g\in GL(3,C)$ and simplify the coherent state wave functions to

$$\psi(z)=\Big\langle\sigma|\exp\Big(\frac{1}{2}\sum_{ij}z_{ij}B_{ij}\Big)|\Psi\Big\rangle. \tag{4.9}$$

We find the inner product

$$\langle\Psi_1|\Psi_2\rangle=k\int_{-1}^{1}\psi_1^*(z)\psi_2(z)\,\det(I-zz^*)^{\sigma-4}\,d^2z, \tag{4.10}$$

where

$$k^{-1}=\int_{-1}^{1}\det(I-zz^*)^{\sigma-4}\,d^2z, \tag{4.11}$$

$$d^2z=\prod_{i<j}\frac{1}{1+\delta_{ij}}\,dz_{ij}\,dz_{ij}^*, \tag{4.12}$$

and the integrations are over the domain $-1\leqslant|z_{ij}|^2<1$.

We have not succeeded in finding the sp(3,R) measure for the generic $(\sigma_1,\sigma_2,\sigma_3)$ case. However, we have succeeded in finding inner products that avoid its use and which are more useful as a consequence.

An sp(3,R) state $|\Psi_1\rangle$ can be expanded

$$|\Psi_1\rangle=\sum_\alpha P^1_\alpha(A)|\sigma\alpha\rangle, \tag{4.13}$$

where (P^1_α) are polynomials and α indexes a basis of U(3) states for the $\sigma=(\sigma_1,\sigma_2,\sigma_3)$ representation. It follows that

$$\langle\Psi_1|\Psi_2\rangle=\sum_\alpha\int\langle\sigma\alpha|z,g\rangle\langle z,g|P^1_\alpha(B)|\Psi_2\rangle dv_s(z,g), \tag{4.14}$$

where dv_s is the unknown sp(3,R) measure. Now it is clear that

$$\langle\sigma\alpha|z,g\rangle=\langle\sigma\alpha|z=0,g\rangle\equiv\langle\sigma\alpha|g\rangle.$$

Therefore we have the inner product

$$\langle\Psi_1|\Psi_2\rangle=\sum_\alpha\int\langle\sigma\alpha|g\rangle\,[P^1_\alpha(\nabla)\langle z,g|\Psi_2\rangle]_{z=0}\,dv(g), \tag{4.15}$$

where $dv(g)$ is the U(3) measure (cf. Sec. III).

An equivalent inner product, which employs the U(3)-boson measure, is further obtained by introducing the U(3)-boson counterpart to $|\Psi_1\rangle$

$$|\phi_1\rangle=\sum_\alpha P^1_\alpha(a^\dagger)|\sigma\alpha\rangle. \tag{4.16}$$

Then, since

$$\langle \sigma\alpha|g\rangle \equiv (\sigma\alpha|g), \tag{4.17}$$

Eq. (4.15) can be augmented to the fully integral form

$$\langle \Psi_1|\Psi_2\rangle = \int\int (\phi_1|z,g)\langle z,g|\Psi_2\rangle d\mu(z)d\nu(g), \tag{4.18}$$

where

$$d\mu(z) = \exp\Big(-\frac{1}{2}\sum_{ij}|z_{ij}|^2\Big)\prod_{i<j}\frac{1}{1+\delta_{ij}}dz_{ij}\,dz_{ij}^* \tag{4.19}$$

is the Bargmann measure.

In what follows, however, we shall not need the explicit form of the inner product.

Note that, from the definition (4.2), sp(3,R) coherent state wave functions are holomorphic functions of z and g. Furthermore, since the U(3) coherent states are isomorphic to vectors in the original U(3) representation space, it follows that the sp(3,R) wave functions are isomorphic to holomorphic vector-valued functions of just the six z variables with vector values in the U(3) representation space.

V. CHANGE OF MEASURE

It was noted in Sec. II that whereas the realization Γ of sp(3,R), given by Eq. (2.1), is unitary with respect to the sp(3,R) coherent state measure, it is not unitary with respect to the U(3)-boson measure. Consequently, the induced Dyson realization (2.11) is likewise nonunitary. To obtain a unitary Holstein–Primakoff realization, we need to make a transformation of Γ to take account of the different sp(3,R) and U(3)-boson measures.

Let $V^\sigma_{\rm sp}$ denote the Hilbert space of coherent state wave functions with inner product defined by the sp(3,R) measure and let $V^\sigma_{\rm ub}$ denote the corresponding Hilbert space with respect to the U(3)-boson measure. Let

$$K: V^\sigma_{\rm ub} \to V^\sigma_{\rm sp} \tag{5.1}$$

transform an orthonormal basis for $V^\sigma_{\rm ub}$ into an orthonormal basis for $V^\sigma_{\rm sp}$, such that the U(3) symmetry of states is conserved, i.e.,

$$K^{-1}\Gamma(C)K = \Gamma(C), \quad C\in U(3). \tag{5.2}$$

Theorem: If κ is the positive Hermitian square root of $KK^\dagger$, i.e.,

$$\kappa^2 = KK^\dagger, \quad \kappa = \kappa^\dagger, \tag{5.3}$$

where Hermitian adjoints are defined with respect to the U(3)-boson measure, then γ, defined by

$$\gamma(X) = \kappa^{-1}\Gamma(X)\kappa, \quad X\in \text{sp}(3,R), \tag{5.4}$$

is a unitary realization of sp(3,R).

Proof: First observe that for X, $Y\in$ sp(3,R),

$$[\gamma(X),\gamma(Y)] = \kappa^{-1}[\Gamma(X),\Gamma(Y)]\kappa. \tag{5.5}$$

Since Γ is a realization of sp(3,R), it follows that

$$[\Gamma(X),\Gamma(Y)] = \Gamma([X,Y]), \tag{5.6}$$

and hence

$$[\gamma(X),\gamma(Y)] = \gamma([X,Y]). \tag{5.7}$$

Thus γ is a realization. Next observe that $\bar\gamma$, defined by

$$\bar\gamma(X) = K^{-1}\Gamma(X)K, \quad X\in \text{sp}(3,R), \tag{5.8}$$

is a unitary realization, by construction, implying that

$$\bar\gamma(A_{ij}) = \bar\gamma(B_{ij})^\dagger. \tag{5.9}$$

Now

$$\Gamma(B_{ij})^\dagger = z_{ij} \tag{5.10}$$

[with respect to the U(3)-boson measure]. Therefore Eq. (5.8) implies

$$\Gamma(A_{ij}) = KK^\dagger z_{ij}(KK^\dagger)^{-1} = \kappa^2 z_{ij}\kappa^{-2}. \tag{5.11}$$

Thus, from the definition (5.4) of γ and the fact that K and hence $KK^\dagger$ commute with $\Gamma(C)$, we obtain Eq. (2.12) for $\gamma(X)$, which is manifestly unitary.

VI. sp (3,R) MATRIX ELEMENTS

Matrix elements in $V^\sigma_{\rm ub}$ of $\gamma(x)$, $x\in$ U(3)-boson, are by the coherent state isomorphism (2.4) identical to the corresponding matrix elements of x in $\mathbb{H}^\sigma_{\rm ub}$, which are known. Thus, if we determine the matrix elements of κ, we can, by Eq. (2.12), deduce the matrix elements in $V^\sigma_{\rm ub}$ of $\gamma(X)$, $X\in$ sp(3,R). These in turn, again by isomorphism, can be identified with matrix elements in $\mathbb{H}^\sigma_{\rm sp}$ of X by which identification we implicitly define a basis for $\mathbb{H}^\sigma_{\rm sp}$. The outstanding problem then is to determine κ.

The basis (2.9) defines a decomposition of the coherent state space into a direct sum of U(3) invariant subspaces

$$V^\sigma_{\rm ub} = \sum_\omega V^{\sigma\omega}_{\rm ub} = \sum_{\omega,i} V^{\sigma i\omega}_{\rm ub}, \tag{6.1}$$

where $i=(n\rho)$ indexes the multiplicity of U(3) unirreps ω contained in $V^{\sigma\omega}_{\rm ub}$. Since the operator κ, by definition, is U(3) invariant, it follows that $V^{\sigma\omega}_{\rm ub}$ is κ invariant. Thus the decomposition (6.1) block-diagonalizes κ. The calculation of κ is thereby reduced to a set of calculations for $\{\kappa(\omega)\}$, where $\kappa(\omega)$ is the restriction of κ to $V^{\sigma\omega}_{\rm ub}$.

Now in the decomposition (6.1), many U(3) representations ω are multiplicity free. For example, the lowest weight state σ and the second weight states $(2,0,0)\times\sigma$ are always multiplicity free. We shall call such states and such representations "simple."

Theorem: If ω is simple then $\kappa(\omega)$ is a multiple of the identity I_ω on $V^{\sigma\omega}_{\rm ub}$, i.e.,

$$\kappa(\omega) = k(\omega)I_\omega. \tag{6.2}$$

Proof: The proof follows immediately from Schur's lemma and the fact that, if ω is simple, $V^{\sigma\omega}_{\rm ub}$ carries an irreducible U(3) representation.

Thus, when ω is simple, the evaluation of $\kappa(\omega)$ reduces to the evaluation of the single number $k(\omega)$. A simple extension of this argument shows that, if ω has multiplicity n, then $\kappa(\omega)$ is defined by an $n\times n$ Hermitian matrix.

The matrix elements of $\kappa(\omega)$ are easily derived starting from the identity $\Gamma(A_{ij}) = [\Lambda, z_{ij}]$, with Λ given by Eq. (2.14). This identity is easily proved by direct evaluation of the commutator using the Eqs. (2.3), from which one derives, for example,

$$[(z\nabla)_{ij}, z_{lk}] = \delta_{jl}z_{ik} + \delta_{jk}z_{il}. \tag{6.3}$$

From Eq. (5.11), one then obtains the fundamental equation

$$[\Lambda, z_{ij}] = \kappa^2 z_{ij}\kappa^{-2}. \tag{6.4}$$

TABLE I. $\Omega(\sigma n\omega)$ and $\kappa^2(\omega)/\kappa^2(\sigma)$ for some U(3) representations ω contained in the sp(3,R) unirrep (20,13,10).

Index	n	ω	$\Omega(\sigma n\omega)$	$\kappa^2(\omega)/\kappa^2(\sigma)$
1	(0,0,0)	(20,13,10)	$36\frac{1}{3}$	1
2	(2,0,0)	(22,13,10)	$76\frac{1}{3}$	40
3	(2,0,0)	(21,14,10)	$67\frac{1}{3}$	31
4	(2,0,0)	(20,15,10)	$60\frac{1}{3}$	24
8,1	(2,2,0)	(22,15,10)	$101\frac{1}{3}$	$\begin{pmatrix} 1014 & -2\sqrt{210} \\ -2\sqrt{210} & 940 \end{pmatrix}$
8,2	(4,0,0)	(22,15,10)	$98\frac{1}{3}$	

Thus, since Λ is diagonal in the U(3)-boson basis (2.9), Eq. (6.4) gives

$$(\sigma' n'\rho'\omega')\|\kappa^2(\omega')z\kappa^{-2}(\omega)\|\sigma n\rho\omega) = (\Omega(\sigma n'\omega') - \Omega(\sigma n\omega))(\sigma n'\rho'\omega'\|z\|\sigma n\rho\omega), \tag{6.5}$$

where the matrix elements are SU(3)-reduced matrix elements in V^{σ}_{ub}.

It follows that, if ω and ω' are both simple,

$$\frac{k(\omega')}{k(\omega)} = (\Omega(\sigma n'\omega') - \Omega(\sigma n\omega))^{1/2}, \tag{6.6}$$

giving immediately the expression (2.18) for the matrix elements of $\gamma(A_{ij}) = \kappa z_{ij}\kappa^{-1}$ between simple states. Hence we derive the analytic expression (2.19) for matrix elements of the sp(3,R) algebra between simple states.

In general, from Eq. (2.17), we derive the recursion relation

$$(i\omega|\kappa^2|j\omega) = \frac{1}{N(j)}\sum_{k\omega'\delta}[\Omega(i\omega) - \Omega(k\omega')](k\omega'|\kappa^2|l\omega') \times (i\omega\|a^{\dagger}\|k\omega')_{\delta}(j\omega\|a^{\dagger}\|l\omega')^{*}_{\delta}, \tag{6.7}$$

where

$$N(j) = \sum_{ik}(j\omega|a^{\dagger}_{ik}a_{ki}|j\omega) = \sum_{l} n_l(j),$$

for the matrix elements of $\kappa^2(\omega)$ which are easily solved at each weight starting from the lowest.

If we neglect the off-diagonal elements of $\kappa^2(\omega)$, then Eq. (6.5) gives the approximate expression

$$\frac{\kappa_{ii}(\omega')}{\kappa_{ll}(\omega)} \cong (\Omega(i\omega') - \Omega(l\omega))^{1/2}, \tag{6.8}$$

and hence the approximate sp(3,R) matrix elements of Eq. (2.27).

The procedure is illustrated in Tables I and II which give some matrix elements for the representation $\sigma = (20,13,10)$ [i.e., the representation $N_0(\lambda_0\mu_0) = 43(7,3)$ in a U(1)×SU(3) classification] calculated with a hand calculator using the above expressions. The exact matrix elements for simple states are obtained quite trivially from the known boson matrix elements, listed in Table II. As shown in Table I, the $\omega = (22,15,10)$ states have a twofold multiplicity and one must evaluate the expressions given by Eq. (6.7), for the matrix elements of $\kappa^2(\omega)$. The result is given in Table I and used to calculate the exact matrix elements given in Table II, from the expression (2.26).

The matrix elements given in Table II under "improved approximation" were calculated directly using Eq. (2.27). One confirms that they are precise for simple states and, as expected from previous comparisons with numerically computed sp(3,R) matrix elements,[22] they are accurate for the representation in question to ~0.1%. For large $(\sigma_1,\sigma_2,\sigma_3)$, such as occur in the application of the sp(3,R) model to rare-earth rotational nuclei, they are more accurate. For example, for the representation $N_0(\lambda_0\mu_0) = 733(82,0)$ used to describe rotational states in ^{154}Sm, the approximate analytic expression (2.27) gave matrix elements accurate to two parts in 10^6, which is more than sufficient for most conceivable practical applications. For smaller λ_0 and larger $N_0 = 3\sigma$, the results are still more accurate.

VII. AN APPROXIMATE REALIZATION OF sp(3,R) AND A HOLSTEIN–PRIMAKOFF EXPANSION

A simple algorithm was given in Sec. VI for the computation of the matrix elements of the Hermitian operator κ, which appears in the expression (2.12) for the realization γ of sp(3,R). It would be very useful to obtain an explicit realization for κ. The fact that we obtain exact analytic matrix elements for the representation $\sigma_1 = \sigma_2 = \sigma_3$ suggests that, for such representations, it should be possible. This is indeed the case, as we now show. Furthermore, the fact that accurate approximate matrix elements are given by an analytic expression in the general case suggests a corresponding approximate realization of κ, which we also find.

First observe that if we define $\Lambda''\cdot z$ and $\nabla\cdot\Lambda''$ by Eq.

TABLE II. Comparison of approximate (analytic) and exact sp(3,R) matrix elements evaluated for the (20,13,10) representation by the method described in the text.

			$\langle i\|\|A\|\|j\rangle$		
i	j	$(i\|\|a^{\dagger}\|\|j)$	u(3)-boson approximation	improved approximation	exact
2	1	1	$\sqrt{86/3} = 5.541$	$\sqrt{40} = 6.3245$	$\sqrt{40} = 6.3245$
3	1	−1	$-\sqrt{86/3} = -5.541$	$-\sqrt{31} = -5.5678$	$-\sqrt{31} = -5.5678$
4	1	1	$\sqrt{86/3} = 5.541$	$\sqrt{24} = 4.8990$	$\sqrt{24} = 4.8990$
8,1	2	$\sqrt{5/6}$	$\sqrt{215/9} = 4.888$	$\sqrt{125/6} = 4.5644$	4.5634
8,1	3	$\sqrt{2/3}$	$\sqrt{172/9} = 4.371$	$\sqrt{68/3} = 4.7610$	4.7631
8,1	4	$\sqrt{1/2}$	$\sqrt{43/3} = 3.786$	$\sqrt{41/2} = 4.5277$	4.5264
8,2	2	$\sqrt{7/36}$	$\sqrt{301/54} = 2.361$	$\sqrt{77/18} = 2.0683$	2.0704
8,3	3	$-\sqrt{80/63}$	$-\sqrt{6880/189}$ $= -6.033$	$-\sqrt{2480/63}$ $= -6.2742$	−6.2725
8,4	4	$\sqrt{15/28}$	$\sqrt{215/14} = 3.919$	$\sqrt{285/14} = 4.5119$	4.5132

(2.29) then, from Eq. (2.15) and its Hermitian adjoint, we obtain

$$\Lambda\cdot z_{ij} = \kappa^2 z_{ij}\kappa^{-2}, \quad \nabla_{ij}\cdot\Lambda = \kappa^{-2}\nabla_{ij}\kappa^2. \tag{7.1}$$

Thus, if, as in the derivation of the approximate expression (2.27) for sp(3,R) matrix elements, we assume that Λ and κ are simultaneously diagonal and hence commute, we obtain

$$\Lambda^n\cdot z_{ij} = \kappa^{2n} z_{ij}\kappa^{-2n}, \quad \nabla_{ij}\cdot\Lambda^n = \kappa^{-2n}\nabla_{ij}\kappa^{2n}. \tag{7.2}$$

If we then formally extend the definition (2.29) to half-integer n, we obtain, with the approximation $[\Lambda,\kappa]\cong 0$,

$$\Lambda^{1/2}\cdot z_{ij} \cong \kappa z_{ij}\kappa^{-1}, \quad \nabla_{ij}\cdot\Lambda^{1/2} \cong \kappa^{-1}\nabla_{ij}\kappa, \tag{7.3}$$

and hence Eq. (2.28).

To define $\Lambda^{1/2}$, note that we may write

$$\begin{aligned} &\Lambda = \Lambda_0 + \Lambda_1, \\ &\Lambda_0 = \tfrac{1}{3}\mathrm{tr}(\mathbb{C})\mathrm{tr}(z\nabla), \\ &\Lambda_1 = \Lambda - \Lambda_0. \end{aligned} \tag{7.4}$$

Within a representation $(\sigma_1,\sigma_2,\sigma_3)$,

$$\Lambda_0\cdot z_{ij} = 2\sigma z_{ij}.$$

Thus we may equate Λ_0 with the operator $2\sigma I$, where I is the identity. Hence we define

$$\Lambda^{1/2} = (\Lambda_0 + \Lambda_1)^{1/2} = \sqrt{2\sigma}\sum_{k=0}^{\infty}\binom{1/2}{k}(2\sigma)^{-k}\Lambda_1^k, \tag{7.5}$$

which is meaningful whenever the series is convergent. From the expression (2.16) for the eigenvalues of Λ, we deduce that matrix elements of

$$\Lambda_1^k\cdot z_{ij} \quad \text{and} \quad \nabla_{ij}\cdot\Lambda_1^k,$$

between low weight states, are of order of magnitude $(\lambda_\sigma+\mu_\sigma)^k$, where λ_σ and μ_σ are given by Eq. (2.24). Thus, the series is well-defined for $(\lambda_\sigma+\mu_\sigma)\ll 2\sigma$.

The extent to which this approximate realization satisfies the sp(3,R) commutation relations is a measure of its accuracy. As shown in the Appendix,

$$\begin{aligned} &[\nabla_{ij}\cdot\Lambda^{1/2}, \Lambda^{1/2}\cdot z_{lk}] \\ &\quad = \gamma([B_{ij},A_{lk}]) - [1/8(2\sigma)^3]X_2^{(2)} + \cdots, \\ &[\Lambda^{1/2}\cdot z_{ij}, \Lambda^{1/2}\cdot z_{lk}] \\ &\quad = \gamma([A_{ij},A_{lk}]) \\ &\quad - [1/8(2\sigma)^3](Y_0^{(4)} + 2Y_1^{(3)} + Y_2^{(2)}) + \cdots, \end{aligned} \tag{7.6}$$

where

$$\begin{aligned} X_n^{(m)} &= [\nabla_{ij}, \Lambda_1^m\cdot z_{lk}]\cdot\Lambda_1^n, \\ Y_n^{(m)} &= [z_{ij}, \Lambda_1^m\cdot z_{lk}]\cdot\Lambda_1^n. \end{aligned} \tag{7.7}$$

The commutation relations of $\Lambda^{1/2}\cdot z_{ij}$ and $\nabla_{ij}\cdot\Lambda^{1/2}$ with $\gamma(C_{lk})$ are exactly satisfied. Since the matrix elements of $\Lambda^{1/2}\cdot z_{ij}$ and $\nabla_{ij}\cdot\Lambda^{1/2}$ are each of order $\sqrt{2\sigma}$, it follows that, these approximations for $\gamma(A_{ij})$ and $\gamma(B_{ij})$, respectively, are in error by a factor of order $[(\lambda_\sigma+\mu_\sigma)/2\sigma]^4$. We note too that, when $\sigma_1=\sigma_2=\sigma_3$ (i.e., $\lambda_\sigma=\mu_\sigma=0$), Λ commutes precisely with κ and the approximation becomes exact.

Finally, to obtain the generic Holstein–Primakoff realization, one must make the substitution $z_{ij}\to a_{ij}^\dagger, \nabla_{ij}\to a_{ij}$ in the expressions for $\gamma(X)$, $X\in$ sp(3,R). Since the approximate expression (2.28) is accurate to fourth order, it clearly generates the first three terms of the expansion correctly. The first two terms are given in Eq. (2.30). For the $\sigma_1=\sigma_2=\sigma_3$ representations, Eq. (2.28) generates precisely the infinite Holstein–Primakoff expansion.

VIII. CONCLUDING REMARKS

The coherent state realization of the sp(3,R) algebra has proved to be remarkably powerful. However, its utility appears not to be so much in providing a Hilbert space, like Bargmann space, in which calculations are simplified, as in providing an intermediary step towards finding the relationship between the sp(3,R) and the simpler u(3)-boson algebras. As a result we have been able to discover analytic expressions for the sp(3,R) matrix elements that are exact between simple states and remarkably accurate generally. This is a result of considerable interest which saves substantial computer time in sp(3,R) model calculations[23] by obviating the numerical methods used previously.[24,21]

It is of interest to note that our coherent state representations are similar to, but of a different type than, those employed in the previous works on the subject.[17–19] Deenen and Quesne,[19] for example, found for the special case $(\sigma_1=\sigma_2=\sigma_3=N/2)$, the realization of the sp(3,R) algebra

$$\begin{aligned} &Q(A_{ij}) = \omega_{ij}, \\ &Q(B_{ij}) = N\Delta_{ij} - 4\Delta_{ij} + (\Delta\omega\Delta)_{ij}, \\ &Q(C_{ij}) = (N/2)\delta_{ij} + (\omega\Delta)_{ij}, \end{aligned} \tag{8.1}$$

where $\Delta_{ij} = (1+\delta_{ij})\partial/\partial\omega_{ij}$. This realization was obtained, following Barut and Girardello,[17] by a change of variables

$$\omega_{ij} = \sum_{n=1}^{N} x_{in}x_{jn}, \tag{8.2}$$

from the standard Bargmann coherent state realization[2]

$$A_{ij} = \sum_n b_{in}^\dagger\, b_{jn}^\dagger \to \sum_n x_{in}x_{jn}. \tag{8.3}$$

It is sometimes called a Barut realization. One notes that the Barut realization is evidently isomorphic to ours for the special case $(\sigma_1=\sigma_2=\sigma_3=N/2)$ in which $\mathbb{C}_{ij}\to\delta_{ij}N/2$. However, the differences are interesting. As Deenen and Quesne showed, their (Barut) coherent states are eigenstates of the B_{ij} lowering operators. In contrast, our coherent states are translates of the lowest weight state by the group action, in accord with the definitions of Perelomov[9] and Onofri.[10] For the Bargmann coherent states of the Weyl algebra,[2] the two definitions are, of course, equivalent but they extend differently to more complicated groups. It seems rather remarkable, though, that the two distinct definitions of coherent states should result in such transparently isomorphic results, even for the special case $(\sigma_1=\sigma_2=\sigma_3)$. It would be very interesting therefore to see if this correspondence extends to the generic representations and to other groups.

In conclusion, we note another interesting characteristic. Whereas the coherent state theory for the noncompact sp(3,R) group has some complications over that for a compact group, it also has some extra nice features. In particular, the (coherent state related) boson expansions for a compact Lie algebra must be restricted to what is sometimes called

the "physical subspace."[6] For example, in the bosonization of the SO(2N) fermion pair algebra,[7] one must restrict the bosons to the image of states accessible to fermions obeying the Pauli exclusion principle. In contrast, the generic ($\sigma_3 > 3$) U(3)-boson representation spaces are isomorphic to those of sp(3,R). Thus they contain no spurious "nonphysical" states. A consequence of this is that the Holstein–Primakoff realization of the sp(3,R) algebra is manifestly Hermitian whereas in general, as observed by Dobaczewski,[7] the Holstein–Primakoff realization is Hermitian only on the physical subspace.

Further applications of these techniques will be given in a paper (with B. G. Wybourne and P. H. Butler) to follow.

APPENDIX: ERROR ANALYSIS

The objective is to show, by proving Eq. (7.6), that the approximate realization (2.28) of the sp(3,R) algebra satisfies the sp(3,R) commutation relations up to terms of order $[(\lambda_\sigma + \mu_\sigma)/2\sigma]^4$.

From the definition (7.5) of $\Lambda^{1/2}$, it follows that

$$\nabla_{ij}\cdot\Lambda^{1/2} = \sqrt{2\sigma}\sum_{m=0}^{\infty}\binom{1/2}{m}(2\sigma)^{-m}F_{ij}^{(m)}, \tag{A1}$$

$$\Lambda^{1/2}\cdot z_{lk} = \sqrt{2\sigma}\sum_{n=0}^{\infty}\binom{1/2}{n}(2\sigma)^{-n}G_{lk}^{(n)}, \tag{A2}$$

where

$$F_{ij}^{(m)} = \nabla_{ij}\cdot\Lambda_1^m, \tag{A3}$$

$$G_{lk}^{(n)} = \Lambda_1^n\cdot z_{lk}. \tag{A4}$$

In expanding the commutator $[\nabla_{ij}\cdot\Lambda^{1/2},\Lambda^{1/2}\cdot z_{lk}]$, terms are encountered of the type

$$X^{(m,n)} = [F_{ij}^{(m)},G_{lk}^{(n)}]. \tag{A5}$$

Now, from the Jacobi identity,

$$[F_{ij}^{(m)},G_{lk}^{(n)}] = [F_{ij}^{m-1},G_{lk}^{(n+1)}] + [[F_{ij}^{(m-1)},G_{lk}^{(n)}],\Lambda_1], \tag{A6}$$

which by repeated use leads to the equation

$$X^{(m,n)} = X^{(0,m+n)} + \sum_{k=1}^{m}[X^{(m-k,n+k-1)},\Lambda_1]. \tag{A7}$$

Repeated use of Eq. (A7) then leads to the expression

$$X^{(m,n)} = \sum_{k=0}^{m}\binom{m}{k}\ X_k^{(m+n-k)}, \tag{A8}$$

where $X_k^{(m)}$ is defined by

$$X_{k+1}^{(m)} = [X_k^{(m)},\Lambda_1], \quad X_0^{(m)} = X^{(0,m)}. \tag{A9}$$

Since

$$[\nabla_{ij},\Lambda^{1/2}\cdot G_{lk}^{(n)}] = \sqrt{2\sigma}\sum_{m=0}^{\infty}\binom{1/2}{m}\ (2\sigma)^{-m}X^{(0,m+n)}, \tag{A10}$$

it follows from Eqs. (A1) and (A8) that

$$[\nabla_{ij}\cdot\Lambda^{1/2},G_{lk}^{(n)}] = [\nabla_{ij},\Lambda^{1/2}\cdot G_{lk}^{(n)}] + \sqrt{2\sigma}\sum_{m}\sum_{k=1}^{m}\binom{1/2}{m}\binom{m}{k} \times(2\sigma)^{-m}X_k^{(m+n-k)}. \tag{A11}$$

Hence we obtain

$$[\nabla_{ij}\cdot\Lambda^{1/2},\Lambda^{1/2}\cdot z_{lk}] = [\nabla_{ij},\Lambda\cdot z_{lk}] + 2\sigma\sum_{m,n}\sum_{k=1}^{m}\binom{1/2}{m}\binom{1/2}{n}\binom{m}{k} \times(2\sigma)^{-(m+n)}X_k^{(m+n-k)}. \tag{A12}$$

Now the first term on the rhs is just $\Gamma([B_{ij},A_{lk}])$ which by Eq. (2.12) is equal to $\gamma([B_{ij},A_{lk}])$. Hence

$$[\nabla_{ij}\cdot\Lambda^{1/2},\Lambda^{1/2}\cdot z_{lk}] = \gamma([B_{ij},A_{lk}]) + 2\sigma\sum_{m,n}\sum_{k=1}^{m}\binom{1/2}{m}\binom{1/2}{n}\binom{m}{k} \times(2\sigma)^{-(m+n)}X_k^{(m+n-k)}. \tag{A13}$$

Thus the second term on the rhs is a measure of the error in the approximation

$$\gamma(B_{ij})\cong\nabla_{ij}\cdot\Lambda^{1/2}, \quad \gamma(A_{lk})\cong\Lambda^{1/2}\cdot z_{lk}. \tag{A14}$$

An order by order evaluation of this error term, in inverse powers of 2σ, reveals that the first three terms vanish identically, which is why the approximation is so accurate. The fourth-order correction term is given in Eq. (7.6).

The commutator $[\Lambda^{1/2}\cdot z_{ij},\Lambda^{1/2}\cdot z_{lk}]$ is evaluated in a similar way. The commutator

$$Y^{(m,n)} = [G_{ij}^{(m)},G_{lk}^{(n)}] \tag{A15}$$

is shown to satisfy the equation

$$Y^{(m,n)} = (-1)^m\sum_{k=0}^{m}\binom{m}{k}Y_k^{(m+n-k)}, \tag{A16}$$

with $Y_k^{(m)}$ defined by

$$Y_{k+1}^{(m)} = [Y_k^{(m)},\Lambda_1] \quad Y_0^{(m)} = Y^{(0,m)}. \tag{A17}$$

Thus, in parallel with the derivation of Eq. (A13), we obtain

$$[\Lambda^{1/2}\cdot z_{ij},\Lambda^{1/2}\cdot z_{lk}] = 2\sigma\sum_{m,n}\sum_{k=0}^{m}(-1)^m\binom{1/2}{m}\binom{1/2}{n}\binom{m}{k} \times(2\sigma)^{-(m+n)}\ Y_k^{(m+n-k)}. \tag{A18}$$

In evaluating the rhs, it is important to use the identity

$$[\Lambda\cdot z_{ij},\Lambda\cdot z_{lk}] = \Gamma([A_{ij},A_{lk}]) = 0, \tag{A19}$$

which implies the relationship

$$Y^{(2)} + Y_1^{(1)} = 0. \tag{A20}$$

An order by order evaluation of the rhs of Eq. (A18) then reveals that the first three terms vanish identically. The fourth-order term is given in Eq. (7.6)

Note that for a representation of the type ($\sigma_1 = \sigma_2 = \sigma_3 = \sigma$), in which the u(3) operators reduce to

$$\gamma(\mathbb{C}_{ij}) = \delta_{ij}\sigma I + (z\nabla)_{ij}, \tag{A21}$$

where I is the identity, all the correction terms in the expansions (A13) and (A18) vanish and the approximations (2.28) become exact.

[1] R. I. Glauber, Phys. Rev. **130**, 2529 (1963); **131**, 2766 (1963).
[2] V. Bargmann, Commun. Pure Appl. Math. **14**, 187 (1961); **20**, 1(1967).
[3] D. J. Rowe and R. Basserman, Can. J. Phys. **54**, 1941 (1976); D. J. Rowe and A. G. Ryman, Phys. Rev. Lett. **45**, 406 (1980); D. J. Rowe, A. Ryman, and G. Rosensteel, Phys. Rev. A **22**, 2362 (1980); P. Kramer and M. Saraceno, *Lecture Notes in Physics* (Springer, Berlin, 1981), Vol. 140.
[4] G. Rosensteel and D. J. Rowe, Phys. Rev. Lett. **38**, 10 (1977); Ann Phys.

(N.Y.) **126**, 343 (1980).
[5]G. F. Filippov, Yu. V. Teryoshin, and V. S. Vasilevsky, "On generating invariants of Sp(6,R) model," Kiev report ITP-81-103E, Kiev, 1981.
[6]P. Garbaczewski, Phys. Rep. **36**, 65 (1978); E. R. Marshalek, Prog. Theor. Phys. Suppl. **74** and **75**, 89 (1983).
[7]J. Dobaczewski, Nucl. Phys. A **369**, 213 (1981); A **369**, 237 (1981); A **380**, 1 (1982).
[8]K. T. Hecht, E. J. Reske, T. H. Seligman, and W. Zahn, Nucl. Phys. A **356**, 146 (1981); Y. Suzuki, Nucl. Phys. A **405**, 40 (1983).
[9]A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[10]E. Onofri, J. Math. Phys. **16**, 1087 (1975).
[11]G. W. Mackey, Bull. Am. Math. Soc. **69**, 628 (1963).
[12]R. Godement, "Seminaire Cartan," École Nationale Supérieure, Paris, 1958; G. Rosensteel and D. J. Rowe, Int. J. Theor. Phys. **16**, 63 (1977).
[13]G. Rosensteel and D. J. Rowe, Phys. Rev. Lett. **47**, 223 (1981); D. J. Rowe and G. Rosensteel, Phys. Rev. C **25**, 3236 (1982).
[14]Yu F. Smirnov (private communication); cf. also C. Burdik, M. Havlicek, Yu F. Smirnov, V. N. Tolstoy, and A. A. Sachuruk, "About one construction of the boson realization for the simple Lie algebra," Charles University report, 1983.
[15]F. J. Dyson, Phys. Rev. **102**, 1217 (1956).
[16]T. Holstein and H. Primakoff, Phys. Rev. **58**, 1098 (1940).
[17]A. O. Barut and L. Girardello, Commun. Math. Phys. **21**, 41 (1971).
[18]L. D. Mlodinow and N. Papanicolau, Ann. Phys. (N.Y.) **131**, 1 (1981).
[19]J. Deenen and C. Quesne, J. Math. Phys. **23**, 878 (1982); **23**, 2004 (1982).
[20]O. Castãnos, E. Chacón, and M. Moshinsky, J. Math. Phys. **25**, 1211 (1984).
[21]G. Rosensteel and D. J. Rowe, J. Math. Phys. **24**, 2461, (1983); cf. also C. Quesne, J. Math. Phys. **22**, 1482 (1981).
[22]D. J. Rowe, G. Rosensteel, and R. Carr, J. Phys. A: Math. Gen. **17**, L399 (1984).
[23]P. Park, J. Carvalho, M. Vassanji, D. J. Rowe, and G. Rosensteel, Nucl. Phys. A **414**, 93 (1984).
[24]G. Rosensteel, J. Math. Phys. **21**, 924 (1980).

Remarks about inverse diffraction problem[a)]

A. G. Ramm
Mathematics Department, Kansas State University, Manhattan, Kansas 66506

(Received 10 August 1983; accepted for publication 20 January 1984)

From the scattering data one finds the support function or the principal curvatures of the surface of a reflecting obstacle. From either of these data the surface is effectively reconstructed.

PACS numbers: 03.80. + r

I. INTRODUCTION

Let D be a bounded domain (an obstacle) with a smooth boundary Γ, and $\Omega = R^3 \setminus D$ be the exterior domain. It is well known (see, e.g. Ref. 1) that the scattering amplitude f for the problem

$$(\nabla^2 + k^2)u = 0 \quad \text{in } \Gamma, \quad k > 0, \tag{1}$$

$$u = 0 \quad \text{on } \Gamma, \tag{2}$$

$$u = \exp[ik(v,x)] + v, \tag{3}$$

$$v \sim \frac{\exp(ikr)}{r} f(n,v,k) \quad \text{as } |x| = r \to \infty, \quad \frac{x}{r} = n, \tag{4}$$

in the high-frequency ($k \to \infty$) approximation determines the Gaussian curvature K of Γ.

It was shown in Ref. 1 (see also Ref. 2) that f determines the support function $a(l)$ of Γ explicitly and the parametric equations of Γ are

$$x_j = -\frac{\partial a(l)}{\partial \alpha_j}, \quad j = 1,2,3. \tag{5}$$

Here $l = (\alpha_1,\alpha_2,\alpha_3)$ is a unit vector of normal to Γ at the point s_0, which is uniquely determined by v and n if Γ is strictly convex. Namely, $l = (n - v)/|n - v|$ and s_0 is the point at which the expression (l,s) is stationary on Γ_+, where Γ_+ is the illuminated part of Γ. The function $a(l)$ was defined in Ref. 2 as $a(l) = -\max_{s \in \Gamma_+}(l,s)$. The function $a(l)$ is a homogeneous function of $\alpha_1,\alpha_2,\alpha_3$ of order 1:

$$\frac{\partial a}{\partial \alpha_j}\alpha_j = a, \tag{6}$$

one should sum over the repeated indices here and below. The basic result of Ref. 2 gives an algorithm for a stable calculation of Γ from the knowledge of the scattering amplitude $f(n,v,k)$ for n,v such that $(n - v)/|n - v|$ runs through all of the unit sphere S^2 (e.g., for $n = -v$ and v runs through S^2, the backscattering case) and for two values of k, $k = k_1$, and $k_2 \neq k_1$ such that the high-frequency asymptotic formula for f holds[2]:

$$f_j = f(n,v,k_j) = -\frac{\exp[2ik_j(n,l)a(l)]}{2[K(s_0)]^{1/2}}(1 + \epsilon_j), \quad \epsilon_j \to 0 \quad \text{as } k_j \to \infty. \tag{7}$$

Here $K(s_0)$ is the Gaussian curvature of Γ at the point r_0. If the f_j are measured with the accuracy δ, i.e., $|f_{j\delta} - f_j| < \delta$ and $f_{j\delta}$ are the measurements, and if $K(s_0) \leqslant d^2$, $d = \text{const}$, then

$$\frac{f_{1\delta}}{f_{2\delta}} = \exp[2i(n,l)a(l)(k_1 - k_2)] \times [1 + O(|\epsilon_1| + |\epsilon_2|) + O(\delta d)].$$

Thus

$$a(l) = \frac{1}{2i(n,l)(k_1 - k_2)} \ln \frac{f_{1\delta}}{f_{2\delta}} + O\left(\frac{|\epsilon_1| + |\epsilon_2| + \delta d}{|k_1 - k_2|\,|(n,l)|}\right). \tag{8}$$

The knowledge of $a(l)$ with a known error allows one to recover the surface stably using the following estimates of the derivatives of a function $a(l)$ known with an error

$$\eta = O\left(\frac{|\epsilon_1| + |\epsilon_2| + \delta d}{|k_1 - k_2|\,|(n,l)|}\right).$$

Assume that $|a''| \leqslant M$, where a'' denotes any second derivative of $a(\alpha_1,\alpha_2,\alpha_3)$, and assume that a_η is known, $|a_\eta - a| < \eta$. Let $h = (2\eta/M)^{1/2}$, $\epsilon = \sqrt{2M\eta}$. Then $|[a_\eta(\alpha + hb) - a_\eta(\alpha - hb)]/2h - \partial a/\partial b| < \epsilon = \sqrt{2M\eta}$ (Ref. 3). Here b is any unit vector, $\partial a/\partial b = (\nabla_\alpha a, b)$ is the derivative of a in the direction b. All these facts are proved in Ref. 2 and provide a stable and effective algorithm for recovering the surface from the scattering data. One should be able to measure the phase of the scattered field in order to use the above algorithm. This is not very easy in practice. Therefore one can think of recovering the support function of Γ from some other data.

It is clear from (7) that the Gaussian curvature $K = K_1K_2$, where the K_j are the principal curvatures, can be determined from the measurements of $|f|^2$. Let us assume that the quantity $K_1 - K_2$ can be measured.[4] (In Ref. 4 a possibility to determine $K_1 - K_2$ from the measurements of the scattered field was reported for electromagnetic scattering from a metallic body. It is not clear if $K_1 - K_2$ can be found from the measurements of the scattered field in acoustic problems.) Then the quantity $h = K_1^{-1} + K_2^{-1}$ is known as a function of the unit normal $l = (\alpha_1,\alpha_2,\alpha_3)$ to Γ. From this data the support function of Γ can be recovered and then the parametric equation of Γ is given by (5). Notice, that if the origin of the coordinate system is placed inside the convex obstacle D, then $a(l) = \max_{s \in \Gamma}(l,f)$, and (5) takes the form

$$x_j = \frac{\partial a}{\partial \alpha_j}, \quad 1 \leqslant j \leqslant 3. \tag{5'}$$

II. RECOVERING THE SURFACE FROM $h = K_1^{-1} + K_2^{-1}$

In this section a construction of Γ from h is given. A similar construction can be found in the literature.[5] Since it

[a)] Prepared while the author was visiting Schlumberger–Doll Research, P. O. B. 307, Ridgefield, CT 06877.

 0022-2488/84/092672-03$02.50

was not used in the inverse scattering problem, the construction is described in sufficient detail for convenience of the reader. It is worth mentioning that the problems of recovering a surface from its Gaussian curvature K_1K_2 or its mean curvature $(K_1+K_2)/2$ were studied in the literature extensively. Both problems lead to nonlinear elliptic equations which cannot be solved explicitly. In contrast, the problem we are discussing here will lead to the Laplace equation which is effectively solvable:

$$\frac{\partial^2 a}{\partial\alpha_1^2}+\frac{\partial^2 a}{\partial\alpha_2^2}+\frac{\partial^2 a}{\partial\alpha_3^2}\equiv\nabla_\alpha^2 a(\alpha_1,\alpha_2,\alpha_3)=h. \tag{9}$$

To derive (9) one takes the Rodriguez formula

$$dx_j - R\,d\alpha_j = 0, \quad 1\leqslant j\leqslant 3, \tag{10}$$

in which $x=x(\alpha_1,\alpha_2,\alpha_3)$ is the radius vector of a point of the surface Γ, and the unit vector $l=(\alpha_1,\alpha_2,\alpha_3)$ serves as a parameter. There is a 1–1 correspondence between the parameter $l=(\alpha_1,\alpha_2,\alpha_3)$ and the parameter $(\theta,\phi)\in S^2$. Namely, l is determined by the point $(\theta,\phi)\in S^2$ and determines this point. Since l is also the outer unit normal to Γ, we write $d\alpha_j$ in (10) instead of $-dN_j$ as in Ref. 5, where N is the interior unit normal to Γ. Finally, R in (10) is the radius of the curvature of the normal cut of Γ in the principal direction $(d\alpha_1,d\alpha_2,d\alpha_3)$. From (5′) and (10) it follows that

$$\begin{aligned}0&=\frac{\partial^2 a}{\partial\alpha_j\partial\alpha_i}d\alpha_i - R\,d\alpha_j\\&=\left(\frac{\partial^2 a}{\partial\alpha_j\partial\alpha_i}-R\,\delta_{ji}\right)d\alpha_i,\\ \delta_{ji}&=\begin{cases}1, & i=j,\\ 0, & i\neq j.\end{cases}\end{aligned} \tag{11}$$

Since the vector $(d\alpha_1,d\alpha_2,d\alpha_3)\neq 0$ one concludes from (11) that

$$\det\left(\frac{\partial^2 a}{\partial\alpha_j\,\partial\alpha_i}-R\delta_{ji}\right)=0. \tag{12}$$

But $\det \partial^2 a/\partial\alpha_j\,\partial\alpha_i = 0$. Indeed, differentiate (6) in α_i to obtain

$$\frac{\partial^2 a}{\partial\alpha_j\,\partial\alpha_i}\alpha_j+\frac{\partial a}{\partial\alpha_j}\delta_{ij}=\frac{\partial a}{\partial\alpha_i} \quad\text{or}\quad \frac{\partial^2 a}{\partial\alpha_j\,\partial\alpha_i}\alpha_j=0. \tag{13}$$

Since $(\alpha_1,\alpha_2,\alpha_3)\neq 0$ one concludes from (13) that

$$\det\frac{\partial^2 a}{\partial\alpha_j\,\partial\alpha_i}=0. \tag{14}$$

From (14) it follows that Eq. (12) has a root $R=0$. Its two other roots are the principal radii $R_j=K_j^{-1}$, where K_j are the principal curvatures. Since the body is assumed to be convex, $R_j>0$. The trace of the matrix $\partial^2 a/\partial\alpha_j\,\partial\alpha_i$ is equal to R_1+R_2+0, where R_1, R_2, 0 are the roots of (12). Since $R_1+R_2\equiv h$ one obtains Eq. (9).

It remains to solve this equation. Let us expand $h=h(\alpha_1,\alpha_2,\alpha_3)=h(\theta,\phi)$ in spherical harmonics:

$$h=\sum_{n>0}h_{nm}Y_{nm}(\theta,\phi), \tag{15}$$

where the Y_{nm} are the normalized in $L^2(S^2)$ spherical harmonics,

$$Y_{nm}=\left(\frac{(2n+1)(n-m)!}{4\pi(n+m)!}\right)^{1/2}P_{n,m}(\cos\theta)e^{im\phi},$$
$$-n\leqslant m\leqslant n,$$

$P_{n,m}(x)$ are associated Legendre functions, and the h_{nm} are the Fourier coefficients of h,

$$h_{nm}=\int_{S^2}h(\theta,\phi)\,\overline{Y_{nm}}\,d\omega, \quad d\omega=\sin\theta\,d\theta\,d\phi. \tag{16}$$

The function $a(\alpha_1,\alpha_2,\alpha_3)$ as a homogeneous function of order 1 can be expanded in the series $(A=a(r\alpha_1,r\alpha_2,r\alpha_3)=ra(\alpha_1,\alpha_2,\alpha_3))$.

$$A=r\sum_{n>0}A_{nm}Y_{nm}, \tag{17}$$

where $r>0$, and the function (17) is homogeneous of order 1. The point (r,θ,ϕ) corresponds to $(r\alpha_1,r\alpha_2,r\alpha_3)$.

Let us consider (r,θ,ϕ) as spherical coordinates and (y_1,y_2,y_3) as the corresponding Cartesian coordinates. Then $r=(\Sigma_{j=1}^3 y_j^2)^{1/2}$,

$$\begin{aligned}\nabla_y^2 rY_{nm}&=\left(\frac{\partial^2}{\partial y_1^2}+\frac{\partial^2}{\partial y_2^2}+\frac{\partial^2}{\partial y_3^2}\right)rY_{nm}\\&=-\frac{(n-1)(n+2)}{r}Y_{nm},\end{aligned} \tag{18}$$

and from (17) and (18) it follows that

$$\nabla_y^2 A=-\sum_{n>0}A_{nm}\frac{(n-1)(n+2)}{r}Y_{nm}. \tag{19}$$

Substitute (19) and (15) in (9), take $r=1$, and equate the coefficients in front of Y_{nm} to obtain

$$A_{nm}=-h_{nm}/(n-1)(n+2), \quad n\neq 1. \tag{20}$$

A necessary condition for a function $h(\theta,\phi)$ to be equal to R_1+R_2, where R_1,R_2 are the principal radii of a closed surface, is the equation

$$h_{1m}=0, \quad m=0,\pm 1. \tag{21}$$

This will be proved shortly. Assuming (21), one obtains an effective formula for the support function from (17) with $r=1$,

$$a(l)=-\sum_{\substack{n>0\\ n\neq 1}}^{\infty}\frac{h_{nm}}{(n-1)(n+2)}Y_{nm}(\theta,\phi). \tag{22}$$

Here l is the unit vector corresponding to the point $(\theta,\phi)\in S^2$, and the h_{nm} were defined in (16).

It remains to check (21). Let $d\sigma$ be the element of the surface area of Γ, $d\sigma=R_1\,ds_1\,R_2\,ds_2$, where $ds_1\,ds_2=d\omega$ is the element of the area of S^2, $|R_j\,ds_j|$, $j=1,2$, are the elements of the length along the lines of principal curvatures on Γ, and $ds_1\,ds_2$ is the inner product of two vectors. Consider a family Γ_z of parallel surfaces (for which the normals are the same as for $\Gamma_0=\Gamma$). Then

$$\begin{aligned}d\sigma_z&=(R_1+z)ds_1(R_2+z)ds_2\\&=d\sigma+z(R_1+R_2)d\omega+z^2\,d\omega.\end{aligned} \tag{23}$$

From the Gauss formula it follows that

$$\int_{\Gamma_z}N_j\,d\sigma_z=0, \tag{24}$$

where N_j is the jth component of the outer unit normal N to Γ_z, $0<z<z_0$, z is arbitrary, and z_0 small, so that one can

assume that N does not depend on z. Substitute (23) in (24) to obtain

$$\int_{S^2} h(l)N_j\, d\omega = \int_{S^2} (R_1 + R_2)N_j\, d\omega = 0, \quad 1 \leqslant j \leqslant 3. \tag{25}$$

But $N_j = l_j$ are three linearly independent vectors in the space spanned by Y_{1m}, $m = 0, \pm 1$. Therefore (25) is equivalent to (21) because $h = R_1 + R_2$ and h_{1m} is defined in (16). This completes the proof.

Example. If Γ is a sphere, then $R_1 = R_2 = R = \text{const}$, $h_{nm} = 0$ for $n > 0$. Therefore, $A_{nm} = 0$ for $n > 0$, $a(l) = A_0 Y_0, Y_0 = 1/\sqrt{4\pi}$, by formula (20), $A_0 = h_0/2$, $h_0 = 2R\sqrt{4\pi}$. Thus $a(l) = R$ and one recovers Γ from $h = 2R = \text{const}$.

In the above construction it is not guaranteed that the support function $a(l)$ found from the given function $h(l)$, satisfying the necessary conditions (21), will correspond to a convex body D. In Ref. 6 a necessary and sufficient condition on h for the corresponding $a(l)$ to be the support function of a convex body is given.

The main result from Ref. 6 for the three-dimensional case says: let $h(l)$ be a continuously differentiable function of the outer unit normal l on the unit sphere S^2. For $h(l)$ to be the sum of the principal radii of curvature of a convex surface Γ at the point at which the outer unit normal to Γ is l, it is necessary and sufficient that the condition (21) and the following condition (26) hold:

$$\int_{S^2} \frac{(l,l'')(\nabla h(l),l'')}{1 - (l,l')}\, d\omega \leqslant 0, \tag{26}$$

for all $l', l'' \in S^2$ for which $(l',l'') = 0$ with strict inequality in (26) for some choices of l' and l''. Here (l,l'') is the inner product, and $\nabla h(l)$ is to be calculated as follows: define $h(l)$ for all $l \in R^3$ by the rule

$$h(l) = \frac{1}{|l|} h\left(\frac{1}{|l|}\right), \quad |l| = \sqrt{\alpha_1^2 + \alpha_2^2 + \alpha_3^2}, \quad l = (\alpha_1,\alpha_2,\alpha_3), \tag{27}$$

calculate $\nabla h(l)$, and then set $|l| = 1$.

The reason for extending h as a positively homogeneous function of order -1 is that $a(l)$ is homogeneous of degree 1, $\nabla^2 a$ is homogeneous of degree -1, and thus Eq. (9) holds for all $(\alpha_1,\alpha_2,\alpha_3) \in R^3$.

III. ADDITIONAL GEOMETRICAL CONSIDERATIONS

Let $x(l)$ be a vector function on S^2. It is called the normal representation of a convex nondegenerate body D with the surface Γ if $x(l)$ is that point of Γ at which the outer normal to Γ is l. If Γ admits a tangent plane at any point then $x(l)$ is defined on all of S^2. Let us extend $x(l)$ from S^2 to R^3 by setting $x(u) = x(u/|u|)$, where $u \in R^3$, $|u| = (u_1^2 + u_2^2 + u_3^2)^{1/2}$, $u \neq 0$. The value $x(0)$ does not play any role. The support function of Γ is $(l,x(l)) = a(l)$ provided that the origin is inside D. One has

$$x_j = \frac{\partial a}{\partial \alpha_j}, \quad l = (\alpha_1,\alpha_2,\alpha_3), \quad 1 \leqslant j \leqslant 3,$$

$$\frac{\partial^2 a}{\partial \alpha_j\, \partial \alpha_i} = \frac{\partial x_i}{\partial \alpha_j} = \frac{\partial x_j}{\partial \alpha_i}, \tag{28}$$

the matrix $a_{ij} \equiv \partial^2 a/\partial\alpha_i \partial\alpha_j$ is semi-positive-definite. Since $a(\lambda l) = \lambda a(l), \lambda > 0$, we define $a(0) = 0$.

Differentiate Eq. (9) to obtain

$$\nabla^2 x_j(u) = \frac{\partial h}{\partial \alpha_j}, \tag{29}$$

where $h(u) = (1/|u|)h(u/|u|)$ [see (27)].

In Ref. 6 the following proposition is proved: a continuously differentiable vector function $x(u)$ is a normal representation of a convex Γ if and only if $\partial x_i(u)/\partial u_j$ is a semi-positive-definite symmetric matrix not identically zero.

The necessity follows from (28), and $a(l) = (l,x(l))$ is the support function of a convex Γ. The sufficiency also follows from (28): construct $a(l) = (l,x(l))$, Eq. (28) shows that the Hessian a_{ij} is a semi-positive-definite matrix. Since $x(u)$ is homogeneous of order 0, one has $(u,\partial x(u)/\partial u_j) = 0$, $1 \leqslant j \leqslant 3$. Therefore $a(u)$ is convex and, since $a(l)$ is convex and, since $a(l)$ is the support function of Γ, the surface Γ is convex.

This condition can be formulated in terms of the coefficients h_{nm} [see (15)] as semi-positive-definiteness of the matrix

$$a_{ij} = -\sum_{\substack{n \geqslant 0 \\ n \neq 1}} \frac{h_{nm}}{(n-1)(n+2)} \times \left(\frac{\partial^2}{\partial y_i\, \partial y_j} r Y_{nm}(\theta,\phi) \right)\Big|_{r=1} \geqslant 0, \quad \forall (\theta,\phi) \in S^2,$$

where $a_{ij} \geqslant 0$ means that $a_{ij} t_i t_j \geqslant 0$, $\forall t_i$, and y_j are the same as in formula (18).

Note added in proof: In Ref. 10 it was shown that $a(l)$ can be recovered stably from the knowledge of the scattering amplitude $f(n,\nu,K_1)$ at one high frequency and $n,\nu \in S^2$ such that $l = (n - \nu)/|n - \nu|$ runs through all of S^2. In Ref. 11 a uniqueness theorem for inverse diffraction problems with two-dimensional data is given.

ACKNOWLEDGMENT

The author thanks Professor K. Leichtweiss from the University of Stuttgart for a discussion and pointing out Refs. 5–9.

[1] A. G. Ramm, *Theory and Applications of Some New Classes of Integral Equations* (Springer, New York, 1980).
[2] A. G. Ramm, J. Math. Anal. Appl. (to appear, 1984).
[3] A. G. Ramm, Math. Meth. Appl. Sci. **3**, 336 (1981).
[4] S. Chaudhuri, W. Boerner, and B. Foo, "A High Frequency Inverse Scattering Model to Recover the Specular Point Curvature from Parametric Scattering Data," Department of Electrical Engineering, University of Illinois, Chicago, preprint, 1983.
[5] W. Blaschke, *Vorlesungen uber Differential Geometrie* (Springer–Verlag, Berlin, 1945), Vol. 1, p. 204.
[6] W. Firey, Mathematika **14**, 1 (1967).
[7] A. Hurwitz, Ann. Ecol. Norm. **19**, 357 (1902).
[8] E. Christoffel, J. Reine Angew. Math. **64**, 193 (1865).
[9] C. Hsuing, Math. Z. **64**, 41 (1956).
[10] A. G. Ramm, "Inverse diffraction problem," in *Proceedings of NATO-ARW-IMEI-1983* (D. Reidel, Dordrecht, 1984).
[11] A. G. Ramm, Phys. Rev. Lett. **52**, 13 (1984); Phys. Lett. (to appear, 1984).

Cylindrically symmetric solitary wave solutions to the Einstein equations

P. S. Letelier
Departamento de Física, Universidade de Brasília, 70.910, Brasília, D. F., Brazil

(Received 20 July 1983; accepted for publication 19 January 1984)

The integration of the Einstein equations for cylindrically symmetric solitary waves is reduced to a single quadrature when the "seed" solution is diagonal. Also in this case, explicit formulas that show the solitary wave character of the one- and two-soliton solutions are studied. A particular case of n-soliton solution is exhibited. Two theorems that show how to construct new solutions from known ones are presented.

PACS numbers: 04.20.Jb

I. INTRODUCTION

One of the most powerful methods used to generate new cylindrically symmetric solutions to the Einstein equations from a given solution (seed solution) is the inverse scattering method (ISM).[1-3] The new solutions generated by this method present similar properties to those of the solitary waves studied in fluid dynamics and in classical field theory.[4] In the actual application of the ISM, one finds two main difficulties. First, the ISM requires the explicit integration of an overdetermined system of linear partial differential equations. To find integrals that can be expressed in a closed form for the above-mentioned system of equations is not an easy task.[5] Second, the algebraic complexity of the solutions seldom allows the display of their main features. The purpose of this paper is to study the overcoming of the above-mentioned difficulties in the application of the Belinsky–Zakharov ISM for the special class of diagonal seed solutions.

The vacuum Einstein equations for the cylindrically symmetric metric

$$ds^2 = (e^{\Sigma}/\sqrt{t})(dt^2 - dr^2) - tf(d\theta + h\,dz)^2 - (t/f)dz^2, \tag{1.1}$$

where f, h, and Σ are functions of t and r only, split in two groups.[6] The first group can be written as

$$[t(f_{,t}/f - f^2hh_{,t})]_{,t} - [t(f_{,r}/f - f^2hh_{,r})]_{,r} = 0, \tag{1.2a}$$

$$(tf^2h_{,t})_{,t} - (tf^2h_{,r})_{,r} = 0, \tag{1.2b}$$

where $(\ )_{,t} \equiv \partial_t$ and $(\ )_{,r} \equiv \partial_r$, and the second as,

$$2\Sigma = \int t\{[f^{-2}(f_{,t}^2 + f_{,r}^2) + f^2(h_{,t}^2 + h_{,r}^2)]dt + 2(f^{-2}f_{,t}f_{,r} + f^2h_{,t}h_{,r})dr\}. \tag{1.3}$$

The existence of Σ is guaranteed by Eqs. (1.2). By a diagonal solution we mean a solution to (1.2) with $h = 0$.

In Sec. II, we present a summary of the ISM and the explicit integration of (1.3) when f and h are n-soliton solutions to (1.2). In Sec. III, we study the "Schrödinger equations" for the wave function ψ_0, and in the diagonal case we reduce the integration of these equations along the poles' trajectories to a single quadrature. Also in this section we give examples of particular functions ψ_0, and we present a superposition "principle" that allows the explicit computation of new functions ψ_0 from known ones.

In Secs. IV and V, we study the one- and two-soliton solutions, respectively. In particular we give compact formulas that reveal the soliton character of each solution in an explicit way. In Sec. VI, we show how to construct n-soliton solutions. Finally in Sec. VII, we discuss some of the results.

II. THE INVERSE SCATTERING METHOD

The starting point of the ISM used to solve Eqs. (1.2) is the fact that those equations can be written as the integrability condition for an overdetermined system of differential equations.[1] Equations (1.2) can be written in a completely equivalent form as the matrix equation[7]

$$(t\gamma_{,t}\gamma^{-1})_{,t} - (t\gamma_{,r}\gamma^{-1})_{,r} = 0, \tag{2.1}$$

with

$$\gamma = t\begin{vmatrix} f & fh \\ fh & h^2f + f^{-1} \end{vmatrix}. \tag{2.2}$$

Two important properties of the matrix γ are

$$\det\gamma = t^2, \quad \gamma = \gamma^T. \tag{2.3}$$

The integrability condition for the system of equations,

$$D_t\psi = \frac{tU + \lambda V}{t^2 - \lambda^2}\psi, \tag{2.4a}$$

$$D_r\psi = \frac{tV + \lambda U}{t^2 - \lambda^2}\psi, \tag{2.4b}$$

where

$$D_t = \partial_t + \frac{2\lambda t}{t^2 - \lambda^2}\partial_\lambda, \tag{2.5a}$$

$$D_r = \partial_r + \frac{2\lambda^2}{t^2 - \lambda^2}\partial_\lambda, \tag{2.5b}$$

$$U = t\gamma_{,t}\gamma^{-1}, \quad V = t\gamma_{,r}\gamma^{-1}, \tag{2.6}$$

is just the same Eq. (2.1). ψ is a 2×2 complex matrix function of t, r, and the spectral parameter λ. Putting $\lambda = 0$ in (2.4), we have $\psi(\lambda = 0) = \gamma$.

Solutions with pure soliton character are associated with solutions of Eqs. (2.4) of the form

$$\psi = \chi\psi_0, \tag{2.7}$$

$$\chi = I + \sum_{k=1}^{n}\frac{R_k}{\lambda - \mu_k}, \tag{2.8}$$

where ψ_0 is a solution to Eqs. (2.4) for a given γ, say γ_0. R_k are complex matrix functions of t and r only, and μ_k are scalar complex functions of t and r only. The pure soliton character of the solution is associated with the particular form of χ given by (2.8), i.e., with the existence of simple poles in the matrix χ. The number of poles will tell us the number of

 0022-2488/84/092675-07$02.50

solitons appearing in the solution. Note that by letting $\lambda = 0$ in (2.7), we get

$$\gamma = (\chi|_{\lambda=0})\gamma_0. \tag{2.9}$$

A condition that guarantees the fact that $\gamma = \gamma^T$ is

$$\gamma = \chi(t^2/\lambda)\gamma_0\chi^T(\lambda). \tag{2.10}$$

From (2.1)–(2.10) we find[1]

$$\gamma_{ab} = (\gamma_0)_{ab} - \sum_{k,l=1}^{n} \frac{N_a^{(l)}(\Gamma^{-1})_{lk}N_b^{(k)}}{\mu_k\mu_l}, \tag{2.11}$$

$$\Gamma_{kl} = \frac{m_a^{(k)}(\gamma_0)_{ab}m_k^{(l)}}{\mu_k\mu_l - t^2}, \tag{2.12}$$

$$N_a^{(k)} = m_b^{(k)}(\gamma_0)_{ba}, \tag{2.13}$$

$$m_a^{(k)} = m_{0b}^{(k)}M_{ba}^{(k)}, \tag{2.14}$$

$$M^{(k)} = \psi_0^{-1}(\mu_k), \tag{2.15}$$

$$\mu_k = \alpha_k - r \pm [(\alpha_k - r)^2 - t^2]^{1/2}. \tag{2.16}$$

The sum convention on the indices a and b is assumed. a and b take the values 1 and 2. $m_{0b}^{(k)}$ and α_k are sets of arbitrary complex constants. Note that the solution (2.11) is completely determined by γ_0, ψ_0 and these sets of constants. Regardless of the fact that the matrix whose elements are (2.11) is symmetric, in general, it is not possible to cast it in the form (2.2), since the matrix (2.2) has determinant equal to t^2. To remedy this problem, we can define a new matrix

$$\gamma^{Ph} = t\gamma/(\det\gamma)^{1/2} \tag{2.17}$$

that satisfies the two conditions (2.3). Taking the trace of (2.1), one can prove that the new γ^{Ph} is also a solution to (2.1) whenever γ is a solution. The determinant of (2.11) can be explicitly computed:

$$\det\gamma_n = t^{2n}\prod_{k=1}\mu_k^{-2}\det\gamma_0. \tag{2.18}$$

As in the elliptic case, one can compute the integral (1.3). For the solution (2.11) we find

$$\Sigma_n^{Ph} = \Sigma_0 + \ln\left[t^{-n^2/2}\left(\prod_{k=1}^{n}\mu_k^{n+1}\right)\times\prod_{\substack{k,l=1\\k>l}}^{n}(\mu_l-\mu_k)^{-2}\det\Gamma\right] + \ln C_n, \tag{2.19}$$

where C_n is an arbitrary constant and the factor

$$\prod_{\substack{k,l=1\\k>l}}^{n}(\mu_l-\mu_k)^{-2}$$

should be set equal to 1 for $n = 1$. The method that we use to compute (2.19) is essentially the same method employed in the elliptic case.[8]

III. THE FUNCTION ψ_0

The function ψ_0 obeys the differential equations (2.4) with γ replaced by γ_0, i.e.,

$$\left(\partial_t + \frac{2\lambda t}{t^2-\lambda^2}\partial_\lambda\right)\psi_0 = \frac{tU_0+\lambda V_0}{t^2-\lambda^2}\psi_0, \tag{3.1a}$$

$$\left(\partial_r + \frac{2\lambda^2}{t^2-\lambda^2}\partial_\lambda\right)\psi_0 = \frac{tV_0+\lambda U_0}{t^2-\lambda^2}\psi_0, \tag{3.1b}$$

where $U_0 \equiv t(\gamma_0)_{,t}\gamma_0^{-1}$ and $V_0 = t(\gamma_0)_{,r}\gamma_0^{-1}$. Furthermore, ψ_0 must satisfy the initial condition

$$\psi_0|_{\lambda=0} = \gamma_0. \tag{3.2}$$

Equations (3.1) can be written as

$$(t\partial_r - \lambda\partial_t)\psi_0 = V_0\psi_0, \tag{3.3a}$$

$$(t\partial_t - \lambda\partial_r + 2\lambda\partial_\lambda)\psi_0 = U_0\psi_0. \tag{3.3b}$$

In this section we study the system of equations (3.3) with the boundary condition (3.2) when γ_0 is a diagonal matrix, i.e., $(\gamma_0)_{12} = (\gamma_0)_{21} = 0$. If γ_0 is diagonal, one may assume that ψ_0 is also a diagonal matrix with these assumptions; (3.3) and (3.2) yield

$$(t\partial_r - \lambda\partial_t)\ln\det\psi_0 = 0, \tag{3.4a}$$

$$(t\partial_r - \lambda\partial_r + 2\lambda\partial_\lambda)\ln\det\psi_0 = 2, \tag{3.4b}$$

$$\ln\det\psi_0|_{\lambda=0} = 2\ln t. \tag{3.5}$$

In finding Eqs. (3.4) we have made use of the identities

$$\mathrm{Tr}\,U_0 = 2, \quad \mathrm{Tr}\,V_0 = 0. \tag{3.6}$$

A solution to Eqs. (3.4) with the condition (3.5) is

$$\det\psi_0 = t^2 + \lambda^2 + 2r\lambda. \tag{3.7}$$

A more general solution to (3.4) can be obtained adding $c\lambda$ to the rhs of (3.7), where c is a constant. We have omitted such a term because in the final results it will only introduce a redefinition of arbitrary constants. Since, for the diagonal case $\det\psi_0 = (\psi_0)_{11}(\psi_0)_{22}$, we have

$$(\psi_0)_{22} = [(t^2 + \lambda^2 + 2r\lambda)/(\psi_0)_{11}]. \tag{3.8}$$

It is convenient to introduce the notation

$$e^{\phi} \equiv (\gamma_0)_{11}/t, \tag{3.9}$$

$$\Lambda \equiv (t^2 + 2\lambda r + \lambda^2)^{-1/2}(\psi_0)_{11}. \tag{3.10}$$

Thus,

$$(\psi_0)_{11} = (t^2 + 2\lambda r + \lambda^2)^{1/2}\Lambda, \tag{3.11a}$$

$$(\psi_0)_{22} = (t^2 + 2\lambda r + \lambda^2)^{1/2}\Lambda^{-1}. \tag{3.11b}$$

For the diagonal case, Eqs. (1.2) and (3.9) tell us that

$$\phi_{,tt} + \phi_{,t}/t - \phi_{,rr} = 0. \tag{3.12}$$

The solution of the Einstein equations for the diagonal case are known as Einstein–Rosen waves.[9] From (3.2)–(3.4), (3.9) and (3.10) we get

$$(t\partial_t - \lambda\partial_r + 2\lambda\partial_\lambda)\ln\Lambda = t\phi_{,t}, \tag{3.13a}$$

$$(t\partial_r - \lambda\partial_t)\ln\Lambda = t\phi_{,r}, \tag{3.13b}$$

$$\ln\Lambda|_{\lambda=0} = \phi. \tag{3.13c}$$

The integrability condition of the system of equations (3.13) is just the cylindrical wave equation (3.12).

The ISM requires the explicit integration of Eqs. (3.13). The function Λ is known for the following particular solutions to (3.12):

$$\phi = \ln t, \tag{3.14a}$$

$$\phi = r, \tag{3.15a}$$

$$\phi = \tfrac{1}{2}t^2 + r^2, \tag{3.16a}$$

$$\phi = \ln[-r \pm (r^2 - t^2)^{1/2}]. \tag{3.17a}$$

$$\phi = \ln[\tfrac{1}{2}\tan(k\xi)], \tag{3.18a}$$

where

$$t = (a^2/2k^2)\sin(2k\tau)\sin(2k\xi),$$

$$r = -(a^2/2k^2)\cos(2k\tau)\cos(2k\xi),$$

a and k are constants. The corresponding functions Λ are

$$\Lambda = (t^2 + 2\lambda r + \lambda^2)^{1/2}, \tag{3.14b}$$

$$\Lambda = \exp(r + \tfrac{1}{2}\lambda), \tag{3.15b}$$

$$\Lambda = \exp[\tfrac{1}{2}t^2 + (r + \tfrac{1}{2}\lambda)^2], \tag{3.16b}$$

$$\Lambda = -r \pm [r^2 - t^2]^{1/2} - \lambda, \tag{3.17b}$$

$$\Lambda = \left\{\frac{a^4 \sin^2(2k\tau)\sin^2(k\xi) - a^2k^2\lambda\cos(2k\xi)}{k^6(t^2 + 2\lambda r + \lambda^2)} - \frac{1}{k^2}\right\}^{1/2}. \tag{3.18b}$$

The solutions (3.14), (3.17), and (3.18) are implicit in Refs. 1 and 10. Solutions (3.15) and (3.16) are the hyperbolic version of two solutions implicit in Ref. 5. The solution (3.18) is particularly interesting since the case $k = 1$ represents a closed Friedman–Robertson–Walker model and the case $k = i$ an open one. Since Eq. (3.12) as well as Eqs. (3.13) are invariant under a translation in the variable r we can always introduce a constant in the solutions (3.14a), (3.14b) etc., letting $r \to r + c$.

From the particular form of Eqs. (3.13) it is easy to show the following theorems:

Let $\phi_{(1)}$ and $\phi_{(2)}$ be solutions to (3.12) and $\Lambda_{(1)}$ and $\Lambda_{(2)}$ their corresponding solutions to (3.13). Then, the Λ associated to

$$\phi = a\phi_{(1)} + b\phi_{(2)}, \tag{3.19}$$

where a and b are arbitrary constants is

$$\Lambda = \Lambda_{(1)}^{a}\Lambda_{(2)}^{b}. \tag{3.20}$$

An important corollary is obtained by putting $b = 0$ in (3.19) and (3.20), i.e., the Λ associated to $\phi = a\phi_{(1)}$ is $\Lambda_{(1)}^{a}$.

Let Λ be the solution to (3.13) corresponding to $a\phi$ solution to (3.12). The Λ's associated to the new solutions ϕ_R and ϕ_I defined by

$$\phi_R = \tfrac{1}{2}[\phi(t,r + ir_0) + \phi(t,r - ir_0)], \tag{3.21}$$

$$\phi_I = (1/2i)[\phi(t,r + ir_0) - \phi(t,r - ir_0)], \tag{3.22}$$

where r_0 is arbitrary constant, are

$$\Lambda_R = [\Lambda(t,r + ir_0,\lambda)\Lambda(t,r - ir_0,\lambda)]^{1/2}, \tag{3.23}$$

$$\Lambda_I = \left[\frac{\Lambda(t,r + ir_0,\lambda)}{\Lambda(t,t - ir_0,\lambda)}\right]^{1/2i}. \tag{3.24}$$

For a real λ,

$$\Lambda_R = |\Lambda(t,r + ir_0,\lambda)|, \tag{3.25}$$

$$\Lambda_I = \exp\arctan\frac{\operatorname{Im}\Lambda(t,r + ir_0,\lambda)}{\operatorname{Re}\Lambda(t,r + ir_0,\lambda)}. \tag{3.26}$$

The "complex translational method" presented here can be used to obtain exact perturbations to the Friedman–Robertson–Walker models and their corresponding solitonic solutions. Work along these lines will be soon reported.

In the final formulas (2.11)–(2.15) the matrix ψ_0 appears only in the form $\psi_0|_{\lambda=\mu_k}$. Thus, to construct the solution (2.11) we *only* need,

$$\Lambda_k \equiv \Lambda|_{\lambda=\mu_k}. \tag{3.27}$$

The equations that give the poles' trajectories (2.16) are[1]

$$\mu_{k,t} = \frac{2\mu_k t}{t^2 - \mu_k^2}, \tag{3.28a}$$

$$\mu_{k,r} = \frac{2\mu_k^2}{t^2 - \mu_K^2}. \tag{3.28b}$$

From (3.28) and the equations that are obtained by doing $\lambda = \mu_k$ in (3.13a) and (3.13b) we get

$$\partial_t \ln\Lambda_k = (t/2\mu_k)(\mu_{k,t}\phi_{,t} + \mu_{k,r}\phi_{,r}), \tag{3.29a}$$

$$\partial_r \ln\Lambda_k = (t/2\mu_k)(\mu_{k,t}\phi_{,r} + \mu_{k,r}\phi_{,t}). \tag{3.29b}$$

Thus,

$$\ln\Lambda_k = \int \frac{t}{2\mu_k}[(\mu_{k,t}\phi_{,t} + \mu_{k,r}\phi_{,r})dt + (\mu_{k,r}\phi_{,t} + \mu_{k,t}\phi_{,r})dr]. \tag{3.30}$$

Using (3.28), one can prove that (3.30) is consistent with (3.13c). The existence of (3.30) is guaranteed by Eq. (3.12) and the fact that $\ln\mu_k$ is also a solution of (3.12). In other words, the overdetermined system of Eqs. (3.1) for diagonal matrices U_0, V_0, and ψ_0 is completely determined along the poles' trajectories. Thus, in principle, *to any solution* of (3.12) we can associate a n-soliton solution (2.11). Finally, we want to point out that the application of the closely related method for Bäcklund transformations (BT) in the case of diagonal seed solutions also reduces to the finding of a single function.[11,12]

IV. ONE-SOLITON SOLUTIONS

One-soliton solutions are defined[1] as those solutions obtained using the ISM with a "scattering matrix" χ with one simple real pole μ_1. In the diagonal case, these solutions can be written in a simple form. From (2.11)–(2.19), (3.11), and (3.27) we get

$$\gamma_{11}^{Ph} = \frac{(m_1)^2 te^{\phi}\mu^{-1}\Lambda_1^{-2} + (m_2)^2 t^{-1}e^{-\phi}\mu\Lambda_1^2}{(m_1)^2 e^{\phi}\Lambda_1^{-2} + (m_2)^2 e^{-\phi}\Lambda_1^2}(\gamma_0)_{11}, \tag{4.1a}$$

$$\gamma_{12}^{Ph} = -\frac{m_1 m_2 t(\mu/t - t/\mu)}{(m_1)^2 e^{\phi}\Lambda_1^{-2} + (m_2)^2 e^{-\phi}\Lambda_1}, \tag{4.1b}$$

$$\gamma_{22}^{Ph} = \frac{(m_1)^2\mu e^{\phi}t^{-1}\Lambda_1^{-2} + (m_2)^2\mu^{-1}e^{-\phi}t\Lambda_1^2}{(m_1)^2 e^{\phi}\Lambda_1^{-2} + (m_2)^2 e^{-\phi}\Lambda_1}(\gamma_0)_{22}, \tag{4.1c}$$

$$\Sigma_1^{Ph} = \Sigma_0 + \ln\left\{\frac{t^{1/2}[(m_1)^2 e^{\phi}\Lambda_1^{-2} + (m_2)^2 e^{-\phi}\Lambda_1^2]}{[(\alpha_1 - r)^2 - t^2]^{1/2}}\right\} + \ln C_1, \tag{4.2}$$

where we have set $m_a \equiv m_{oa}^{(1)}$ and $\mu \equiv \mu_1$. The constants m_a are real in this case. Equations (4.1)–(4.2) can be cast in a more appealing form by defining the new variables

$$x = \phi - 2\ln\Lambda_1, \tag{4.3a}$$

$$p = \phi - \ln(\mu\Lambda_1^2/t), \tag{4.3b}$$

$$q = \phi - \ln(t\Lambda_1^2/\mu), \tag{4.3c}$$

$$2y = q - p, \tag{4.3d}$$

and the new constants

$$\eta = m_1 m_2/|m_1 m_2|, \tag{4.4a}$$

$$\tanh\delta = \frac{(m_1)^2 - (m_2)^2}{(m_1)^2 + (m_2)^2}. \tag{4.4b}$$

From (4.1)–(4.6) we get

$$\gamma_{11}^{Ph} = \frac{\cosh(p+\delta)}{\cosh(x+\delta)}(\gamma_0)_{11}, \tag{4.5a}$$

$$\gamma_{12}^{Ph} = -\frac{\eta t \sinh y}{\cosh(x+\delta)}, \tag{4.5b}$$

$$\gamma_{22}^{Ph} = \frac{\cosh(q+\delta)}{\cosh(x+\delta)}(\gamma_0)_{22}, \tag{4.5c}$$

$$\Sigma_1^{Ph} = \Sigma_0 + \ln[t^{-1/2}\cosh(x+\delta)/\sinh y] + \ln C_1, \tag{4.6}$$

where we have denoted the "renormalized" integration constant by the same symbol used in (4.2), a practice that we shall follow in this paper. Note that the structure of the solution does not depend on the seed solution's γ_0 particular form, as long as $(\gamma_0)_{12} = (\gamma_0)_{21} = 0$. The soliton character of the solution is given by the appearance of the function $\cosh(x+\delta)$ in the denominator of the rhs of (4.5). Particular cases of one-soliton solutions are studied in some detail in Refs. 1 and 10.

V. TWO-SOLITON SOLUTIONS

Two-soliton solutions are defined[1] as those solutions obtained using the ISM with a "scattering matrix" χ with two simple poles. In this case the poles are either real or complex conjugate. From (2.11)–(2.19), (3.11), (3.27), and $(\gamma_0)_{12} = (\gamma_0)_{21} = 0$, we find after some algebra

$$\gamma_{11}^{Ph} = \frac{[t(\mu_2-\mu_1)P_1]^2 + [(\mu_1\mu_2 - t^2)P_2]^2}{[t(\mu_2-\mu_1)S_1]^2 + [(\mu_1\mu_2 - t^2)S_2]^2}(\gamma_0)_{11}, \tag{5.1a}$$

$$\gamma_{12}^{Ph} = \frac{t(\mu_2-\mu_1)(\mu_1\mu_2 - t^2)}{\mu_1\mu_2} \times \frac{\mu_2(\mu_1^2 - t^2)m_{01}^{(1)}m_{02}^{(1)}T_1 - \mu_1(\mu_2^2 - t^2)m_{01}^{(2)}m_{02}^{(2)}T_2}{[t(\mu_2-\mu_1)S_1]^2 + [(\mu_1\mu_2 - t^2)S_2]^2}, \tag{5.1b}$$

$$\gamma_{22}^{Ph} = \frac{[t(\mu_2-\mu_1)Q_1]^2 + [(\mu_1\mu_2 - t^2)Q_2]^2}{[t(\mu_2-\mu_1)S_1]^2 + [(\mu_1\mu_2 - t^2)S_2]^2}(\gamma_0)_{22}, \tag{5.1c}$$

$$\Sigma_2^{Ph} = \Sigma_0 - \ln[(\mu_1^2 - t^2)(\mu_2^2 - t^2) \times (\mu_1\mu_2 - t^2)^2(1/\mu_1 - 1/\mu_2)^2] + \Sigma_2' + \ln C_2, \tag{5.2}$$

where

$$S_1 \equiv m_{01}^{(1)}m_{01}^{(2)}\frac{e^{\phi}}{\Lambda_1\Lambda_2} + m_{02}^{(1)}m_{02}^{(2)}\frac{\Lambda_1\Lambda_2}{e^{\phi}}, \tag{5.3}$$

$$S_2 \equiv m_{01}^{(1)}m_{02}^{(2)}\frac{\Lambda_2}{\Lambda_1} - m_{02}^{(1)}m_{01}^{(2)}\frac{\Lambda_1}{\Lambda_2}, \tag{5.4}$$

$$P_1 \equiv m_{01}^{(1)}m_{01}^{(2)}\frac{te^{\phi}}{(\mu_1\mu_2)^{1/2}\Lambda_1\Lambda_2} + m_{02}^{(1)}m_{02}^{(2)}\frac{(\mu_1\mu_2)^{1/2}\Lambda_1\Lambda_2}{te^{\phi}}, \tag{5.5}$$

$$P_2 \equiv m_{01}^{(1)}m_{02}^{(2)}\frac{\Lambda_2\sqrt{\mu 2}}{\Lambda_1\sqrt{\mu_1}} - m_{02}^{(1)}m_{01}^{(2)}\frac{\Lambda_1\sqrt{\mu_1}}{\Lambda_2\sqrt{\mu_2}}, \tag{5.6}$$

$$T_1 \equiv (m_{01}^{(2)}e^{\phi/2}/\Lambda_2)^2 + (m_{02}^{(2)}\Lambda_2/e^{\phi/2})^2, \tag{5.7}$$

$$T_2 \equiv (m_{01}^{(1)}e^{\phi/2}/\Lambda_1)^2 + (m_{02}^{(1)}\Lambda_1/e^{\phi/2})^2, \tag{5.8}$$

$$Q_1 \equiv m_{02}^{(1)}m_{02}^{(2)}\frac{t\Lambda_1\Lambda_2}{(\mu_1\mu_2)^{1/2}e^{\phi}} + m_{01}^{(1)}m_{01}^{(2)}\frac{(\mu_1\mu_2)^{1/2}e^{\phi}}{t\Lambda_1\Lambda_2}, \tag{5.9}$$

$$Q_2 \equiv m_{01}^{(1)}m_{02}^{(2)}\frac{\Lambda_2\sqrt{\mu_1}}{\Lambda_1\sqrt{\mu_2}} - m_{02}^{(1)}m_{01}^{(2)}\frac{\Lambda_1\sqrt{\mu_2}}{\Lambda_2\sqrt{\mu_1}}, \tag{5.10}$$

$$\Sigma_2' \equiv \ln\{[t(\mu_2-\mu_1)S_1]^2 + [(\mu_2\mu_1 - t^2)S_2]^2\}. \tag{5.11}$$

The fact that the poles μ_1 and μ_2 are either real or complex conjugate can be used to simplify the previous formulas. For real μ_1 and μ_2 and real constants $m_{oa}^{(k)}$ such that

$$m_{01}^{(1)}m_{02}^{(1)}m_{01}^{(2)}m_{02}^{(2)} > 0, \tag{5.12}$$

we find

$$\gamma_{11}^{Ph} = \frac{[t(\mu_2-\mu_1)\cosh(p_1+\delta_1)]^2 + [(\mu_1\mu_2 - t^2)\sinh(p_2+\delta_2)]^2}{[t(\mu_2-\mu_1)\cosh(x_1+\delta_1)]^2 + [(t^2 - \mu_1\mu_2)\sinh(x_2+\delta_2)]^2}(\gamma_0)_{11}, \tag{5.13a}$$

$$\gamma_{12}^{Ph} = \frac{(\mu_2-\mu_1)(\mu_1\mu_2 - t^2)}{2\mu_1\mu_2}\frac{\eta_1\mu_1(\mu_2^2 - t^2)\cosh(y_1+\epsilon_1) + \eta_2\mu_2(\mu_1^2 - t^2)\cosh(y_2+\epsilon_2)}{[t(\mu_2-\mu_1)\cosh(x_1+\delta_1)]^2 + [(\mu_1\mu_2 - t^2)\sinh(x_2+\delta_2)]^2}, \tag{5.13b}$$

$$\gamma_{22}^{Ph} = \frac{[t(\mu_2-\mu_1)\cosh(q_1+\delta_1)]^2 + [(\mu_1\mu_2 - t^2)\sinh(q_2+\delta_2)]^2}{[t(\mu_2-\mu_1)\cosh(x_1+\delta_1)]^2 + [(t^2 - \mu_1\mu_2)\sinh(x_2+\delta_2)]^2}(\gamma_0)_{22}, \tag{5.13c}$$

$$\Sigma_2' = \ln\{[t(\mu_2-\mu_1)\cosh(x_1+\delta_1)]^2 + [(\mu_1\mu_2 - t^2)\sinh(x_2+\delta_2)]^2\}, \tag{5.14}$$

where

$$x_1 \equiv \phi - \ln(\Lambda_1\Lambda_2), \tag{5.15}$$

$$x_2 \equiv \ln(\Lambda_2/\Lambda_1), \tag{5.16}$$

$$p_1 \equiv \phi - \ln[\Lambda_1\Lambda_2(\mu_1\mu_2)^{1/2}/t], \tag{5.17}$$

$$p_2 \equiv \ln[\Lambda_2\sqrt{\mu_2}/\Lambda_1\sqrt{\mu_1}], \tag{5.18}$$

$$q_1 \equiv 2x_1 - p_1, \quad q_2 \equiv 2x_2 - p_2, \tag{5.19}$$

$$y_1 \equiv x_1 + x_2, \quad y_2 \equiv x_1 - x_2, \tag{5.20}$$

and

$$\tanh \delta_1 \equiv \frac{m_{01}^{(1)} m_{01}^{(2)} - m_{02}^{(1)} m_{02}^{(2)}}{m_{01}^{(1)} m_{01}^{(2)} + m_{02}^{(1)} m_{02}^{(2)}}, \tag{5.21}$$

$$\tanh \delta_2 \equiv \frac{m_{01}^{(1)} m_{02}^{(2)} - m_{01}^{(2)} m_{02}^{(1)}}{m_{01}^{(1)} m_{02}^{(2)} + m_{01}^{(2)} m_{02}^{(1)}}, \tag{5.22}$$

$$\tanh \epsilon_1 \equiv \frac{(m_{01}^{(1)})^2 - (m_{02}^{(1)})^2}{(m_{01}^{(1)})^2 + (m_{02}^{(1)})^2}, \tag{5.23}$$

$$\tanh \epsilon_2 \equiv \frac{(m_{01}^{(2)})^2 - (m_{02}^{(2)})^2}{(m_{01}^{(2)})^2 + (m_{02}^{(2)})^2}, \tag{5.24}$$

$$\eta_1 \equiv - \frac{m_{01}^{(1)} m_{02}^{(1)}}{|m_{01}^{(1)} m_{02}^{(1)}|}, \quad \eta_2 = \frac{m_{01}^{(2)} m_{02}^{(2)}}{|m_{01}^{(2)} m_{02}^{(2)}|}. \tag{5.25}$$

In the complementary case of Eq. (5.12), i.e.,

$$m_{01}^{(1)} m_{02}^{(1)} m_{01}^{(2)} m_{02}^{(2)} < 0, \tag{5.26}$$

we find that the relations (5.13a), (5.13c), (5.14), (5.21), and (5.22) keep the same form, but changing the hyperbolic functions by their respective cofunctions. The relations (5.15)–(5.20) and (5.23)–(5.25) remain the same. The component γ_{12}^{Ph}, now reads

$$\gamma_{12}^{Ph} = - \frac{t(\mu_2 - \mu_1)^2(\mu_1\mu_2 - t^2)}{2\mu_1\mu_2} \frac{\eta_1\mu_1(\mu_2^2 - t^2)\cosh(y_1 + \epsilon_1) + \eta_2\mu_2(\mu_1^2 - t^2)\cosh(y_2 + \epsilon_2)}{[t(\mu_2 - \mu_1)\sinh(x_1 + \delta_1)]^2 + [(\mu_1\mu_2 - t^2)\cosh(x_2 + \delta_2)]^2}. \tag{5.27}$$

In the case that μ_1 and μ_2 are complex conjugates we set

$$\mu \equiv \mu_1, \quad \mu^* = \mu_2, \tag{5.28}$$

$$\Lambda(\mu) \equiv \Lambda_1, \quad \Lambda(\mu^*) = \Lambda_2. \tag{5.29}$$

To end up with a real metric we need

$$m_1 \equiv m_{01}^{(1)}, \quad m_1^* = m_{01}^{(2)}, \tag{5.30}$$

$$m_2 \equiv m_{02}^{(1)}, \quad m_2^* = m_{02}^{(2)}. \tag{5.31}$$

From (5.1)–(5.11) we find

$$\gamma_{11}^{Ph} = \frac{[2t|\mu|\sin y_2 \cosh(p + \delta_1)]^2 + [(|\mu|^2 - t^2)\sin(\omega_1 + \delta_2)]^2}{[2t|\mu|\sin y_2 \cosh(x + \delta_1)]^2 + [(|\mu| - t^2)\sin(\omega_3 + \delta_2)]^2} (\gamma_0)_{11}, \tag{5.32a}$$

$$\gamma_{12}^{Ph} = -2t(|\mu|^2 - t^2)\sin y_2 \frac{(|\mu|^2 - t^2)\cos y_2 \sin y_1 \sinh(x + \delta_1) + (|\mu|^2 + t^2)\sin y_2 \cos y_1 \cosh(x + \delta_1)}{[2t|\mu|\sin y_2 \cosh(x + \delta_1)]^2 + [(|\mu|^2 - t^2)\sin(\omega_3 + \delta_2)]^2}, \tag{5.32b}$$

$$\gamma_{22}^{Ph} = \frac{[2t|\mu|\sin y_2 \cosh(q + \delta_1)]^2 + [(|\mu|^2 - t^2)\sin(\omega_2 + \delta_2)]^2}{[2t|\mu|\sin y_2 \cosh(x + \delta_1)]^2 + [(|\mu|^2 - t^2)\sin(\omega_3 + \delta_2)]^2} (\gamma_0)_{22}, \tag{5.32c}$$

$$\Sigma_2^{Ph} = \Sigma_0 - 2\ln[|\mu^2 - t^2|(|\mu|^2 - t^2)\sin(y_2)/|\mu|] + \ln\{[2t|\mu|\sin y_2 \cosh(x + \delta_1)]^2 + [(|\mu|^2 - t^2)\sin(\omega_3 + \delta_2)]^2\} + \ln C_2, \tag{5.33}$$

where

$$x = \phi - 2\ln|\Lambda(\mu)|, \tag{5.34}$$

$$p = \phi - \ln(|\mu|\,|\Lambda(\mu)|^2/t), \tag{5.35}$$

$$q = \phi - \ln(t|\Lambda(\mu)|^2/|\mu|), \tag{5.36}$$

$$\omega_1 = 2\arg\Lambda + \arg\mu, \tag{5.37}$$

$$\omega_2 = 2\arg\Lambda - \arg\mu, \tag{5.38}$$

$$\omega_3 = \tfrac{1}{2}(\omega_1 + \omega_2), \tag{5.39}$$

$$y_1 = \tfrac{1}{2}(\omega_1 + \omega_2) + \delta_2, \tag{5.40}$$

$$y_2 = \tfrac{1}{2}(\omega_1 - \omega_2), \tag{5.41}$$

and

$$\tanh \delta_1 = \frac{|m_1|^2 - |m_2|^2}{|m_1|^2 + |m_2|^2}, \tag{5.42}$$

$$\delta_2 = \arg m_1^* m_2. \tag{5.43}$$

As in the one-soliton case the structure of the two-soliton solution does not depend on the seed solution's γ_0, particular form, as long as $(\gamma_0)_{12} = (\gamma_0)_{21} = 0$. In the case of real poles, the two-soliton character of the solution can be perceived by noticing the special form of the denominator that appears in each component of γ^{Ph}. But, for complex conjugated poles, the two-soliton character of γ^{Ph} given by (5.32) is not clear. Even though the localized character of the solu-

tion is insured by the presence of the function $\cosh^2(x+\delta_1)$ in the denominator of each component of γ^{Ph}, the appearance of two "bumps" is not as clear as in the case of real poles. The case of complex conjugate poles looks as if the two bumps were "mixed" forming a "bound state." Of course the actual form of the functions γ_{ab}^{Ph} will depend heavily on the special form of $(\gamma_0)_{ab}$, or to be more precise on the special form of the functions ϕ and on its functionally related function A. The study of a special case of two-soliton solutions can be found in Ref. 13.

Letting $\mu_1=\mu_2$ in (5.1)–(5.11) we find that $\gamma^{Ph}=\gamma_0$ and $\Sigma^{Ph}=\Sigma_0$, i.e., the two poles cancel out and we end up with the original seed solution γ_0. This result can be easily proved for a nondiagonal γ_0 using the general formalism of Ref. 1. In the method of BT we have the exact opposite behavior of the poles, i.e., the coincidence of poles can be used to generate new solutions.[14]

VI. n-SOLITON SOLUTIONS

In this section we study how to construct a particular n-soliton solution to the Einstein equation using as a seed solution a diagonal γ_0. First, let us consider the particular one-soliton solution obtained by putting either $m_1=0$ or $m_2=0$ in (4.1),

$$\gamma_{11}^{Ph}=(\mu_1/t)^{\epsilon_1}(\gamma_0)_{11}, \tag{6.1}$$

$$\gamma_{12}^{Ph}=0,\quad \gamma_{22}^{Ph}=t^2/\gamma_{11}^{Ph}, \tag{6.2}$$

where $\epsilon_1=\pm 1$. Similarly letting either $m_{01}^{(1)}=m_{02}^{(2)}=0$, $m_{01}^{(1)}=m_{01}^{(2)}=0$, $m_{02}^{(1)}=m_{02}^{(2)}=0$ or $m_{02}^{(1)}=m_{01}^{(2)}=0$ in (5.1) we get

$$\gamma_{11}^{Ph}=(\mu_1/t)^{\epsilon_1}(\mu_2/t)^{\epsilon_2}(\gamma_0)_{11}; \tag{6.3}$$

γ_{12}^{Ph} and γ_{22}^{Ph} are given by (6.2) and ϵ_1 and ϵ_2 are constants that can take the values ± 1. Moreover, the two-soliton solution (6.3) can also be obtained from the one-soliton solution by considering the seed solution γ_0 as a one-soliton of the same form (6.1), i.e., taking

$$(\gamma_0)_{11}=(\mu_2/t)^{\epsilon_2}(\gamma_0')_{11} \tag{6.4}$$

in (6.1), where γ_0' is a new seed solution. This procedure can be repeated n times to give

$$\gamma_{11}^{Ph}=\left[\prod_{l=1}^{n}\left(\frac{\mu_l}{t}\right)^{\epsilon_l}\right](\gamma_0)_{11}; \tag{6.5}$$

γ_{12}^{Ph} and γ_{22}^{Ph} are obtained as in Eq. (6.1b).

The function Σ_n^{Ph} can be found either using Eq. (2.19) or by direct integration of Eq. (1.3). We shall use the second method in this case because it is simpler. From (6.5) and (3.9) we get

$$\phi_n=\phi-\ln t\sum_{l=1}^{n}\epsilon_l+\sum_{l=1}^{n}\epsilon_l\ln\mu_l. \tag{6.6}$$

From Eq. (1.3) and (3.30) after a direct, but rather long, computation we get[15,16]

$$\begin{aligned}\Sigma_n^{Ph}=\Sigma_0-\phi\sum_{l=1}^{n}\epsilon_l+2\sum_{l=1}^{n}\epsilon_l\ln A_l+\frac{1}{2}\ln t\left(\sum_{l=1}^{n}\epsilon_l\right)^2\\ -\sum_{l,k=1}^{n}\epsilon_l\epsilon_k\ln\mu_l+\sum_{l}^{n}\epsilon_l^2\ln\frac{\mu_l^2}{\mu_l^2-t^2}\\ +2\sum_{l>k=1}^{n}\epsilon_l\epsilon_k\ln(\mu_l-\mu_k)+\ln C_n.\end{aligned} \tag{6.7}$$

The function ϕ_n given by (6.6) with ϵ_l arbitrary constants is also a solution to (3.12). In computing (6.7) we did not assume any particular value of ϵ_l. Then (6.7) gives us the corresponding Σ_n^{Ph} for the function (6.6) with arbitrary constants ϵ_l. Although the solution (6.6) with arbitrary ϵ_l is a solution to the Einstein equations, we cannot say that this new solution is an n-soliton solution since the soliton character is given by the existence of simple poles in the "scattering" matrix χ and for simple poles we have $\epsilon_l=\pm 1$.

Let us consider a simple particular case of a seed solution,

$$\phi=a_1\ln t+a_2(r-r_1)+a_3\left[\tfrac{1}{2}t^2+(r-r_2)^2\right], \tag{6.8}$$

where a_1, a_2, a_3, r_1, and r_2 are arbitrary constants. The construction of the particular n-soliton associated to this particular seed solution requires the actual computation of Σ_0 and A_l only. From Eq. (1.3) we find

$$\begin{aligned}\Sigma_0=\tfrac{1}{2}a_1^2\ln t+\tfrac{1}{4}(a_2^2+2a_1a_3)t^2+a_1a_2r+a_1a_3(r-r_2)^2\\ +a_2a_3(r-r_2)t^2+\tfrac{1}{8}a_3^2t^4+a_3^2t^2(r-r_2)^2.\end{aligned} \tag{6.9}$$

To obtain A_l we use the superposition "principle" presented in Sec. III. First, we have that associated to each

$$\phi_1=\ln t, \tag{6.10a}$$

$$\phi_2=r-r_1, \tag{6.11a}$$

$$\phi_3=\tfrac{1}{2}t^2+(r-r_2)^2, \tag{6.12a}$$

we have the corresponding A functions

$$A_{(1)}=(t^2+2\lambda r+\lambda^2)^{1/2}, \tag{6.10b}$$

$$A_{(2)}=\exp(r-r_1+\tfrac{1}{2}\lambda), \tag{6.11b}$$

$$A_{(3)}=\exp\left[\tfrac{1}{2}t^2+(r-r_2+\tfrac{1}{2}\lambda)^2\right]. \tag{6.12b}$$

From the superposition theorem we have

$$A(\mu_l)=A_{(1)}^{a_1}(\mu_l)A_{(2)}^{a_2}(\mu_l)A_{(3)}^{a_3}(\mu_l). \tag{6.13}$$

From (6.10b)–(6.12b), (6.13), and (2.16) we get

$$\begin{aligned}A_l=(2\alpha_l\mu_l)^{a_1/2}\exp\{a_2(r-r_1+\tfrac{1}{2}\mu_l)\\ +a_3\left[\tfrac{1}{2}t^2+(r-r_2+\tfrac{1}{2}\mu_l)^2\right]\}.\end{aligned} \tag{6.14}$$

Also, this particular n-soliton solution can be used as a new seed solution to construct a new solution, i.e., we can use Σ^{Ph} given by (6.7) as Σ_0 and the function A_l associated to the full solution (6.6) can be found again using the superposition principle, and so on.

VII. DISCUSSION

The main results of this paper are that the finding of the soliton solutions associated with Einstein–Rosen waves is reduced to a single quadrature Eq. (3.30) (we consider the quadrature that defines Σ_0 as known) and that the one- and two-soliton solutions admit explicit "canonical" forms that exhibit the soliton character in a simple way.

The timelike and spacelike character of the coordinates t and r can be inverted by adding $i\pi$ to the integration constant that appears in Σ_n^{Ph}. Depending on the timelike or spacelike character of t, the soliton solutions can be interpreted as an exact finite perturbation of either a cosmological model or a cylindrical wave.

A similar discussion to the one presented here can be

done for the case of solitons associated with axially symmetric stationary solutions to the Einstein equations[5] and with solutions to self-dual SU(2) gauge fields on Euclidean space.[17] Work along these lines will be soon reported.

[1]V. A. Belinsky and V. E. Zakharov, Zh. Eksp. Teor. Fiz. **75**, 1955 (1978) [Sov. Phys. JETP **48**, 985 (1978)].

[2]I. Hauser and F. J. Ernst, J. Math. Phys. **21**, 1126 (1980).

[3]The relation between Belinsky–Zakharov and Hauser Ernst ISM can be found in C. M. Cosgrove, J. Math. Phys. **23**, 615 (1982).

[4]See for instance, V. G. Makhankov, Phys. Rep. **35**, 1 (1978).

[5]P. S. Letelier, Phys. Rev. D **26**, 3728 (1982).

[6]See for instance, P. S. Letelier, Phys. Rev. D **26**, 2623 (1982); and references therein.

[7]Equation (2.1) as well as the other equations presented in Sec. II, differ from the equivalent equations of Ref. 1 because they are written in "cylindrical" coordinates and not in characteristic coordinates. Also we have fixed the gauge freedom choosing $\det \gamma = t^2$. This can be done without losing generality as explained in the same Ref. 1.

[8]V. A Belinsky and V. E. Zakharov, Zh. Eksp. Teor. Fiz. **77**, 3 (1979) [Sov. Phys. JETP **50**, 1 (1979)]. Note that Eq. (3.6) of the quoted reference is misprinted. The exponent that appears in the last factor of Eq. (3.6) should read -2.

[9]See for instance, J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1966), p. 352. The coordinates used in this paper are related by a simple change of variables from the ones used in the quoted reference.

[10]V. A. Belinsky, Zh. Eksp. Teor. Fiz. **77**, 1239 (1979) [Sov. Phys. JETP **50**, 623 (1979)].

[11]C. Hoenselaers, W. Kinnersley, and B. C. Xanthopoulos, J. Math. Phys. **20**, 2530 (1979); B. K. Harrison, Phys. Rev. D **21**, 1965 (1980).

[12]The relation between different solution generating techniques can be found in C. M. Cosgrove, J. Math. Phys. **21**, 2417 (1980).

[13]V. A. Belinsky and D. Fargion, Nuovo Cimento B **59**, 143 (1980); V. A. Belinsky and V. E. Zakharov, in *Sources in Gravitational Radiation*, edited by L. Smarr (Cambridge UP, Cambridge, 1979), p. 161.

[14]K. Oohara and H. Sato, Prog. Theor. Phys. **65**, 1891 (1981) and references therein.

[15]A particular case of (6.6), (6.7) can be found in Ref. 8 and G. A. Alekseev and V. A. Belinsky, Zh. Eksp. Teor. Fiz. **78**, 1297 (1980) [Sov. Phys. JETP **51**, 655 (1980)]. See also G. A. Alekseev, Dok. Akad. Nauk SSSR **256**, 827 (1980) [Sov. Phys. Dokl. **26**, 153 (1981)]; P. S. Letelier, Rev. Bras. de Fis. (to appear).

[16]Multiple soliton solutions in the elliptic case obtained using BT can be found in the second citation of Ref. 11 and in G. Neugebauer, J. Phys. A: Math. Gen. **13**, 1737 (1980); D. Kramer and G. Neugebauer, Phys. Lett. A **75**, 259 (1980); W. Dietz and C. Hoenselaers, Proc. R. Soc. London Ser. A **382**, 221 (1982); Phys. Rev. Lett. **48**, 778 (1982); M. Yamazaki, Phys. Rev. Lett. **50**, 1027 (1983); and references therein.

[17]P. S. Letelier, J. Math. Phys. **23**, 1175 (1982).

From i° to the 3 + 1 description of spatial infinity

Abhay Ashtekar[a)]
Physique Théorique, Institut Henri Poincaré, 11, rue P. et M. Curie, 75 231 Paris, France,[b)] and Physics Department, Syracuse University, Syracuse, New York 13210

Anne Magnon
Département de Mathématiques, Université de Clermond-Fd, 63170 Aubière, France, and Physique Théorique, Institut Henri Poincaré, 11, rue P. et M. Curie, 75 231 Paris, France[b)]

(Received 9 November 1983; accepted for publication 30 March 1984)

By carrying out a 3 + 1 decomposition of the spi framework, the expressions of the conserved quantities, defined at i° in terms of the Weyl curvature, are recast in terms of the initial data of the physical space-time. In particular, the analysis brings out the supertranslation ambiguities in the usual 3 + 1 definitions of angular momentum and clarifies, within the 3 + 1 framework, the meaning of the stronger boundary condition needed to remove these ambiguities. The discussion is so arranged that only a minimal acquaintance with the spi framework is necessary to appreciate these issues.

PACS numbers: 04.20.Me, 04.20.Cv

I. INTRODUCTION

In the Arnowitt, Deser, Misner (ADM) formalism,[1] the structure of the gravitational field at spatial infinity is explored by working on three-dimensional spacelike submanifolds of space-time and using the Cauchy data induced on these surfaces by the space-time metric. In the so-called spi framework,[2,3] on the other hand, one conformally completes the four-dimensional space-time by attaching to it the point i° at spatial infinity and uses the curvature tensors of the rescaled metric as the basic variables. The primary motivation in this shift of emphasis was to unify the description of spatial infinity with that of null infinity. However, the spi framework also clarified the role of supertranslations at spatial infinity, introduced a method of reduction of the infinite-dimensional asymptotic symmetry group to the Poincaré group, and provided an expression of angular momentum which is free of supertranslation ambiguities. The status of these issues has continued to remain unclear in the three-dimensional frameworks. Furthermore, even in the case of the 4-momentum, an explicit proof showing the equality of the spi expression involving the Weyl curvature of the 4-metric and the ADM expression involving the Cauchy data on a 3-surface is not yet available. The purpose of this paper is to fill these gaps.

Section II is devoted to preliminaries. The basic definitions are recalled and the supertranslation ambiguities in the ADM definition of angular momentum are pointed out. Section III deals with energy. The spi energy expression is recast in terms of Cauchy data and shown to reproduce not only the ADM expression, but six other expressions, some of which have appeared in the literature in the context of the positive energy theorems.[4] Section IV shows the equality of the spi and the ADM expressions of 3-momentum. Section V deals with angular momentum. The spi procedure for singling out the preferred Poincaré subgroup of the spi group is recast in the 3 + 1 language. This yields a method for singling out the asymptotic Euclidean group associated with any asymptotically flat 3-slice. In the final picture, the method involves only that structure which is available in the ADM framework and makes no reference to four-dimensional fields or conformal completion. The geometrical significance of the additional boundary condition necessary for the Euclidean reduction is discussed. The spatial components of the spi angular momentum are then recast in terms of Cauchy data. The procedure also enables one to see clearly why it is difficult to deal with timelike supertranslations and supertranslation-free boosts in the three-dimensional frameworks. The material on angular momentum is so arranged that users of three-dimensional framework can see the problem [discussion following Eq. (28) in Sec. II] and its resolution (Sec. V) directly in this framework without having to first understand the spi formalism in detail.

II. PRELIMINARIES

We shall first recall briefly the basic ideas of the spi framework,[2,3] then summarize the results on the existence of asymptotically flat initial data sets in space-times admitting i°, and finally discuss the supertranslation ambiguities in the ADM expression of angular momentum.

Definition 1[3]: A space-time (M, g_{ab}) will be said to be asymptotically empty and flat at spatial infinity if there exists a space-time $(\hat{M}, \hat{g}_{ab})$ which is smooth everywhere except at a point i°, where $\hat{g}_{ab}$ is $C^{>0}$, together with an imbedding of M into $\hat{M}$ (with which we identify M with its image in $\hat{M}$) satisfying the following conditions: (i) $\overline{J}(i^\circ) = \hat{M} - M$; (ii) there exists a function Ω on $\hat{M}$ which is $C^{>2}$ at i° and smooth elsewhere such that on M, $\hat{g}_{ab} = \Omega^2 g_{ab}$, and at i°, $\Omega = 0$, $\hat{\nabla}_a \Omega = 0$ and $\hat{\nabla}_a \hat{\nabla}_b \Omega = 2\hat{g}_{ab}$; and, (iii) the Ricci tensor R_{ab} of g_{ab} is such that R_{ab} admits a regular direction-dependent limit at i°.

Here $\overline{J}(i^\circ)$ is the closure of the region in $(\hat{M}, \hat{g}_{ab})$ which is causally related to i°. Thus, the first condition requires that i°

[a)] Alfred P. Sloan Research Fellow.
[b)] Equipe de recherche associée au C.N.R.S.

 0022-2488/84/092682-09$02.50

be spacelike related to all points of the physical space-time. The second ensures that the conformal factor Ω falls off as $1/r^2$ and the third demands that, in the physical space-time, the stress-energy tensor of matter should fall off as $1/r^4$.[5] The $C^{>0}$ condition on $\hat{g}_{ab}$ implies that $\hat{g}_{ab}$ is continuous at i°, smooth in its "angular dependence," but that its radial derivatives may have finite discontinuities at i°. Thus, the Christoffel symbols of $\hat{g}_{ab}$ may have radial discontinuities (although they are smooth in their angular dependence). These discontinuities turn out to be a measure of the total energy momentum of the space-time. This differentiability condition on $\hat{g}_{ab}$ implies that its Riemann tensor, $\hat{R}_{abcd}$, is such that $\Omega^{1/2}\hat{R}_{abcd}$ admits a regular direction-dependent limit $\mathbf{R}_{abcd}(\eta)$ at i°, where η is the unit spacelike tangent vector at i° to the curve along which the limit is taken. The Weyl tensor $\mathbf{C}_{abcd}(\eta)$ can be split into "electric" and "magnetic" parts with respect to η; $\mathbf{E}_{ab}(\eta) = \mathbf{C}_{ambn}(\eta)\eta^m\eta^n$, and, $\mathbf{B}_{ab}(\eta) = {}^*\mathbf{C}_{ambn}(\eta)\eta^m\eta^n$. (Note that η is spacelike rather than timelike; hence the quotation marks in "electric" and "magnetic.") $\mathbf{E}_{ab}$ and $\mathbf{B}_{ab}$ are smooth, symmetric traceless tensor fields on the hyperboloid $\mathscr{D}$ of unit spacelike directions at i° and satisfy

$$\mathbf{D}_{[a}\mathbf{E}_{b]c} = 0 \quad \text{and} \quad \mathbf{D}_{[a}\mathbf{B}_{b]c} = 0 \tag{2.1}$$

in virtue of the asymptotic vacuum equations satisfied by the physical metric g_{ab}. (Here, $\mathbf{D}$ is the derivative operator on $\mathscr{D}$ compatible with its intrinsic metric $\mathbf{h}_{ab} = \mathbf{g}_{ab} - \boldsymbol{\eta}_a\boldsymbol{\eta}_b$, where $\mathbf{g}_{ab} = \hat{g}_{ab|i^\circ}$.) $\mathbf{E}_{ab}$ and $\mathbf{B}_{ab}$ capture $1/r^3$-part of the physical Weyl tensor. The Ricci part, $\mathbf{R}_{ab}(\eta)$ of $\mathbf{R}_{abcd}(\eta)$ only provides potentials for $\mathbf{E}_{ab}$ and $\mathbf{B}_{ab}$. Set $\mathbf{S}_{ab}(\eta) = \mathbf{R}_{ab}(\eta) - \frac{1}{6}\mathbf{R}(\eta)\,\mathbf{g}_{ab}$ and, $\mathbf{K}_{ab} = \mathbf{h}_a{}^m\mathbf{h}_b{}^n\mathbf{S}_{mn} - \mathbf{S}_{mn}\,\eta^m\eta^n\mathbf{h}_{ab}$. Then $\mathbf{K}_{ab}$ is a smooth tensor field on $\mathscr{D}$ and satisfies

$$\mathbf{B}_{ab} = \tfrac{1}{4}\boldsymbol{\epsilon}_{amn}\mathbf{D}^m\mathbf{K}^n{}_b\,, \tag{2.2}$$

where $\boldsymbol{\epsilon}_{abc}$ is the alternating tensor on $(\mathscr{D},\mathbf{h}_{ab})$. (The potential for $\mathbf{E}_{ab}$ will not be needed in what follows.)

The 4-momentum $\mathbf{P}_a$ is a covector at i°, constructed from $\mathbf{E}_{ab}$: Given any vector $\mathbf{V}^a$ at i°, we have

$$\mathbf{P}_a\mathbf{V}^a := \frac{1}{8\pi}\oint_C \mathbf{E}_{ab}\mathbf{V}^a\,d\mathbf{S}^b\,, \tag{2.3}$$

where the integral is performed on *any* cross section C of $\mathscr{D}$. One may imagine defining a "magnetic-type" (or, NUT[6]) 4-momentum ${}^*\mathbf{P}_a$ by replacing $\mathbf{E}_{ab}$ by $\mathbf{B}_{ab}$ in Eq. (2.3). However, because of Eq. (2.2) and because $\mathbf{K}_{ab}$ is smooth, ${}^*\mathbf{P}_a$ vanishes identically. (The situation is analogous to that in electrodynamics. To have a nonzero "magnetic" 4-momentum, one would have to weaken boundary conditions and allow "wire singularities" in $\mathbf{K}_{ab}$.) The group of asymptotic symmetries preserving the boundary conditions of Definition 1 is infinite dimensional: it is the semidirect product of the Lorentz group with the group under addition of functions on $\mathscr{D}$ (i.e., the group of supertranslations at i°). This group admits a preferred 4-parameter abelian group of translations and the 4-momentum $\mathbf{P}_a$ lies in the space dual to the corresponding four-dimensional Lie algebra. To obtain (the familiar, 6-component) angular momentum, however, one has to impose stronger boundary conditions and reduce the infinite-dimensional asymptotic symmetry group—called the spi group—to Poincaré. The required additional condition is

$$\mathbf{B}_{ab} \equiv \lim_{\to i^\circ} \Omega^{1/2}{}^*\hat{C}_{ambn}(\hat{\nabla}^m\Omega^{1/2})(\hat{\nabla}^n\Omega^{1/2}) = 0\,. \tag{2.4}$$

(This condition is automatically satisfied if the space-time is stationary *or* axisymmetric.[7]) If this condition is satisfied, Eq. (2.2) implies that we can always require

$$\mathbf{K}_{ab} = 0\,. \tag{2.5}$$

This last condition is left invariant by translations but not by (other) supertranslations. Hence, if it is included in the boundary conditions to be preserved by asymptotic symmetries, the asymptotic symmetry group reduces from the spi group to the Poincaré. [At null infinity, one may analogously demand: $\lim \Omega^{-1}\,{}^*\hat{C}_{ambn}\hat{n}^m\hat{n}^n = 0$, where $\hat{n}^m = \hat{\nabla}^m\Omega$ is the (null) normal to $\mathscr{I}$. This implies that the intrinsic connection $\{D\}$ on $\mathscr{I}$ is free of Bondi news. If we include $\{D\}$ in the universal structure that should be preserved by asymptotic symmetries, the Bondi–Metzner–Sachs (BMS) group reduces to the Poincaré.[8] Unfortunately, unlike (2.4), the above condition is too strong at null infinity since it rules out radiation.] Finally, to obtain the expression of angular momentum, one requires that $\hat{\beta}_{ab} := \hat{C}_{ambn}(\hat{\nabla}^m\Omega^{1/2})(\hat{\nabla}^n\Omega^{1/2})$ should admit a regular direction-dependent limit $\hat{\boldsymbol{\beta}}_{ab}$ at i° and defines $\mathbf{M}_{ab}$ at i° by

$$\mathbf{M}_{ab}\mathbf{F}^{ab} := \frac{1}{8\pi}\oint_C \hat{\boldsymbol{\beta}}_{ab}\boldsymbol{\zeta}^a\,d\mathbf{S}^b \tag{2.6}$$

for arbitrary skew tensor $\mathbf{F}_{ab}$ at i°, where $\boldsymbol{\zeta}^a := \frac{1}{2}\boldsymbol{\epsilon}^{abcd}\mathbf{F}_{cd}\,\boldsymbol{\eta}_b$ is a Killing vector on $(\mathscr{D},\mathbf{h}_{ab})$. Under the action of translations, $\hat{\boldsymbol{\beta}}_{ac}$ transforms in just such a way as to change $\mathbf{M}_{ab}$ in the familiar fashion. The condition $\mathbf{B}_{ab} = 0$ also features in Penrose's recent approach to conserved quantities[9]: it is precisely the condition needed in the construction of the asymptotic twistor space which enables one to take limits of Penrose's quasilocal quantities to spatial infinity.[10] The limiting quantities agree with $\mathbf{P}^a$ and $\mathbf{M}_{ab}$ constructed above. (The condition $\mathbf{B}_{ab} = 0$ also arises in Hawkings[11] definition of "asymptotically De Sitter space-times.") From theorems proved by Choquet-Bruhat, Christodoulou, and O'Murchadha on the Cauchy problem in general relativity[12] it follows that there is a large class of space-times which satisfy Definition 1 and have $\mathbf{B}_{ab} = 0$. (For details up to this point, see Ref. 3.)

Let (M,g_{ab}) satisfy Definition 1 with completion $(\hat{M},\hat{g}_{ab})$. Let Σ be any spacelike submanifold of $\hat{M}$, passing through i°, which is $C^{>1}$ at i° and smooth elsewhere. This means[3] that the metric $\hat{q}_{ab}$ induced on $\hat{\Sigma}$ by $\hat{g}_{ab}$ is $C^{>0}$ at i° and that the extrinsic curvature $\hat{\pi}_{ab}$ admits regular direction-dependent limits there. In Appendix B of Ref. 2 it was shown that $(\hat{q}_{ab},\hat{\pi}_{ab})$ satisfies the Geroch[13] version of the ADM asymptotic conditions. In Appendix A of this paper, we consider the pair (q_{ab},π_{ab}) induced on $\Sigma = (\hat{\Sigma} - i^\circ)$ by g_{ab} and show that there exists a flat metric f_{ab} outside some compact set of Σ, such that, in the Cartesian chart defined by f_{ab}, the components of tensor fields $q_{ab} - f_{ab}$, $\partial_a q_{bc}$, π_{ab}, and $\partial_a\partial_b q_{cd}$ fall off at least as fast as $1/r$, $1/r^2$, $1/r^2$ and $1/r^3$, respectively, as r tends to infinity, where ∂_a and r are the derivative operator and the radial coordinate defined by f_{ab}. Thus, the Cauchy data (q_{ab},π^{ab}) satisfy the usual ADM con-

ditions. Furthermore, since the stress energy of matter sources falls off as $1/r^4$ [condition (iii) in Definition 1], one can use the Hamiltonian arguments in the spirit of ADM to obtain conserved quantities. The ADM formula for energy is

$$E = \frac{1}{16\pi} \lim_{r_0 \to \infty} \oint_{r=r_0} (\partial_a q_{bc} - \partial_b q_{ac}) f^{ac}\, dS^b . \tag{2.7}$$

To compute the 3-momentum and the angular momentum, one has to select suitable vector fields which generate spatial translations and rotations at infinity. Let N^a be any Killing field of the flat metric f_{ab}. The diffeomorphism generated by N^a is a spi asymptotic symmetry (restricted to Σ). Let us set

$$Q_{N^a} = \frac{1}{8\pi} \lim_{r_0 \to \infty} \oint_{r=r_0} (\pi_{ab} - \pi^m{}_m q_{ab}) N^a\, dS^b . \tag{2.8}$$

If N^a is an asymptotic translation ($\partial_a N_b = 0$), ADM interpret Q_{Na} to be the component of the 3-momentum along N^a, and if N^a is an asymptotic rotation ($\partial_{(a} N_{b)} = 0$, $\partial_a N_b \neq 0$), to be the component of the angular momentum along the axis of N^a. Note, however, that the entire procedure is tied down to the initial choice of the flat metric f_{ab}. What happens under the change of f_{ab}? To investigate this issue we proceed as follows. Let X_1, X_2, X_3 be a set of Cartesian coordinates on Σ defined by f_{ab}: $f_{ab}\, dX^a\, dX^b = dX_1^2 + dX_2^2 + dX_3^2$. Introduce a new set $X'_{\mathbf{K}}$ of coordinates, supertranslated with respect to $X_{\mathbf{K}}$,

$$X'_{\mathbf{K}} = X_{\mathbf{K}} + \Lambda_{\mathbf{K}}(\theta,\varphi) \tag{2.9}$$

and consider the metric f'_{ab} defined by $f'_{ab} dX'^a\, dX'^b = dX'^2_1 + dX'^2_2 + dX'^2_3$. Clearly, f'_{ab} is also flat but not equal to f_{ab}. To compare the two metrics, let us compute the components $f'_{\mathbf{ab}}$ of f'_{ab} in the X_K chart[14]:

$$f'_{\mathbf{ab}} = \frac{\partial X'^c}{\partial X^a} \frac{\partial X'^d}{\partial X^b} f'_{c'd'} = \delta_{\mathbf{ab}} + O(1/r) .$$

Hence $f_{ab} - f'_{ab}$ (and therefore, $q_{ab} - f'_{ab}$) falls off as $1/r$. Thus, f'_{ab} is as good a flat metric as f_{ab}. The energy expression (2.7) is unchanged if f'_{ab} and ∂'_a are used in place of f_{ab} and ∂_a. Since the translations N'^a of f'_{ab} differ from the translations N^a of f_{ab} by terms which fall off as $1/r$, the expression for 3-momentum is also unchanged. For rotations, however, the situation is more complicated. Let $N^a = \partial/\partial\varphi$ be the Z-directional rotation with respect to f_{ab} and $N'^a = \partial/\partial\varphi'$ be the Z'-directional rotation with respect to f'_{ab}. Asymptotically, Z direction coincides with the Z' direction. However, $N^a - N'^a$ does not go to zero: the components of $N^a - N'^a$ in the X'_K chart are

$$\left(\frac{\partial \Lambda_1}{\partial \varphi} + \Lambda_2, \frac{\partial \Lambda_2}{\partial \varphi} - \Lambda_1, \frac{\partial \Lambda_3}{\partial \varphi} \right).$$

Thus, at infinity, $N^a - N'^a$ yields an angle-dependent translation, or a supertranslation, whence, $Q_{N^a} - Q_{N'^a}$ is nonzero. (Recall that π_{ab} falls off as $1/r^2$ and the 2-surface volume element blows up like r^2. Hence, the integral $\oint(\pi_{ab} - \pi q_{ab})(N^a - N'^a) dS^b$ will not, in general, vanish even in the limit.) Furthermore, by choosing suitable $\Lambda_{\mathbf{K}}(\theta,\varphi)$, one can get, for the z-component of the angular momentum, any answer one wishes (except in special cases in which π^{ab} is, e.g., trace-free). Thus, as it stands, the ADM formula for angular momentum is simply not well defined. The supertranslation ambiguities in this formula have to be removed by imposing stronger boundary conditions which, for example, will permit only a preferred class of flat metrics, related to one another by asymptotic *translations* only. As it stands, the ADM angular momentum is defined only up to additions of supermomenta and hence cannot even transform correctly under boosts of the 3-surface. (One knows[2] from the spi framework that supermomenta are not conserved under boosts; in fact they can change by arbitrary amounts.) Why then, have explicit applications of the ADM angular momentum formula not shown these ambiguities? Apparently, explicit evaluations of angular momentum have been carried out only in the axisymmetric case, where the rotational Killing field—rather than a supertranslated version thereof—offers itself as a natural candidate for being N^a.[15]

III. THE ENERGY

Let, as before, $\hat{\Sigma}$ be a spacelike submanifold (of $\hat{M}$) which is $C^{>1}$ at i° and smooth elsewhere. Let $\mathbf{t}^a$ be the unit normal to $\hat{\Sigma}$ at i°. The spi energy, associated with $\hat{\Sigma}$, is given by[16]

$$-\mathbf{P}_a \mathbf{t}^a = -\frac{1}{8\pi} \oint_C \mathbf{E}_{ab} \mathbf{t}^a \mathbf{t}^b\, \mathbf{d}^2\mathbf{S} \tag{3.1}$$

where $\mathbf{d}^2\mathbf{S}$ is the volume element of a unit 2-sphere. We wish to rewrite this expression in terms of the Cauchy data (q_{ab}, π^{ab}), induced on $\Sigma := \hat{\Sigma} - i^\circ$ by g_{ab}. We therefore begin by expressing the integrand in terms of fields defined within the space-time. By definition, we have $\mathbf{E}_{ab} = \lim \Omega^{1/2} \hat{C}_{ambn} \hat{\eta}^m \hat{\eta}^n$, while a simple calculation shows that, if t^a is the unit normal to Σ with respect to g_{ab}, $\lim_{\to i^\circ} \Omega^{-1} t^a = \mathbf{t}^a$. Furthermore, as is shown in Appendix A, Σ admits a flat metric f_{ab} the inverse square, r^{-2}, of whose radial coordinate r can be taken, without loss of generality, to be the (restriction to Σ of the) conformal factor Ω. Let us make this choice. Then, we have

$$-\mathbf{P}_a \mathbf{t}^a = -\frac{1}{8\pi} \lim_{r_0 \to \infty} \int_{r=r_0} r^3 C_{abcd}\, \eta^b \eta^d t^a t^c\, \mathbf{d}^2\mathbf{S} , \tag{3.2}$$

where C_{abcd} is the Weyl tensor of g_{ab} and $\eta^b = f^{ab}\partial_a r$, the unit radial normal to the 2-spheres $r = r_0$, with respect to f_{ab}. Now, $C_{abcd} t^a t^c = e_{bd}$, the (usual) electric part of the Weyl tensor with respect to Σ. This electric part is related to the Cauchy data (q_{ab}, π^{ab}) via

$$e_{ab} = \mathcal{R}_{ab} - \pi_a{}^m \pi_{mb} + \pi\pi_{ab} - \tfrac{1}{2}(q_a{}^m q_b{}^n + q_{ab} q^{mn})(R_{mn} - \tfrac{1}{6} R_{cd} g^{cd} g_{mn}) , \tag{3.3}$$

where $\mathcal{R}_{ab}$ and R_{ab} are, respectively, the Ricci tensors of q_{ab} and g_{ab}. Now, it is shown in Appendix A that π_{ab} falls off as $1/r^2$ and it follows from condition (iii) of Definition 1 that R_{ab} falls off as $1/r^4$. Hence, we have

$$-\mathbf{P}_a t^a = -\frac{1}{8\pi} \lim_{r_0 \to \infty} \oint_{r=r_0} r^3 \mathcal{R}_{ab}\, \eta^a \eta^b\, \mathbf{d}^2\mathbf{S} . \tag{3.4}$$

One sees already that the energy is independent of π_{ab}. Let the derivative operators D and ∂, compatible with q_{ab} and f_{ab}, respectively, be related by

$$(D^a - \partial^a)K_b = C_{ab}{}^c K_c\,, \quad \forall K_b\,. \tag{3.5}$$

Then

$$C_{ab}{}^c = -\tfrac{1}{2}q^{cd}(\partial_a q_{bd} + \partial_b q_{ad} - \partial_d q_{ab}) \tag{3.6}$$

and the Riemann tensor $\mathscr{R}_{abc}{}^d$ of q_{ab} is given by

$$\mathscr{R}_{abc}{}^d = 2(\partial_{[a}C_{b]c}{}^d + C_{c[a}{}^m C_{b]m}{}^d)\,. \tag{3.7}$$

Substituting (3.7) in (3.4), using the fact that $(q_{ab} - f_{ab})$, $\partial_a q_{bc}$, and $\partial_a\partial_b q_{cd}$ fall off as $1/r$, $1/r^2$, and $1/r^3$, respectively, and using the fact that $\partial_a\partial_b r = (1/r)(f_{ab} - \partial_a r\partial_b r)$, one obtains, after some simplification

$$-\mathbf{P}_a\mathbf{t}^a = \frac{1}{16\pi}\lim_{r_0\to\infty}\oint_{r=r_0}(\partial_c q_{ab} - \partial_b q_{ac})f^{ac}\,dS^b\,. \tag{3.8}$$

This is precisely the ADM energy formula. The simplification involves the application of Stokes' theorem to set certain integrals equal to zero and hence calculation does *not* imply that the integrands of Eq. (3.4) and of (3.8) are equal even in the limit.

Remarks 1: The same techniques as used above enable one to show that the energy integral (3.1) can be reduced to other useful forms, some of which have appeared in the literature in the context of the positive energy theorem. We shall just list these alternate expressions.

(*a*) *Expressions in terms of physical fields on* Σ: In addition to the ADM expression (2.7), we have the expression in terms of the radial–radial component of the Ricci tensor,[17]

$$E = -\frac{1}{8\pi}\lim_{r_0\to\infty}\oint_{r=r_0} r\mathscr{R}_{ab}\,\eta^a\eta^b\,d^2V_q\,, \tag{3.4}$$

which suggests that the component may be thought of as the "mass aspect" in $3+1$ frameworks, and the following expression in terms of the tensor $C_{ab}{}^c$ of Eqs. (3.5) and (3.6)

$$E = \frac{1}{4\pi}\lim_{r_0\to\infty}\oint_{r=r_0} C_{ab}{}^c\eta_c\,\eta^b\,dS^a\,. \tag{3.9}$$

This expression involves the asymptotic behavior of the metric connection D and is therefore very similar to Witten's[4] expression involving spin coefficients.

(*b*) *Expressions involving 2-sphere fields*: Consider the foliation of Σ by a family of metric 2-spheres of f_{ab} (surfaces $r = \text{const}$). Using Gauss–Codazzi equations, one can express $\mathscr{R}_{ab}\,\eta^a\eta^b$ in terms of the intrinsic and the extrinsic curvature induced on these 2-spheres by q_{ab}, and obtain, using Eq. (3.4)

$$E = \frac{1}{16\pi}\lim_{r_0\to\infty}\oint_{r=r_0} r({}^2R - ({}^2\pi)^2 + {}^2\pi^{ab}{}^2\pi_{ab})\partial^2V_q\,, \tag{3.10}$$

where ${}^2R$ is the scalar curvature of the 2-sphere $r = r_0$; ${}^2\pi_{ab}$, the extrinsic curvature, and ${}^2\pi$ its trace. If f_{ab} is so chosen that the metrics on $r = r_0$ 2-spheres induced by q_{ab} and f_{ab} differ by terms of the order of $1/r^2$ rather than $1/r$,[18] this expression simplifies to the Geroch expression used by Jang[4] in the context of positivity of energy:

$$E = \frac{1}{32\pi}\lim_{r_0\to\infty}\oint_{r=r_0} r({}^2R - ({}^2\pi)^2)\partial^2V_q\,. \tag{3.11}$$

(*c*) *Expressions in terms of conformally rescaled fields on* $\hat{\Sigma}$: As remarked earlier, the initial data (q_{ab}, π^{ab}) on Σ satisfy the Geroch version[13] of the ADM asymptotic conditions.[2] In the Geroch approach, one uses the conformally rescaled fields $\hat{q}_{ab} = \Omega^2 q_{ab}$, $\hat{\pi}^{ab} = \Omega^{-3}\pi^{ab}$ and $\mathbf{\Omega}_{ab} = \lim_{\to i^\circ}\Omega^{-1/2}(\hat{D}_a\hat{D}_b\Omega - 2\hat{q}_{ab})$. Starting from (3.4), one can show[17]:

$$E = -\frac{1}{24\pi}\oint_C \mathbf{\Omega}_{ab}\,\boldsymbol{\eta}^a\boldsymbol{\eta}^b\,d^2S \tag{3.12}$$

$$= -\frac{1}{24\pi}\oint_C \mathbf{\Omega}_{ab}\,\mathbf{q}^{ab}\,d^2S\,, \tag{3.13}$$

where the integrals are taken on the cross section C of the hyperboloid $\mathscr{D}$ (of unit spacelike vectors in the tangent space of i°) obtained by the intersection of $\mathscr{D}$ with (the tangent space at i° to) Σ.

Remark 2: The "magnetic" analog of energy—the NUT charge—can be obtained by substituting $\mathbf{B}_{ab}$ in (3.1) in place of $\mathbf{E}_{ab}$ and is nonzero only if the present asymptotic conditions are weakened to allow K_{ab} [of Eq. (2.2)] to develop "wire singularities." In the $3+1$ framework this corresponds to keeping the boundary conditions on q_{ab} as they are but allowing π_{ab} to develop "wire singularities." The $3+1$ reduction of the spi formula then yields, for the NUT charge,[16]

$$-{}^*\mathbf{P}_a\mathbf{t}^a = -\frac{1}{8\pi}\oint_C \mathbf{B}_{ab}\mathbf{t}^a\mathbf{t}^b\,d^2\mathbf{S} \tag{3.14}$$

$$= -\frac{1}{4\pi}\lim_{r_0\to\infty}\oint_{r=r_0} r(D_m\pi_{nb}\,\eta^b)dS^{mn}\,. \tag{3.15}$$

IV. THE 3-MOMENTUM

The steps leading from the spi-expression of the 3-momentum to the ADM expression are completely analogous to those which led us from Eq. (3.1) to Eq. (3.8). For completeless, we sketch them briefly.

The component of the spi 3-momentum along a spacelike vector $\mathbf{N}^a$ at i° (which is tangential to $\hat{\Sigma}$) is given by

$$\mathbf{P}_a N^a = \frac{1}{8\pi}\oint_C \mathbf{E}_{ab}\mathbf{N}^a\,d\mathbf{S}^b\,. \tag{4.1}$$

As before, one rewrites this expression as a limit of integrals over metric 2-spheres of f_{ab} within Σ to obtain

$$\mathbf{P}_a\mathbf{N}^a = \frac{1}{8\pi}\lim_{r_0\to\infty}\oint_{r=r_0} r^3C_{abcd}\,\eta^b\eta^d N^a t^c\,d^2\mathbf{S}\,, \tag{4.2}$$

where N^a is a translational Killing field of f_{ab} which induces the spi translation corresponding to $\mathbf{N}^a$ at i°. [Note that, at i°, the limit of $\Omega^{-1}N^a$ is direction dependent and is given by $\mathbf{N}^a - 2(\mathbf{N}\cdot\boldsymbol{\eta})\eta^a$. However, the term proportional to η^a does not contribute to (4.2) since the Weyl tensor is already contracted with η^b and η^d.] One can now expand the physical Weyl tensor C_{abcd} in terms of its (usual) electric and magnetic parts relative to Σ, $e_{ab} = C_{ambn}t^m t^n$ and $b_{ab} = {}^*C_{ambn}t^m t^n$ to obtain

$$C_{abcd}\,\eta^b\eta^d N^a t^c = -\epsilon_{adm} b^{m}{}_{b}\eta^b\eta^d N^a\,, \tag{4.3}$$

where ϵ_{abc} is the alternating tensor on (Σ, q_{ab}). Next, using the identity

$$b_{ab} = \epsilon_{mn(a} D^{m}\pi^{n}{}_{b)} \tag{4.4}$$

one gets

$$\mathbf{P}_a\mathbf{N}^a = \frac{1}{8\pi}\lim_{r_0\to\infty}\oint_{r=r_0} r^3 \times\{D_{[d}\,\pi_{a]b} + \tfrac{1}{2}\epsilon_{dap}\epsilon_{mnb} D^{m}\pi^{np}\}\eta^b\eta^d N^a\, d^2\mathbf{S}\,. \tag{4.5}$$

One now simplifies the integral using the facts stated below Eq. (3.7) to show the desired result:

$$\mathbf{P}_a\mathbf{N}^a = \frac{1}{8\pi}\lim_{r_0\to\infty}\oint_{r=r_0} N^a(\pi_{ab} - \pi q_{ab})dS^b\,. \tag{4.6}$$

Again, the simplification involves the use of Stokes' theorem to discard certain integrals, whence the calculation does *not* show that the integrand of Eq. (4.2) [or (4.5)] is equal to that of Eq. (4.6) even in the limit. Finally, the techniques used in the simplification also yield other, equivalent expressions for the 3-momentum in terms of the initial data. Perhaps the most useful among these for explicit evaluations in special cases is

$$\mathbf{P}_a\mathbf{N}^a = \frac{1}{4\pi}\lim_{r_0\to\infty}\oint_{r=r_0}(N_a\,\eta^a)\pi_{cb}\,\eta^c\,dS^b\,. \tag{4.7}$$

V. THE SPATIAL ANGULAR MOMENTUM

This section is divided into two parts. In the first, we discuss the issue of reducing the infinite-dimensional asymptotic symmetry group associated with $(\Sigma, q_{ab}, \pi^{ab})$ (the "spatial spi-group") to the Euclidean group, and in the second, re-express the spi expression for the spatial component of the angular momentum in terms of initial data. For the convenience of readers who may not be familiar with the details of the spi construction, here the Euclidean reduction will be carried out entirely in the 3 + 1 framework, spi formalism being used only to motivate the required additional fall-off condition. The proof that the Euclidean group so obtained is the same as the one which results from the spi construction is given in Appendix B.

A. Euclidean reduction

The additional condition needed to reduce the spi group to the Poincaré group is $\mathbf{B}_{ab} = 0$ [Eq. (2.4)]. Now we are interested only in the 3-surface Σ, rather than the entire space-time, and hence, in removing only the ambiguities associated with spatial supertranslations within Σ. An examination of the discussion surrounding Eqs. (2.4) and (2.5) yields the relevant condition. It is[19]

$$\mathbf{B}_{ab}\mathbf{t}^a\gamma^b{}_m = 0\,, \tag{5.1}$$

where $\mathbf{t}^a$ is, as before (the vector field induced on $\mathscr{D}$ by) the unit normal to $\hat{\Sigma}$ at i^o, and $\gamma_b{}^m$, the metric on the 2-sphere cross section of $\mathscr{D}$, perpendicular to $\mathbf{t}^a$. To translate this condition in terms of physical space-time, let us introduce a flat metric f_{ab} in a neighborhood of infinity of Σ and consider a family of metric 2-spheres of f_{ab}. If γ_{ab} denotes the metric induced on this family *by the physical metric* g_{ab}, and r, the radial coordinate with respect to f_{ab}, Eq. (5.1) reduces to

$$\lim_{r\to\infty} r^3\mathscr{R}_{ab}\,\eta^a\gamma^b{}_m = 0 \tag{5.2}$$

in virtue of the definition of $\mathbf{B}_{ab}$, Eq. (3.3), and the fall-off conditions satisfied by π_{ab} and the matter sources. Thus, the spi formalism implies that the Euclidean reduction is possible if the radial-angular components of the three-dimensional Ricci tensor fall-off faster than what is guaranteed by the fall-off of q_{ab}. (Note that the radial–radial component of $\mathscr{R}_{ab}$ cannot be required to fall-off faster than $1/r^3$ unless the energy associated with q_{ab} is zero.) It is straightforward to check that (5.2) is supertranslation invariant: if it is satisfied by the metric 2-spheres of f_{ab}, it is satisfied by the metric 2-spheres of *any* supertranslated flat metric f'_{ab}.

We are now ready to carry out the reduction. Equation (5.2) is, via Gauss–Codazzi equations, equivalent to

$$\lim_{r\to\infty} r^3\,{}^2D_a({}^2\pi^{ab} - {}^2\pi\gamma^{ab}) = 0\,, \tag{5.3}$$

where ${}^2D$ is the derivative operator compatible with γ_{ab}, ${}^2\pi_{ab}$ is the extrinsic curvature with respect to q_{ab} of the 2-spheres under consideration, and ${}^2\pi$ its trace. The idea is to solve this equation and choose the preferred Euclidean group by imposing suitable conditions on the solution. For this purpose, we first express ${}^2D_a$ in terms of the derivative operator $\eth_a$ induced on the 2-spheres by the *flat* metric f_{ab} to obtain

$$\eth_a({}^2\pi^{ab} - {}^2\pi\gamma^{ab}) - (1/r)\eth_a(\mathring{\gamma}^{am}\mathring{\gamma}^{bn}(\gamma_{mn} - \mathring{\gamma}_{mn})) = o(1/r^3)\,, \tag{5.4}$$

where $\mathring{\gamma}_{ab}$ is the metric induced on 2-spheres by f_{ab}, $\mathring{\gamma}^{ab}$ is its inverse, and where the remainder $o(1/r^3)$ satisfies $\lim_{r\to\infty} r^3 o(1/r^3) = 0$. Now, on a 2-sphere of radius r (with respect to f_{ab}), $\eth$ is the standard 2-sphere derivative operator (compatible with $\mathring{\gamma}_{ab}$) and the general solution to the equation $\eth_a\mathbf{p}^{ab} = 0$ with $\mathbf{p}^{ab} = p^{(ab)}$ is (Appendix C)

$$\mathbf{p}^{ab} = -\eth^a\eth^b g + [\mathring{\gamma}^{mn}\eth_m\eth_n g + (1/r^2)g]\mathring{\gamma}^{ab} \tag{5.5}$$

for some function g on the 2-sphere. Hence, it follows that the general solution to (5.4) is

$$({}^2\pi^{ab} - {}^2\pi\gamma^{ab}) = -\eth^a\eth^b g + (\eth^2 + (1/r^2)g)\mathring{\gamma}^{ab} + (1/r)(\mathring{\gamma}^{am}\mathring{\gamma}^{bn}(\gamma_{mn} - \mathring{\gamma}_{mn})) + o(1/r^2)\,, \tag{5.6}$$

where g is a sum of terms of the type[20] $g = \Sigma h_i(r)K_i(\theta,\varphi)$, and where $\eth^2 = \mathring{\gamma}^{mn}\eth_m\eth_n$. Now, a direct calculation of ${}^2\pi^{ab}$ yields [irrespective of whether (5.2) is imposed]

$${}^2\pi^{ab} = (1/r)\gamma^{ab} + O(1/r^2)\,, \tag{5.7}$$

where $\lim_{r\to\infty} r^2 O(1/r^2)$ exists but is not necessarily zero. Hence, it follows that there exists a function f of the type $f = K(\theta,\varphi) + \Sigma\tilde{h}_i(r)\tilde{K}_i(\theta,\varphi)$ where $\lim_{r\to\infty}\tilde{h}_i(r) = 0$, such that

$$\begin{aligned}({}^2\pi^{ab} - {}^2\pi\gamma^{ab}) &= -\eth^a\eth^b f + \left(\eth^2 + \frac{1}{r^2}\right)f\mathring{\gamma}^{ab} + \frac{1}{r}\mathring{\gamma}^{am}\mathring{\gamma}^{bn}(\gamma_{mn} - \mathring{\gamma}_{mn}) + o\left(\frac{1}{r^2}\right)\\ &= -\frac{1}{r}\mathring{\gamma}^{ab} - \eth^a\eth^b K + \left(\eth^2 + \frac{1}{r^2}\right)K\mathring{\gamma}^{ab} + o\left(\frac{1}{r^2}\right).\end{aligned} \tag{5.8}$$

[f is given by $f = -r + \tilde{g}$, where $\tilde{g}$ is obtained from g by ignoring terms in its expansion for which $h_i(r)$ diverge or fail to admit a limit.] Taking the trace of (5.8) with respect to γ_{ab}, one obtains

$$^2\pi = \frac{2}{r} - \left(\eth^2 + \frac{2}{r^2}\right)K(\theta,\varphi) + o\left(\frac{1}{r^2}\right). \tag{5.9}$$

Note that, had q_{ab} been *equal to* f_{ab}, $^2\pi$ would have been *equal to* $2/r$ while if q_{ab} had not satisfied (5.2), it would be of the type $^2\pi = 2/r + O(1/r^2)$ [given by (5.7)]. Thus, satisfaction of (5.2) yields an in-between situation in which the terms in $1/r^2$ are present but have the specific form, $-(\eth^2 + 2/r^2)K(\theta,\varphi)$.

Let us now carry out a supertranslation (2.9), $X_{\mathbf{K}} \to X'_{\mathbf{K}} = X_{\mathbf{K}} - \Lambda_{\mathbf{K}}(\theta,\varphi)$ and work out $^2\pi'$, the trace of the extrinsic curvature of the metric 2-spheres of f'_{ab}, computed using q_{ab}. A long but straightforward calculation yields

$$\left(^2\pi' - \frac{2}{r'}\right) - \left(^2\pi - \frac{2}{r}\right) = \left(\eth^2 + \frac{2}{r^2}\right)(\eta^{\mathbf{K}}\Lambda_{\mathbf{K}}) + o\left(\frac{1}{r^2}\right), \tag{5.10}$$

where $\eta^{\mathbf{K}}$ are the components of the unit radial vector $f'^{ab}\partial_b r$ in the Cartesian chart $X_{\mathbf{K}}$. (Note that $\eta^{\mathbf{K}}\Lambda_{\mathbf{K}}$ is independent of r, i.e., is a function of θ and φ only.) Hence, it follows that, if we make a supertranslation with $\Lambda_{\mathbf{K}}\eta^{\mathbf{K}} = K(\theta,\varphi)$, we would have

$$^2\pi' = 2/r' + o(1/r'^2). \tag{5.11}$$

Thus, by an appropriate supertranslation, we can get rid of the terms of the order $O(1/r^2)$. Furthermore, if we wish to preserve the absence of such terms, we can only make such further supertranslations for which $\Lambda_{\mathbf{K}}\ \eta^{\mathbf{K}}$ satisfies

$$(\eth^2 + 2/r^2)\Lambda_{\mathbf{K}}\ \eta^{\mathbf{K}} = 0, \tag{5.12}$$

which is precisely the condition defining translations. Going back to the full extrinsic curvature (5.8), we have, for those flat metrics f'_{ab} which satisfy Eq. (5.11)

$$\lim_{r'\to\infty} r'^2(^2\pi'^{ab} - 1/r'\gamma'^{ab}) = 0. \tag{5.13}$$

Thus while, in general, the metric 2-spheres of any flat metric (to which q_{ab} approaches at infinity as specified in Sec. II) satisfy (5.7), if q_{ab} satisfies (5.2) we can find a family of flat metrics which satisfies the stronger condition (5.13). Two flat metrics belong to this family if and only if they are related by an asymptotic *translation*. Hence, if this family is included in the universal structure to be preserved by asymptotic symmetries, the asymptotic symmetry group reduces to the Euclidean group.

To summarize, the supertranslation freedom arises because there are many flat metrics which asymptotically approach q_{ab} as $1/r$. However, if q_{ab} satisfies (5.2), there exists, among these, a preferred family whose metric 2-spheres have "fewer wiggles, as seen by q_{ab}." Killing vectors of any one of these preferred metrics differ by the corresponding Killing vectors of any other by an *asymptotic translation* rather than a supertranslation. One must use *these* Killing vectors to make the ADM formula unambiguous.

B. Spatial angular momentum

To re-express the spatial components of spi angular momentum (2.6) in terms of (q_{ab},π^{ab}) we proceed as follows. Choose a flat metric f_{ab} in the preferred class so that (the unprimed version of) Eq. (5.13) is satisfied. As shown in Appendix B, there exists, in the spi framework, a conformal frame in the preferred class for which $\mathbf{K}_{ab}$ vanishes on $\mathscr{D}$, such that on Σ, $\Omega = 1/r^2$, where r is a radial coordinate of f_{ab}. Let us work with this conformal frame. Then, β_{ab} in Eq. (2.6) is given by[21]

$$\beta_{ab} = \lim\ {}^*C_{ambn}(q^{mp}D_p r)(q^{ns}D_s r)[q^{cd}(D_c r)(D_d r)]^{-1}.$$

Next, on the cross section C of the hyperboloid $\mathscr{D}$, perpendicular to $\mathbf{t}^a$ (i.e., the intersection of $\mathscr{D}$ with the tangent space to $\hat{\Sigma}$ at i°), $\zeta^a := \frac{1}{2}\epsilon^{abcd}\mathbf{F}_{cd}\boldsymbol{\eta}_b$ is necessarily parallel to $\mathbf{t}^a$, if $\mathbf{F}_{cd}$ is a 2-form tangential to $\hat{\Sigma}$. Set $\zeta^a = \mathbf{f}\mathbf{t}^a$. Then, it is easy to verify that the function $\mathbf{f}$ on C is a "spatial translation," i.e., satisfies $^2\mathbf{D}_a\ ^2\mathbf{D}_b\mathbf{f} + \gamma_{ab}\mathbf{f} = 0$, where γ_{ab} and $^2\mathbf{D}$ are, respectively, the metric and the derivative operator induced on C by the metric $\mathbf{h}_{ab}$ on $\mathscr{D}$. (Thus $\mathbf{f}$ is a spherical harmonic with $l = 1$ on the metric 2-sphere C.) Armed with this information about β_{ab} and ζ^a, we can now re-express (2.6) in terms of the initial data. Assuming that $\mathbf{F}_{ab}$ lies within $\hat{\Sigma}$ at i°, we have for the *spatial angular momentum* $\mathbf{M}_{ab}\mathbf{F}^{ab}$

$$\mathbf{M}_{ab}\mathbf{F}^{ab} = \frac{1}{8\pi}\lim_{r_0\to\infty}\oint_{r=r_0} {}^*C_{ambn}\tilde{\eta}^m\tilde{\eta}^n\mathbf{f}t^a t^b r^2\, d^2V_q, \tag{5.14}$$

where $\tilde{\eta}^a = (q^{mn}D_m rD_n r)^{-1/2}q^{ab}D_b r$ is the unit normal to the 2-sphere $r = r_0$; t^a is, as before, the unit normal to Σ (with respect to g_{ab}); and, d^2V_q is the volume element induced by q_{ab} on the 2-sphere $r = r_0$. Thus,

$$\mathbf{M}_{ab}\mathbf{F}^{ab} = \frac{1}{8\pi}\lim_{r_0\to\infty}\oint_{r=r_0} b_{ab}\,\tilde{\eta}^a\tilde{\eta}^b\mathbf{f}r^2\, d^2V_q \tag{5.15}$$

$$= \frac{1}{8\pi}\lim_{r_0\to\infty}\oint_{r=r_0}\epsilon^{amn}(D_m\pi_{nb})\mathbf{f}\tilde{\eta}_a\tilde{\eta}^b r^2\, d^2V_q. \tag{5.16}$$

Now, we only have to simplify this last integral. Since we began with the well-defined expression (2.6), we know that the limit of the integral in (5.21) exists. However, since the integrand now contains an effective factor of r^4—the product of the explicit r^2 and the r^2 hidden in d^2V_q—and since $D_a\pi_{bn}$ falls off only as $1/r^3$, a greater care is needed in the simplification than was necessary before for the 4-momentum.

By integrating by parts, (5.21) can be re-expressed as

$$\begin{aligned}\mathbf{M}_{ab}\mathbf{F}^{ab} = \frac{1}{8\pi}\lim_{r_0\to\infty}\Bigg\{&\oint_{r=r_0} D_m(\epsilon^{amn}\pi_{nb}\,\tilde{\eta}_a\,\tilde{\eta}^b\mathbf{f})r_0^2\, d^2V_q \\ &- \oint_{r=r_0}\epsilon^{amn}(D_m\,\tilde{\eta}_a\,\tilde{\eta}^b)\mathbf{f}\pi_{nb}r_0^2\, d^2V_q \\ &- \oint_{r=r_0}\epsilon^{amn}\tilde{\eta}_a\,\tilde{\eta}^b\pi_{nb}(D_m\mathbf{f})r_0^2\, d^2V_q\Bigg\} \\ = A + B + C, \quad &\text{say}.\end{aligned} \tag{5.17}$$

Let us simplify each term separately. We have

$$A = \frac{1}{8\pi} \lim_{r_0 \to \infty} \oint_{r=r_0} (\gamma^m{}_p + \tilde{\eta}^m \tilde{\eta}_p) D_m \times (\epsilon^{apn} \tilde{\eta}_a \pi_{nb}\, \tilde{\eta}^b \mathbf{f}) r_0{}^2\, d^2V_q$$

$$= \frac{1}{8\pi} \lim_{r_0 \to \infty} \oint_{r=r_0} \epsilon^{pna} \tilde{\eta}_a (D_p \lambda)\mathbf{f} \pi_{nb}\, \tilde{\eta}^b r_0{}^2\, d^2V_q\,, \qquad (5.18)$$

where $\lambda^{-2} = q^{ab}(D_a r)(D_b r)$. Here we have used Stokes' theorem, the fact that $\tilde{\eta}^a$ is unit and the fact that $\lambda = 1 + O(1/r)$. To simplify B, we again use these two facts. The result is

$$B = -\frac{1}{8\pi} \lim_{r_0 \to \infty} \oint_{r=r_0} [\epsilon^{mna} \tilde{\eta}_a (D_m \lambda)\mathbf{f} \pi_{nb} \tilde{\eta}^b r_0{}^2 + \epsilon^{mna} \tilde{\eta}_a \pi_{nb} ({}^2\pi_m{}^b)\mathbf{f} \lambda r_0{}^2]\, d^2V_q\,, \qquad (5.19)$$

where ${}^2\pi_{ab}$ is, as before, the extrinsic curvature of the 2-sphere $r = r_0$. Now, *since we have chosen the flat metric f_{ab} in the preferred class*, we have $({}^2\pi_{ab} - 1/r\gamma_{ab}) = o(1/r^2)$. *Because of this*, the second term in the above integral vanishes [it would *not* vanish if ${}^2\pi_{ab} - (1/r)\gamma_{ab}$ were $O(1/r^2)$!] and we have

$$B = -A\,. \qquad (5.20)$$

Finally, to simplify C, we note that $\mathring{\epsilon}^{ab} = \lambda^{-1} \epsilon^{abc} \eta_c$ is the natural alternating tensor induced by the *flat* metric f_{ab} on the 2-spheres $r = r_0$ and that

$$\xi^a = r_0{}^2 \mathring{\epsilon}^{ab} D_b \mathbf{f} = r_0{}^2 \lambda^{-1} \epsilon^{abc} \eta_c D_b \mathbf{f} \qquad (5.21)$$

is a rotational Killing field of f_{ab} (whose axis is orthogonal to $\mathbf{F}_{ab}$). Hence,

$$C = \frac{1}{8\pi} \lim_{r_0 \to \infty} \oint_{r=r_0} \lambda \pi_{ab} \xi^a\, dS^b\,. \qquad (5.22)$$

Finally, using the Euclidean reduction condition, one can show that $\lambda = 1 + o(1/r)$. [That is, because $({}^2\pi_{ab} - (1/r)\gamma_{ab}) = o(1/r^2)$, the volume elements $\epsilon_{ab} = \epsilon^{abc}\eta_c$ and $\mathring{\epsilon}^{ab} = \epsilon^{abc} D_c r$, induced on the 2-spheres by q_{ab} and f_{ab}, respectively, differ by terms of the order $o(1/r)$ rather than $O(1/r)$.] Hence, we have, combining (5.17), (5.20), and (5.22),

$$\mathbf{M}_{ab}\mathbf{F}^{ab} = \frac{1}{8\pi} \lim_{r_0 \to \infty} \oint_{r=r_0} \pi_{ab} \xi^a\, dS^b\,, \qquad (5.23)$$

which is the ADM expression. Thus, the ADM prescription is not "wrong"; it is incomplete. Finally, note that the simplification used the Euclidean reduction condition (5.13) crucially; had the flat metric f_{ab} not been in the preferred family, the spi expression would *not* have led to the ADM expression.

Remarks: We have seen that, if the angular-radial components of an asymptotically flat metric q_{ab} fall off faster than $1/r^3$, one can invariantly associate to q_{ab} a preferred family of flat metrics f_{ab} which are related to each other by asymptotic translations. The interplay between the curvature of q_{ab} and the extrinsic curvature (with respect to q_{ab}) of the 2-spheres of f_{ab} seems to be interesting even just from the viewpoint of differential geometry, without any reference to the physical problem of angular momentum. Indeed, *a priori*, using only general geometrical considerations, one would have expected that an Euclidean reduction would be possible only if q_{ab} approaches a flat metric as $o(1/r^2)$. It is somewhat surprising—and, from the physical viewpoint, fortunate—that the reduction is possible under a weaker condition which is "in between" the above stronger requirement which rules out 3-metrics with nonzero mass and the original ADM conditions which permit supertranslations.

From a direct physical viewpoint, however, conditions (5.2) and (5.13) are somewhat obscure. A large class of examples satisfying these conditions is provided by the metrics considered by O'Murchudha[22] and Chrusciel[23] in connection with supertranslation ambiguities. Fix any flat 3-metric f_{ab} outside a compact set of Σ and consider any metric q_{ab} whose components in a Cartesian chart of f_{ab} are given by

$$q_{\mathbf{ab}} = (1 + M(\theta,\varphi)/r) f_{\mathbf{ab}} - o(1/r)\,, \qquad (5.24)$$

where θ,φ and r are the spherical polar coordinates associated with the Cartesian chart. Then, it is easy to verify that q_{ac} satisfies (5.2) and the metric 2-spheres of $f_{\mathbf{ab}}$ satisfy (5.13). Thus if the 3-metric is "Schwartzschildean to the leading order," as pointed out in Refs. 22 and 23, the supertranslation ambiguities disappear. But the class of 3-metrics satisfying (5.2) is in fact larger and a characterization of this class along the lines of (5.24) is not yet available.[24]

Finally, note that it seems difficult to eliminate the supertranslation freedom associated with boosts if one works within 3 + 1 frameworks. In the spi framework, supertranslations correspond to *arbitrary* functions on the hyperboloid $\mathscr{D}$, and hence, in the physical space-time, there exist space-time supertranslations which induce identity transformations on the 3-manifold Σ. Obviously, such supertranslations cannot be eliminated by imposing stronger boundary conditions on Cauchy data on Σ. Thus, it is not possible to recover the Poincaré group in a 3 + 1 framework. One may, instead, try to recover just the Poincaré Lie algebra by considering only the infinitesimal boosts off Σ. But the treatment loses its simplicity because the boosts are treated differently from rotations and it seems difficult to interpret geometrically the conditions which arise in the construction.

ACKNOWLEDGMENTS

The need for a 3 + 1 decomposition of the spi-framework was stressed to A.A. by Jim York. A.A. also benefited from discussions with Piotr Chrusciel and Niall O'Murchadha. A.M. thanks Syracuse University for hospitality. A.A. was supported in part by the NSF contract PHY 8310041 with Syracuse University. A.M. was supported in part by crédits ministériels, tranche spéciale.

APPENDIX A: TRANSLATION OF ASYMPTOTIC CONDITIONS IN THE PHYSICAL SPACE LANGUAGE

Let (M, g_{ab}) satisfy Definition 1 with completion $(\hat{M}, \hat{g}_{ab})$. Let $\hat{\Sigma}$ be a spacelike submanifold of $\hat{M}$ passing through i° which is $C^{>1}$ at i° and C^∞ elsewhere. This means that the metric $\hat{q}_{ab}$ induced on $\hat{\Sigma}$ by $\hat{g}_{ab}$ is $C^{>0}$ at i° and that the extrinsic curvature $\hat{\pi}_{ab}$ admits a regular direction-dependent limit there. Consider any chart $\hat{x}^a \equiv (\hat{x}, \hat{y}, \hat{z})$ on $\hat{\Sigma}$, centered at i°, which is $C^{>1}$ at i° and C^∞ elsewhere, and in which the components $\hat{q}_{\hat{a}\hat{b}}$ of $\hat{q}_{ab}$ are given by $q_{\hat{a}\hat{b}} = \delta_{\hat{a}\hat{b}}$. Set

$\hat{r}^2 = \hat{x}^2 + \hat{y}^2 + \hat{z}^2$. Then using the l'Hopital's rule, one can show that $\lim \Omega/\hat{r}^2 = 1$ and that $\Omega/\hat{r}^2$ is $C^{>0}$ at i°. Hence, without loss of generality, we can choose $1/\hat{r}^2$ as the conformal factor. We make this choice and, from now on, use hatted quantities to refer to *this* conformal frame.

Let $\hat{f}_{ab}$ be the flat metric for which $\hat{X}^{\mathbf{a}}$ serves as a Cartesian chart. $\hat{f}_{ab}$ is, by construction, $C^{>0}$ at i° and equals $\hat{q}_{ab}$ at i°. Hence, the difference between $\hat{q}_{ab}$ and $\hat{f}_{ab}$ is of the form

$$\hat{q}_{ab} - \hat{f}_{ab} = \Omega^{1/2}\hat{d}_{ab}\,, \tag{A1}$$

where $\hat{d}_{ab}$ admits regular, direction-dependent limits at i°.

Let us now introduce a C^∞-chart $X^{\mathbf{a}}$ on $\Sigma = \hat{\Sigma} - i^\circ$ by $X^{\mathbf{a}} = \hat{X}^{\mathbf{a}}/\hat{r}^2$ and consider the flat metric f_{ab} on Σ for which $X^{\mathbf{a}}$ is a Cartesian chart.

f_{ab} is badly behaved at i° and is related to $\hat{f}_{ab}$ by $f_{ab} = (1/\hat{r}^4)\hat{f}_{ab} \equiv r^4\hat{f}_{ab}$, where $r^2 = X^2 + Y^2 + Z^2$. Next, let us compute the components of $\hat{q}_{ab}$ in the $X^{\mathbf{a}}$-chart. We have

$$\begin{aligned}\hat{q}_{\mathbf{ab}} &= \frac{\partial\hat{X}^c}{\partial X^a}\frac{\partial\hat{X}^d}{\partial X^b}(\hat{f}_{\hat{c}\hat{d}} + \Omega^{1/2}\hat{d}_{\hat{c}\hat{d}}) \\ &= \frac{1}{r^4}\delta_{ab} + \frac{1}{r^5}d_{\mathbf{ab}}\end{aligned} \tag{A2}$$

where $d_{\mathbf{ab}}$ are again *functions* which admit regular direction-dependent limits to i°. Hence, $q_{ab} = \Omega^{-2}\hat{q}_{ab} = r^4\hat{q}_{ab}$ has, in the $X^{\mathbf{a}}$-chart, the following components:

$$q_{\mathbf{ab}} = f_{\mathbf{ab}} + (1/r)d_{\mathbf{ab}}\,. \tag{A3}$$

Finally, the derivatives of the functions $d_{\mathbf{ab}}$ with respect to $X^{\mathbf{a}}$ are given by

$$\begin{aligned}\frac{\partial}{\partial X^{\mathbf{a}}}d_{\mathbf{bc}} &= \frac{\partial\hat{X}^m}{\partial X^{\mathbf{a}}}\frac{\partial}{\partial\hat{X}^{\mathbf{m}}}d_{\mathbf{bc}} = \frac{1}{r^2}\left(\delta^{\mathbf{m}}{}_{\mathbf{a}} - 2\frac{X^{\mathbf{m}}}{r^2}X_{\mathbf{a}}\right)\frac{\partial}{\partial\hat{X}^m}d_{\mathbf{bc}} \\ &= \frac{1}{r}(\delta^{\mathbf{m}}{}_{\mathbf{a}} - 2\eta^{\mathbf{m}}\eta_{\mathbf{a}})\left[\Omega^{1/2}\frac{\partial}{\partial\hat{X}^m}d_{\mathbf{bc}}\right].\end{aligned} \tag{A4}$$

Since the functions $d_{\mathbf{ab}}$ admit regular direction-dependent limits at i°, the quantity in square brackets also does so. Equations (A3) and (A4) imply that there exists a flat metric f_{ab} such that, in a Cartesian chart $X^{\mathbf{a}}$ defined by it, the components of $q_{ab} - f_{ab}$ are $O(1/r)$ and of $\partial_c q_{ab}$ are $O(1/r^2)$, where ∂ is the derivative operator of f_{ab}. Actually, one can continue in this manner to show that the components of $\partial_c\partial_d q_{ab}$ fall off as $1/r^3$, those of $\partial_m\partial_c\partial_d q_{ab}$ fall off as $1/r^4$, etc. [This is because the fact that d admits a regular direction dependent limit $\mathbf{d}$ implies that[3] $\lim(\Omega^{1/2}\hat{D}_{a_1})(\Omega^{1/2}\hat{D}_{a_2})\cdots(\Omega^{1/2}\hat{D}_{a_n})d$ exists and equals ${}^2\mathbf{D}_{a_1}{}^2\mathbf{D}_{a_2}\cdots{}^2\mathbf{D}_{a_n}\mathbf{d}$ for all n. One can, of course weaken this requirement by fixing the maximum permissible value of n if one so desires.]

Since $\hat{\Sigma}$ is $C^{>1}$, its extrinsic curvature, $\hat{\pi}_{ab}$, with respect to $\hat{g}_{ab}$ admits direction-dependent limits at i°. By converting the components of $\hat{\pi}_{ab}$ from hatted to the unhatted chart and using the relation between $\hat{\pi}_{ab}$ and π_{ab}, it follows that the components of π_{ab} in the $X^{\mathbf{a}}$ chart are $O(1/r^2)$.

APPENDIX B: EUCLIDEAN REDUCTION FROM THE SPI VIEW POINT

In this appendix we shall derive Eq. (5.13) (which led to Euclidean reduction) starting from the spi framework. The derivation shows that the Euclidean group obtained in Sec. V is the Euclidean subgroup (associated with $\hat{\Sigma}$) of the Poincaré group selected at i° by Eq. (2.5).

In the spi framework, one can remove the spatial supertranslational ambiguities associated with Σ by requiring,[19] on $\mathscr{D}$,

$$\gamma_a{}^m\gamma_b{}^n\mathbf{K}_{mn} = 0\,, \tag{B1}$$

where $\gamma_a{}^m$ is the natural metric on the cross section C of $\mathscr{D}$ singled out by Σ. Using the definition[2,3] of $\mathbf{K}_{mn}$ in terms of the Ricci tensor of $\hat{g}_{ab}$ and the fall-off condition on the matter stress energy from Definition 1, (B1) reduces to

$$\lim \Omega^{-1/2}\hat{\gamma}_a{}^m\hat{\gamma}_b{}^n[2\hat{\eta}^c\hat{\eta}^d(\hat{\nabla}_c\hat{\nabla}_d\Omega)\hat{\gamma}_{mn} - 2\hat{\nabla}_m\hat{\nabla}_n\Omega] = 0\,, \tag{B2}$$

where $\hat{\eta}^a = \hat{\nabla}^a\Omega^{1/2}$ and $\hat{\gamma}_{mn}$ is the metric induced on the 2-spheres $\Omega = \text{const}$ by $\hat{q}_{ab}$. This is a restriction on the choice of permissible conformal factors. Next, following Appendix A, we introduce on Σ a flat metric f_{ab} whose radial coordinate r equals $\Omega^{1/2}$. Thus, we now have a restriction on the flat metrics. To recast this condition in a convenient form, let us recast (B2) in the physical space language. For this, we have first to re-express the tensor on the left-hand side in terms of the unhatted differential structure (or, equivalently, re-express it in terms of its components in an unhatted, asymptotically Cartesian chart) and then re-express the connection $\hat{\nabla}$ in terms of ∇. Then, (B2) becomes

$$\begin{aligned}\lim_{r\to\infty} r^2\gamma_a{}^m\gamma_b{}^n(4\lambda^{-1}D_m\tilde{\eta}_n \\ - 4(\tilde{\eta}^c D_c\lambda^{-1})\gamma_{mn} - 4\lambda^{-2}r^{-1}\lambda^{-2}\gamma_{mn}) = 0\,,\end{aligned} \tag{B3}$$

where, as before, $\lambda^{-2} = q^{ab}D_a r D_b r$ and $\eta^a = \lambda q^{ab}D_b r$ is the unit normal to $\Gamma = \text{const}$ 2-spheres. Finally, using the definition of the extrinsic curvature, ${}^2\pi_{ab}$, with respect to q_{ab} of the $r = \text{const}$ 2-spheres, the fact that $({}^2\pi_{ab} - \gamma_{ab}/r) = O(1/r)$ and the fact that $\lambda = 1 + O(1/r)$, one can re-express (B3) as

$$\lim_{r\to\infty} r^2\left({}^2\pi_{ab} - \frac{1}{r}\gamma_{ab}\right) = 0\,. \tag{B4}$$

Thus, Eqs. (B1) and (B4) are *equivalent*. The former selects the Euclidean subgroup (associated with $\hat{\Sigma}$) of the Poincaré group at i°[19] while the latter led us, in Sec. V to select an Euclidean subgroup of the asymptotic symmetry group associated with the physical space Σ.

APPENDIX C: GENERAL SOLUTION TO $\mathring{\partial}_a P^{ab} = 0$

Fix a 2-sphere S of radius r in the Euclidean space (Σ, f_{ab}) and consider the equation $\mathring{\partial}_a\mathbf{p}^{ab} = 0$ on symmetric tensor fields $\mathbf{p}^{ab}$ on S, where $\mathring{\partial}_a$ is, as before, the derivative operator compatible with the metric $\mathring{\gamma}_{ab}$ induced on S by f_{ab}. Let ξ^a be any Killing field on $(S, \mathring{\gamma}_{ab})$. Then, $\mathring{\partial}_a(\mathbf{p}^{ab}\xi_b) = 0$. Thus, $V^a = \mathbf{p}^{ab}\xi_b$ is a divergence-free vector field on S, whence $\mathring{\epsilon}_{ab}V^b$ is curl-free. Since every 1-loop is contractable to zero on S, there exists a function h on S such that $\mathring{\epsilon}_{ab}V^b = D_a h$, whence $V^b = \mathring{\epsilon}^{mb}D_m h$. The function h is defined only up to a constant. Let us eliminate this freedom by fixing an arbitrary point p_0 on S and demanding that $h_{|_{p_0}} = 0$. (The final solution will be of course independent of the choice of

the point p_0.) Thus, every solution $\mathbf{P}^{ab}$ to the equation under consideration defines a mapping Ψ from the space of Killing vectors ξ^a on (S,γ_{ab}) to functions h on S. Recall, however, that every Killing field is completely characterized by its "Killing data" $(\xi^a, \eth_{[a}\xi_{b]} \equiv \xi \mathring{\epsilon}_{ab})$ evaluated at any point p of S.[25] Hence, given any point p in S, we have a linear mapping, $\Psi_p : (\xi^a, \xi)_{|_p} \to h_{|_p}$ which associates with every pair (ξ^a, ξ) consisting of a vector ξ^a and a number ξ at p, a number h at p. Hence, there exists a vector field ν_a and a function ν on S ("a dual Killing data") such that

$$\nu_a \xi^a + \xi\nu = h \tag{C1}$$

for all Killing field ξ^a on S, where ξ is the function on S given by $\eth_a \xi_b = \xi \mathring{\epsilon}_{ab}$. Recalling the definition of h, we have

$$\begin{aligned} P^{ab}\xi_b &= \mathring{\epsilon}^{ba}\eth_b h = \mathring{\epsilon}^{ba}\eth_b(\nu_m \xi^m + \xi\nu) \\ &= \mathring{\epsilon}^{ba}\left[(\eth_b \nu_m)\xi^m + \nu_m \eth_b \xi^m + \xi \eth_b \nu + \nu \eth_b \xi\right]. \end{aligned} \tag{C2}$$

Given a point p, we can always consider two Killing fields whose values at p yield linearly independent vectors at p and whose derivatives vanish at p. Using these Killing fields in Eq. (C2) and the fact that every Killing field ξ^a satisfies $\eth_a \eth_b \xi_c \equiv (\eth_a \xi)\mathring{\epsilon}_{bc} = \frac{1}{2}\,{}^2R \mathring{\epsilon}_{cb}\mathring{\epsilon}_{ad}\xi^d$, we obtain

$$P^{am} = \mathring{\epsilon}^{ba}(\eth_b \nu^m) - \tfrac{1}{2}\nu({}^2R)\mathring{\gamma}^{ma}, \tag{C3}$$

where ${}^2R$ is the scalar curvature of $(S,\mathring{\gamma}_{ab})$. Now, we can use the fact that P^{am} is divergence-free. Thus,

$$\begin{aligned} 0 = \eth_a P^{am} &= \mathring{\epsilon}^{ba}\eth_a \eth_b \nu^m - \tfrac{1}{2}({}^2R)\mathring{\gamma}^{ma}\eth_a \nu \\ &= \tfrac{1}{2}\mathring{\epsilon}^{ba}(\tfrac{1}{2}\,{}^2R \mathring{\epsilon}_{ab}\mathring{\epsilon}^m{}_n \nu^n) - \tfrac{1}{2}({}^2R)\eth^m \nu, \end{aligned} \tag{C4}$$

whence, we conclude

$$\nu_a = \mathring{\epsilon}_{am}\partial^m \nu; \tag{C5}$$

substituting for ν_a in (B3), we now have

$$\begin{aligned} P^{am} &= \mathring{\epsilon}^{ba}\eth_b(\mathring{\epsilon}^{mn}\eth_n \nu) - \tfrac{1}{2}({}^2R)\nu\mathring{\gamma}^{ma} \\ &= (\mathring{\gamma}^{bm}\mathring{\gamma}^{an} - \mathring{\gamma}^{bn}\mathring{\gamma}^{am})\eth_b \eth_n \nu - \tfrac{1}{2}({}^2R)\nu\mathring{\gamma}^{ma} \\ &= \eth^a \eth^m \nu - \mathring{\gamma}^{am}(\eth^2 \nu + \tfrac{1}{2}({}^2R)\nu). \end{aligned} \tag{C6}$$

Finally noting that the scalar curvature, ${}^2R$, of a 2-sphere of radius r is given by ${}^2R = 2/r^2$, we obtain the general solution to $\eth_a P^{am} = 0$:

$$P^{am} = \eth^a \eth^m \nu - \mathring{\gamma}^{am}(\eth^2 \nu + (1/r^2)\nu) \tag{C7}$$

for some function ν on S.

A completely analogous procedure shows that,[2,3] on the hyperboloid $\mathscr{D}$ the solution to the equation $\mathbf{D}_{[a}\mathbf{K}_{b]c} = 0$, $\mathbf{K}_{bc} = \mathbf{K}_{(bc)}$ is given by

$$\mathbf{K}_{ab} = \mathbf{D}_a \mathbf{D}_b \boldsymbol{\nu} + \boldsymbol{\nu} h_{ab} \tag{C8}$$

for some function $\boldsymbol{\nu}$ on $\mathscr{D}$. It is this result which enables the Poincaré reduction of the spi group in the case when $\mathbf{B}_{ab} = 0$.

[1] R. Arnowitt, S. Deser, and C. W. Misner, Phys. Rev. **117**, 1595 (1960); **118**, 1100 (1960); **121**, 1566 (1961); **122**, 997 (1961); the article in *Gravitation, an Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).

[2] A. Ashtekar and R. O. Hansen, J. Math. Phys. **19**, 1542 (1978).

[3] A. Ashtekar, in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980).

[4] P. S. Jang, J. Math. Phys. **17**, 141 (1976); E. Witten, Commun. Math. Phys. **80**, 381 (1981).

[5] Actually, it would suffice for our purposes here if the matter stress energy falls off as $1/r^{3+\epsilon}$ (for some $\epsilon > 0$) in the physical space-time.

[6] E. T. Newman, L. Tamburino, and T. Unti, J. Math. Phys. **4**, 915 (1963); A. Ashtekar and A. Sen, J. Math. Phys. **23**, 2168 (1982).

[7] A. Ashtekar and A. Magnon-Ashtekar, J. Math. Phys. **20**, 793 (1979).

[8] A. Ashtekar, J. Math. Phys. **22**, 2885 (1981).

[9] R. Penrose, Proc. R. Soc. London Ser. A **381**, 53 (1982).

[10] W. Shaw, Proc. R. Soc. London Ser. A **390**, 191 (1983).

[11] S. W. Hawking, Phys. Lett. B **126**, 175 (1983).

[12] Y. Choquet-Bruhat and D. Christodoulou, Acta Math. **146**, 129 (1981); D. Christodoulou and N. O'Murchadha, Commun. Math. Phys. **80**, 271 (1981).

[13] R. Geroch, J. Math. Phys. **13**, 956 (1972).

[14] The boldface indices denote components and take on numerical values 0, 1, 2, 3. Italic indices are "abstract indices" à la Penrose.

[15] It is not clear if there are not other types of ambiguities in the various expressions. In the spi framework, such ambiguities would yield inequivalent completions of the physical space-time, and, although the framework would hold for any one completion, the *relation* between quantities defined using, inequivalent completions may be complicated. For details, see A. Ashtekar, in *The Proceedings of the 10th International Conference on General Relativity and Gravitation*, edited by B. Bertotti *et al.* (Reidel, Dordrecht, The Netherlands, 1984); and in *The Proceedings of the Oregon Conference on Mass and Asymptotic Behavior of Space-time*, edited by F. Flaherty and J. Isenberg (Springer, Berlin, 1984).

[16] Note that our signature is $- + + +$.

[17] Note that one can also show that $\lim \int r\mathscr{R}_{ab}q^{ab}d^2V_q = 0$ and use this fact to obtain alternate expressions. This fact is also used in the passage from Eq. (3.4) to (3.12) and (3.13).

[18] That an additional assumption is involved in the derivation of (3.11) seems not to have been noticed before. This additional condition can always be satisfied if the additional boundary condition, discussed in Sec. V, for angular momentum to be well-defined is satisfied. (Also, our expression differs from Geroch's by a factor of two.)

[19] In view of Eq. (2.2), Eq. (5.1) is equivalent to the requirement that, the pullback, K_{ab} of $\mathbf{K}_{ab}$ to the 2-sphere cross section C of the hyperboloid $\mathscr{D}$ defined by $\hat{\Sigma}$ should satisfy ${}^2\mathbf{D}_{[a}K_{b]c} = 0$, where ${}^2\mathbf{D}$ is the intrinsic derivative operator on C. Hence, one can demand, in place of Eq. (2.5), $K_{ab} = 0$ on S. This condition removes the spatial-supertranslation ambiguity associated with Σ.

[20] r,θ,φ are a set of spherical polar coordinates adapted to f_{ab}. The 2-spheres under consideration are given by $r =$ constant.

[21] In Refs. 2 and 3, β_{ab} is defined to be the limit to i° of ${}^*\hat{C}_{ambn}(\hat{\nabla}^m\Omega^{1/2})(\hat{\nabla}^n\Omega^{1/2})$, i.e., ${}^*C_{ambn}(\nabla^m r)(\nabla^n r)$. But since $\mathbf{B}_{ab}$ vanishes and since η^a is proportional to $\nabla^a r$, with the proportionality factor approaching 1 as $O(1/r)$, one can replace $\nabla^a r$ by η^a.

[22] N. O'Murchadha (private communication).

[23] P. Chrusciel (private communication).

[24] Obvious generalizations of (5.24), e.g., $q_{ab}\,dX^a\,dX^b = (1 + M_1(\theta,\varphi)/r)dr^2 + r^2(1 + M_2(\theta,\varphi)/r)(d\theta^2 + \sin^2\theta\,d\varphi^2) + o(1/r)$ do *not* satisfy (5.2).

[25] See, e.g., A. Ashtekar and A. Magnon-Ashtekar, J. Math. Phys. **19**, 1567 (1978).

Finite energy electric monopoles in an extended theory of gravitation

G. Kunstatter
Department of Physics, University of Toronto, Toronto, Ontario, Canada M5S 1A7

(Received 5 January 1984; accepted for publication 6 April 1984)

We present a one-parameter family of extended Einstein–Maxwell Lagrangians in which an antisymmetric tensor field is nonlinearly coupled to both the gravitational and electromagnetic fields. We show that for arbitrary, positive values of the relevant parameter, the theory admits exact, static spherically symmetric solutions with everywhere finite electric field density and energy density. Asymptotically, the solutions are indistinguishable from the Reissner–Nordstrom solution in general relativity. In addition, we show that a corrected form of the exact solution in the nonsymmetric Kaluza–Klein theory presented in an earlier paper provides a special case of the family of solutions described above.

PACS numbers: 04.50. + h, 11.10.Ef

I. INTRODUCTION

Much work has been done recently on an extended theory of gravitation[1] in which an antisymmetric tensor field $g_{[\mu\nu]}$ (the "skewon" field) is coupled in a highly nontrivial way to the symmetric metric $g_{(\mu\nu)}$ of general relativity. In particular, the two fields are combined to form a new geometrical object: a sesquilinear, Hermitian fiber metric.[2,3] The predictions of the theory are consistent with all solar system data[4] and have interesting consequences for astrophysics[5] and cosmology.[6] In the following we present exact, spherically symmetric, static solutions to the field equations which result when the skewon and graviton fields are coupled to the electromagnetic field $F_{\mu\nu}$. The Lagrangian is similar to the one derived by Kalinowski using a Kaluza–Klein approach.[7] It is different, however, due to the absence of the polarization tensor $H_{\mu\nu}$, which plays an important role in Kalinowski's theory, and it also differs due to the presence of a free parameter λ, which is fixed to be 2 in the Kaluza–Klein version. It is important to note that the interesting properties of our solution do not depend on the particular choice of parameters which result in the Kaluza–Klein approach. They depend only on the presence of a nonlinear coupling between the antisymmetric tensor field $g_{[\mu\nu]}$, and the electromagnetic field.

In Sec. II, we write down the Lagrangian of interest and derive the field equations. The static spherically symmetric solution is found in Sec. III, and its properties discussed. In Sec. IV, we derive a corrected version of the field equations from the nonsymmetric Kaluza–Klein Lagrangian of Kalinowski and show that in the static, spherically symmetric case they reduce to the equations of Sec. III with a particular choice of parameter λ. This section also corrects an earlier paper,[8] in which incorrect field equations were used to obtain a solution. As we shall see, most of the conclusions of that paper remain valid. The exact form of the solution, however, does change. Finally, we summarize our results in Sec. V and present some conclusions.

II. THE LAGRANGIAN AND FIELD EQUATIONS

The extended theory of gravitation we will consider is Moffat's nonsymmetric gravitation theory (NGT). (For a review see Ref. 1.) The theory is based on a Hermitian, hypercomplex-valued fundamental form $g_{\mu\nu} = g^R_{(\mu\nu)} + Jg^I_{[\mu\nu]}$, where $J^2 = 1$[2,3]. The gravitational Lagrangian is constructed from the fiber metric $g_{\mu\nu}$ and a generalized Ricci tensor

$$R_{\mu\nu}(W) = W^{\rho}_{\mu\nu,\rho} - \tfrac{1}{2}(W^{\rho}_{\mu\rho,\nu} + W^{\rho}_{\nu\rho,\mu}) - W^{\rho}_{\mu\lambda}W^{\lambda}_{\nu\rho} + W^{\lambda}_{\rho\lambda}W^{\rho}_{\mu\nu}, \tag{2.1}$$

where $W^{\lambda}_{\mu\nu}$ is a hypercomplex-valued, metrically compatible connection such that

$$W^{\lambda}_{\mu\nu} = \Gamma^{\lambda}_{\mu\nu}(g) - \tfrac{2}{3}\delta^{\lambda}_{\mu}W_{\nu}. \tag{2.2}$$

In Eq. (2.2) $W_{\nu} \equiv W^{\lambda}_{[\nu\lambda]} = \frac{1}{2}(W^{\lambda}_{\nu\lambda} - W^{\lambda}_{\lambda\nu})$, and $\Gamma^{\lambda}_{\mu\nu}(g)$ is a hypercomplex-Hermitian, metrical connection determined uniquely from the metric by the following equation:

$$g_{\mu\nu,\lambda} - g_{\epsilon\nu}\Gamma^{\epsilon}_{\mu\lambda} - g_{\mu\epsilon}\Gamma^{\epsilon}_{\lambda\nu} = 0. \tag{2.3}$$

Note that Eq. (2.2) constrains $\Gamma^{\lambda}_{[\mu\lambda]}$ to be zero.

The full Lagrangian we wish to consider, including the electromagnetic field $F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}$, is (in units $G = c = 1$)

$$L = \mathscr{g}^{\mu\nu}R_{\mu\nu}(\Gamma) + \tfrac{2}{3}\mathscr{g}^{[\mu\nu]}W_{[\mu,\nu]} + 4\pi\sqrt{-g}(F^{\mu\nu}F_{\mu\nu} - \lambda(g^{[\mu\nu]}F_{\mu\nu})^2); \tag{2.4}$$

where $F^{\mu\nu} \equiv g^{\alpha\mu}g^{\beta\nu}F_{\alpha\beta}$, $g^{\mu\nu}$ is defined by $g^{\mu\nu}g_{\mu\rho} = \delta^{\nu}_{\rho}$, and $\mathscr{g}^{\mu\nu} \equiv \sqrt{-g}g^{\mu\nu}$. In addition, we have used the fact that

$$R_{\mu\nu}(W) = R_{\mu\nu}(\Gamma) + \tfrac{2}{3}W_{[\mu,\nu]}, \tag{2.5}$$

which follows from Eq. (2.2). Since we are using a second-order formalism[9] for $g_{\mu\nu}$, the fundamental fields we must vary are $g_{\mu\nu}$, W_{μ}, and A_{μ}. The resulting field equations are, respectively,

$$R_{\mu\nu}(W) - \tfrac{1}{2}g_{\mu\nu}R(W) = 8\pi T_{\mu\nu}, \tag{2.6}$$

$$\mathscr{g}^{[\mu\nu]}{}_{,\nu} = 0, \tag{2.7}$$

$$(\tfrac{1}{2}(\mathscr{g}^{\alpha\mu}g^{\beta\nu} + \mathscr{g}^{\mu\alpha}g^{\nu\beta})F_{\alpha\beta})_{,\nu} = \lambda(\mathscr{g}^{[\mu\nu]}g^{[\alpha\beta]}F_{\alpha\beta})_{,\nu}, \tag{2.8}$$

where

$$T_{\mu\nu} = -(g^{\alpha\beta}F_{\alpha\mu}F_{\beta\nu} - \lambda g^{[\alpha\beta]}F_{\alpha\beta}F_{\mu\nu}) + \tfrac{1}{4}g_{\mu\nu}(F^{\alpha\beta}F_{\alpha\beta} - \lambda(g^{[\alpha\beta]}F_{\alpha\beta})^2). \tag{2.9}$$

 0022-2488/84/092691-05$02.50

Here $T_{\mu\nu}$ reduces to the ordinary Maxwell stress energy tensor when $g_{[\mu\nu]}=0$. In general, it is nonsymmetric but Hermitian, so that $T_{\mu\nu}=T^{R}_{(\mu\nu)}+JT^{I}_{[\mu\nu]}$. In addition the trace of $T_{\mu\nu}$ with respect to the full nonsymmetric vanishes identically: $g^{\mu\nu}T_{\mu\nu}=0$. Another interesting feature of the above field equations is the induced current density

$$J^{\mu}_{\text{ind}}\equiv\lambda\left(\sqrt{-g}g^{[\mu\nu]}g^{[\alpha\beta]}F_{\alpha\beta}\right)_{,\nu}, \tag{2.10}$$

which is identically conserved. This induced current density plays an important role in the nonsingular behavior of the solution to be considered. Note that this induced current vanishes to linearized order because the skewon–photon coupling in the Lagrangian is quadratic in both $g^{[\mu\nu]}$ and $F_{[\mu\nu]}$ (see Ref. 10). Finally we remark that the somewhat complicated form of the left-hand side of Eq. (2.8) is due to the fact that in general $(\sqrt{-g}F^{\mu\nu})_{,\nu}$ is not real (it may have hypercomplex-valued components). Naturally, the physical electromagnetic current must be real:

$$\begin{aligned}J^{\mu}_{\text{em}}&=\text{Re}\left[\left(\sqrt{-g}F^{\mu\nu}\right)_{,\nu}\right]\\&=\left[\left(\mathcal{g}^{(\alpha\mu)}g^{[\beta\nu]}+\mathcal{g}^{[\alpha\mu]}g^{(\beta\nu)}\right)F_{\alpha\beta}\right]_{,\nu}\\&=\tfrac{1}{2}\left[\left(\mathcal{g}^{\alpha\mu}g^{\beta\nu}+\mathcal{g}^{\mu\alpha}g^{\nu\beta}\right)F_{\alpha\beta}\right]_{,\nu}\end{aligned} \tag{2.11}$$

This is precisely the form which appears in Eq. (2.8), and is a direct consequence of the fact that the electromagnetic Lagrangian is real by construction.

III. STATIC SPHERICALLY SYMMETRIC SOLUTION

The most general static spherically symmetric form of the metric $g_{\mu\nu}$ is[11]

$$g_{\mu\nu}=\begin{bmatrix}-\alpha(r) & 0 & 0 & \omega(r)\\ 0 & -\beta(r) & f(r)\sin\theta & 0\\ 0 & -f(r)\sin\theta & -\beta(r)\sin^2\theta & 0\\ -\omega(r) & 0 & 0 & \gamma(r)\end{bmatrix}, \tag{3.1}$$

and the determinant $\sqrt{-g}$ is

$$\sqrt{-g}=(\alpha\gamma-\omega^2)^{1/2}(\beta^2+f^2)^{1/2}\sin\theta. \tag{3.2}$$

For simplicity, we will consider only the case $f=0$. The implications of this restriction are discussed at the end. Furthermore, we will assume that the magnetic field $B(r)=F_{23}=0$ so that the only nonzero components of $F_{\mu\nu}$ are $F_{14}=-F_{41}=E(r)$. By solving Eq. (2.8) we find

$$E(r)=Q\beta(\alpha\gamma-\omega^2)^{1/2}/(\beta^2+2\lambda l^4), \tag{3.3}$$

where Q is a constant of integration. Eq. (2.7) can also be readily solved to yield

$$\omega^2/(\alpha\gamma-\omega^2)=l^4/\beta^2, \tag{3.4}$$

where l^4 is a new constant of integration which acts as a source for the skewon field $g_{[\mu\nu]}$. The sign of l^4 is determined to be positive by the hypercomplex Hermiticity of the metric. The choice $l^4<0$ yields the complex theory ($J^2=-1$). (See Ref. 2.) With the aid of Eqs. (3.3) and (3.4), the stress energy tensor $T_{\mu\nu}$ can be put in the following simple form:

$$T_{\mu\nu}=\frac{1}{2}\frac{Q^2}{(r^4+b^4)}\begin{bmatrix}-\alpha[(r^4-b^4)/(r^4+b^4)] & 0 & 0 & \omega\{[r^4(1-4\lambda)-b^4]/(r^4+b^4)\}\\ 0 & r^2 & 0 & 0\\ 0 & 0 & r^2\sin^2\theta & 0\\ -\omega\{[r^4(1-4\lambda)-b^4]/(r^4+b^4)\} & 0 & 0 & \gamma[(r^4-b^4)/(r^4+b^4)]\end{bmatrix}, \tag{3.5}$$

where we have defined $b^4=2\lambda l^4$ and we have made the standard coordinate choice $r^2=\beta$.

As a consistency check, one can verify that the $T_{\mu\nu}$ given above satisfied the conservation laws of the theory. These conservation laws follow from the fact that $R_{\mu\nu}(\Gamma)$ and $R_{\mu\nu}(W)$ satisfy a set of generalized Bianchi identities[12] which require $T_{\mu\nu}$ to satisfy

$$\left(\mathcal{g}^{\mu\alpha}T_{\mu\rho}+\mathcal{g}^{\alpha\mu}T_{\rho\mu}\right)_{,\alpha}+\sqrt{-g}g^{\alpha\beta}{}_{,\rho}T_{\alpha\beta}=0. \tag{3.6}$$

In the static spherically symmetric case we are considering, it can be shown that the only nontrivial equation in (3.6) takes the simple form

$$\left(\mathcal{g}^{11}T_{11}\right)_{,1}-\tfrac{1}{2}\mathcal{g}^{(\alpha\beta)}T_{(\alpha\beta),1}=0. \tag{3.7}$$

It is interesting to note that only the symmetric part of $T_{\mu\nu}$ plays a role. If the expressions for $T_{\mu\nu}$ given by (3.5) are substituted into the left-hand side of (3.7), the result is identically zero as expected.

We now proceed to solve for α,γ, and ω, following Pant[13] and Kalinowski and Kunstatter.[8] The only nontrivial field equations which emerge from Eq. (2.6) in the special case we are considering are (see Refs. 8 and 13 for the explicit forms of $R_{\mu\nu}$):

$$R_{11}(\Gamma)=8\pi T_{11}, \tag{3.8}$$

$$R_{22}(\Gamma)=8\pi T_{22}, \tag{3.9}$$

$$R_{44}(\Gamma)=8\pi T_{44}. \tag{3.10}$$

Moreover, the generalized Bianchi identities imply that only two of the above are independent. Taking the linear combination

$$(1/\alpha)R_{11}+(1/\gamma)R_{44}=0, \tag{3.11}$$

yields

$$\frac{(\alpha\gamma)'}{(\alpha\gamma)}=\frac{-4}{r}\left(\frac{l^4}{l^4+r^4}\right), \tag{3.12}$$

which implies that

$$(\alpha\gamma)=N(1+l^4/r^4). \tag{3.13}$$

The constant of integration N in Eq. (3.13) must be set to unity in order to recover the Reissner–Nordstrom solution in the large r limit. Substituting Eq. (3.12) into Eq. (3.9) gives

$$\frac{\partial}{\partial r}(r\alpha^{-1})=1-4\pi\frac{Q^2r^2}{(r^4+b^4)}, \tag{3.14}$$

which can readily be integrated to give

$$\alpha^{-1}(r)=(1-2M/r+4\pi Q^2f(r)/r), \tag{3.15}$$

where

$$f(r)=\frac{1}{4\sqrt{2}b}\log\left|\frac{r^2-\sqrt{2}rb+b^2}{r^2+\sqrt{2}rb+b^2}\right| + \frac{1}{2\sqrt{2}b}\left(\tan^{-1}\left(\frac{2r+\sqrt{2}b}{\sqrt{2}b}\right) + \tan^{-1}\left(\frac{2r-\sqrt{2}b}{\sqrt{2}b}\right)\right). \tag{3.16}$$

In Eq. (3.15), M is the usual constant of integration which, in general relativity, provides the source for the asymptotic gravitational field. This completes the solution, since the expressions for $g_{44}=\gamma$ and $g_{[14]}=\omega$ follow from Eqs. (3.4) and (3.13). It is straightforward to verify that the third, as yet unused equation

$$\frac{1}{\alpha}R_{11}-\frac{1}{\gamma}R_{44}=\frac{8\pi}{\alpha}T_{11}-\frac{8\pi}{\gamma}T_{44}$$

is identically satisfied. Moreover, when $\lambda=0$ the above solution reduces to the one found by Moffat.[14]

We now examine in some detail the remarkable properties of this solution. First, we note that Eqs. (3.3), (3.4), and (3.13) yield the following expression for the electric field:

$$E(r)=Qr^2/(r^4+b^4). \tag{3.17}$$

Here $E(r)$ is clearly nonsingular everywhere and is zero at $r=0$. Moreover, $E(r)\to Q/r^2$ as $r\to\infty$, so that it is asymptotically a Coulomb field. The effective charge density, as defined by Eq. (2.11), is

$$4\pi\sqrt{-g}\,\rho(r)\equiv J^4_{\rm em}=\sin\theta\,[4Qr^3b^4/(r^4+b^4)^2], \tag{3.18}$$

which yields the following conserved charge:

$$\int_0^\infty\sqrt{-g}\,\rho(r)dr\,d\theta\,d\phi=Q. \tag{3.19}$$

Thus, the total electric charge Q is precisely equal to the spatial integral of the effective charge distribution.

Next let us examine the properties of the function $f(r)$, as defined by Eq. (3.16). In particular, we note that

$$\lim_{r\to0}f(r)=0. \tag{3.20}$$

Moreover, in the limit that $r\to\infty$,

$$f(r)\to\frac{\pi}{2\sqrt{2}b}-\frac{1}{r}+O\left(\frac{1}{r^2}\right). \tag{3.21}$$

We now find the remarkable result that by setting the source M of the singular gravitational field to zero, we have a solution for $\alpha^{-1}(r)$ of the following form:

$$\alpha^{-1}(r)=(1-4\pi Q^2f(r)/r), \tag{3.22}$$

such that

$$\lim_{r\to0}\alpha^{-1}(r)=1, \tag{3.23}$$

and

$$\lim_{r\to\infty}\alpha^{-1}(r)=1-\frac{2M_Q}{r}+\frac{4\pi Q^2}{r^2}+O\left(\frac{1}{r^3}\right), \tag{3.24}$$

where

$$M_Q\equiv\frac{4\pi Q^2f(\infty)}{2}=\frac{\pi^2Q^2}{\sqrt{2}b}.$$

Thus $\alpha^{-1}(r)$ asymptotically approaches the Reissner–Nordstrom solution with charge Q and mass M_Q which is a fixed function of Q, l, and λ. It is interesting to note that the positivity of the mass is not imposed by hand, but follows from the fact that the constant of integration M in Eq. (3.15) was fixed to be zero so that $\alpha^{-1}(r)$ would be nonsingular.

We now calculate the conserved stress energy density associated with the matter Lagrangian $T_{\mu\nu}$:

$$\sqrt{-g}\,T^4_4\equiv\tfrac{1}{2}(\sqrt{-g}\,g^{4\mu}T_{4\mu}+\sqrt{-g}\,g^{\mu4}T_{\mu4}) =\tfrac{1}{2}Q^2r^2\sin\theta/(r^4+b^4). \tag{3.25}$$

This is again nonsingular everywhere and approaches the Einstein–Maxwell value $(\frac{1}{2}Q^2\sin\theta/r^2)$ as $r\to\infty$. Furthermore, the integral of the energy density is

$$\int\sqrt{-g}\;T^4{}_4\,dr\,d\theta\,d\phi=\frac{4\pi Q^2}{2}f(\infty) =M_Q. \tag{3.26}$$

The Newtonian mass M_Q seen at infinity is precisely equal to the total electric field energy.

Although the solutions presented above have many features in common with Wheeler's geons[15] and with the solutions in the nonlinear Born–Infeld electrodynamics,[16] they are not completely nonsingular. In particular, we have from Eqs. (3.4), (3.13), and (3.22):

$$\omega(r)=Jl^2/r^2, \tag{3.27}$$

and

$$\gamma(r)=(1-4\pi Q^2[f(r)/r])(1+l^4/r^4), \tag{3.28}$$

so that at $r=0$, $g_{[14]}$ and g_{44} have l^2/r^2 and l^4/r^4 singularities, respectively. Nonetheless it is remarkable that the electric field and stress energy density are finite everywhere. Moreover, recall that for simplicity we have set $g_{[23]}=f(r)=0$. When this restriction is relaxed, f appears everywhere in the field equations in the linear combination (β^2+f^2). It is therefore likely that completely nonsingular solutions do exist in which the skewon singularity (l^2/r^2) is replaced by an expression of the form $l^2/(r^4+f^2)$. These solutions are currently being investigated.[17]

IV. THE NONSYMMETRIC KALUZA–KLEIN THEORY

Although the Kaluza–Klein extension of NGT was first considered by Moffat,[18] the Lagrangian we will consider is due to Kalinowski.[7] In particular we wish to examine the relationship between the solutions presented above and analogous solutions[8] in the nonsymmetric Kaluza–Klein theory. Up to a factor of (-4π) multiplying the matter terms, Kalinowski's Lagrangian is[7]

$$L_K=\mathcal{G}^{\mu\nu}R_{\mu\nu}(W)+4\pi\sqrt{-g}\,(H^{\mu\nu}F_{\mu\nu}-2(g^{[\mu\nu]}F_{\mu\nu})^2), \tag{4.1}$$

where $H^{\mu\nu}\equiv g^{\beta\mu}g^{\alpha\nu}H_{\beta\alpha}$ and $H_{\mu\nu}$ is assumed to be an antisymmetric "polarization" tensor given in terms of $g_{\mu\nu}$ and $F_{\mu\nu}$ by

$$g_{\delta\beta}g^{\gamma\delta}H_{\gamma\alpha}+g_{\alpha\delta}g^{\delta\gamma}H_{\beta\gamma}=2g_{\alpha\delta}g^{\delta\gamma}F_{\beta\gamma}. \tag{4.2}$$

Note that $F_{\mu\nu}=H_{\mu\nu}$ when $g_{[\mu\nu]}=0$.

In order to derive the field equations we must vary the

Lagrangian with respect to the independent fields $g^{\mu\nu}$, W_μ, and A_μ. The variation of $H_{\mu\nu}$ is implicitly determined from $\delta g^{\mu\nu}$ and δA_μ by Eq. (4.2). The field equations which result from this variation (see Appendix) are

$$R_{\mu\nu}(W) - \tfrac{1}{2} g_{\mu\nu} R(W) = 8\pi T^K_{\mu\nu}, \tag{4.3}$$

$$\mathscr{g}^{[\mu\nu]}{}_{,\nu} = 0, \tag{4.4}$$

$$(\sqrt{-g}\, H^{\mu\nu})_{,\nu} = 2(\mathscr{g}^{[\mu\nu]} g^{[\alpha\beta]} F_{\alpha\beta})_{,\nu}, \tag{4.5}$$

where

$$T^K_{\mu\nu} \equiv -(g^{\alpha\beta} H_{\beta\nu} H_{\alpha\mu} - 2g^{[\alpha\beta]} F_{\alpha\beta} F_{\mu\nu}) + \tfrac{1}{4} g_{\mu\nu}(H^{\alpha\beta} H_{\alpha\beta} - 2g^{[\alpha\beta]} F_{\alpha\beta}) + \tfrac{1}{2} J_{\mu\nu}, \tag{4.6}$$

and $J_{\mu\nu}$ is the contribution to $T^K_{\mu\nu}$ from $(\delta H_{\mu\nu}/\delta g^{\mu\nu})$:

$$J_{\mu\nu} = 4H_{\alpha\mu} H_{\beta\nu} g^{[\alpha\beta]} - 4H_{\alpha\mu} H_{\tau\epsilon} g^{\tau\alpha} g_{\beta\nu} g^{[\epsilon\beta]}. \tag{4.7}$$

Note that $J_{\mu\nu} = 0$ when $g_{[\mu\nu]} = 0$, as expected. Moreover, $g^{\mu\nu} J_{\mu\nu} = 0$ so that $T^K_{\mu\nu}$ is traceless. In general, however, $J_{\mu\nu}$ does not appear to be zero and may contribute to the field equations. Note that Eqs. (4.3)–(4.7) differ from those in Ref. 7 and subsequent references by a factor of 2 on the right-hand side of Eq. (4.5) and the inclusion of $J_{\mu\nu}$ in Eq. (4.6). Presumably the variation of $H_{\mu\nu}$ was not properly taken into account in those references.

In the static spherically symmetric case with only electric field present, it is straightforward to show that $J_{\mu\nu} = 0$, and $H^{\mu\nu} = F^{\mu\nu}$. Thus Eqs. (4.3)–(4.6) reduce precisely to Eqs. (2.6)–(2.9) but with $\lambda = 2$. The only aspect of the corrected field equations which is relevant to the solution quoted in Ref. 8, is the extra factor of 2. When this correction is taken into account, the solution is seen to be a special case of the family of solutions given in Sec. III above, with $\lambda = 2$. Most of the interesting features quoted in Ref. 8 are therefore correct. The only significant modification is that with the corrected field equations, the Newtonian mass is exactly equal to the integral of $\sqrt{-g}\, T^4{}_4$. In Ref. 8 these quantities incorrectly differed by a factor of $\frac{14}{39}$.

V. CONCLUSIONS

We have shown that a one-parameter family of extended gravitational Lagrangians containing nonlinear skewon–photon couplings admits exact static, spherically symmetric solutions with remarkable geon-like properties. As long as the parameter λ is positive, the electric field, charge density, and energy density in these solutions are all nonsingular. Moreover, the solutions behave exactly like the Reissner–Nordstrom solutions at spatial infinity, with charge and mass precisely equal to the integrated charge and stress-energy densities, respectively. In this sense the solutions describe "mass without mass" and "charge without charge."

The particular choice of photon–skewon coupling was inspired by the nonsymmetric Kaluza–Klein Lagrangian of Kalinowski,[7] although the present Lagrangian is somewhat more general, due to the arbitrariness of the parameter λ. It is interesting to note that the nonsingular properties of our solutions do not depend on the particular value of λ, only on its sign. Although it is a triumph of the Kaluza–Klein ansatz that it uniquely predicts such a term in the Lagrangian, the form of the coupling could also have been derived from more general arguments; namely requiring coordinate invariance, second-order differential equations, and a real Lagrangian. In fact, one can also add another term of the form $4\pi\sqrt{-g}\, g^{[\mu\alpha]} g^{[\nu\beta]} F_{\mu\nu} F_{\alpha\beta}$, which also obeys all of the above criteria. The effects of such a term are currently under investigation.

Much work remains to be done before the physical significance of these solutions is known. It is important to discover whether completely nonsingular solutions exist, and whether they are classically stable. This would have important consequences for the possible existence of quantum particles based on these soliton-like solutions. It would also be interesting to discover whether the nonsingular aspects of the solution are a manifestation of topological properties of the full nonlinear theory. Finally, the consequences of the photon–skewon coupling for photon propagation should be investigated in order to place experimental bounds on the parameters in the theory. These questions will be addressed in future work.

ACKNOWLEDGMENTS

I would like to thank Dan Evens, Martin Green, Rob Mann, John Moffat, and Ted Treml for helpful discussions.

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

APPENDIX: KALUZA–KLEIN FIELD EQUATIONS

We derive the field equations which result from varying L_K in Eq. (4.1) with respect to $g^{\mu\nu}$ and A_μ. The "matter" Lagrangian of interest is

$$L_M \equiv 4\pi\sqrt{-g}(H^{\mu\nu} F_{\mu\nu} - 2(g^{[\mu\nu]} F_{\mu\nu})^2), \tag{A1}$$

where $H^{\mu\nu} \equiv g^{\alpha\mu} g^{\beta\nu} H_{\alpha\beta}$ is given implicitly by Eq. (4.2). Contracting Eq. (4.2) with $g^{\alpha\rho} g_{\rho\mu}$ yields

$$2F_{\beta\mu} = H_{\beta\mu} + K_{\beta\mu}{}^{\gamma\alpha} H_{\gamma\alpha}, \tag{A2}$$

where we have defined

$$K_{\beta\mu}{}^{\gamma\alpha} \equiv g_{\delta\beta} g_{\rho\mu} g^{\gamma\delta} g^{\alpha\rho}. \tag{A3}$$

Note that $K_{\beta\mu}{}^{\gamma\alpha} = \delta_\beta{}^\gamma \delta_\mu{}^\alpha$ when $g^{[\mu\nu]} = 0$. Next we multiply Eq. (A.3) by $H^{\beta\mu}$ to find that

$$F_{\beta\mu} H^{\beta\mu} = H_{\beta\mu} H^{\beta\mu}, \tag{A4}$$

so that

$$\delta L_M = 4\pi\delta[\sqrt{-g}\,](H^{\mu\nu} H_{\mu\nu} - 2(g^{[\mu\nu]} F_{\mu\nu})^2) + 4\pi\sqrt{-g}\,\delta(H^{\mu\nu} H_{\mu\nu}) - 8\pi\sqrt{-g}\,\delta(g^{[\alpha\beta]} F_{\alpha\beta})^2. \tag{A5}$$

We now take the variation of Eq. (A2) and multiply by $H^{\beta\mu}$:

$$2H^{\beta\mu}\delta F_{\beta\mu} = H^{\beta\mu}\delta H_{\beta\mu} + H^{\beta\mu} K_{\beta\mu}{}^{\gamma\alpha}\delta H_{\gamma\alpha} + \delta K_{\beta\mu}{}^{\gamma\alpha} H^{\beta\mu} H_{\gamma\alpha}. \tag{A6}$$

The key to solving the variational problem is to note that

$$
\begin{aligned}
&H^{\beta\mu}\delta H_{\beta\mu} + H^{\beta\mu}K_{\beta\mu}{}^{\gamma\alpha}\delta H_{\gamma\alpha} \\
&\quad = H^{\beta\mu}\delta H_{\beta\mu} + H_{\epsilon\tau}g^{\epsilon\beta}g^{\tau\mu}K_{\beta\mu}{}^{\gamma\delta}\delta H_{\gamma\alpha} \\
&\quad = H_{\epsilon\tau}g^{\epsilon\beta}g^{\tau\mu}\delta H_{\beta\mu} + H_{\beta\mu}g^{\epsilon\beta}g^{\tau\mu}\delta H_{\epsilon\tau} \\
&\quad = \delta(H^{\beta\mu}H_{\beta\mu}) - H_{\beta\mu}H_{\epsilon\tau}\delta(g^{\epsilon\beta}g^{\tau\mu}). \qquad \text{(A7)}
\end{aligned}
$$

Using Eqs. (A6) and (A7), δL_M reduces to

$$
\begin{aligned}
\delta L_M = &-2\pi \not{g}_{\mu\nu}\delta g^{\mu\nu}(H^{\alpha\beta}H_{\alpha\beta} - 2(g^{[\alpha\beta]}F_{\alpha\beta})^2) \\
&+ 8\pi\sqrt{-g}\, H^{\beta\mu}\delta F_{\beta\mu} + 8\pi\sqrt{-g}\, H_{\beta\mu}H_{\epsilon\tau}g^{\epsilon\beta}\delta g^{\tau\mu} \\
&- 4\pi\sqrt{-g}H^{\beta\mu}H_{\gamma\alpha}\delta K_{\beta\mu}{}^{\gamma\alpha} \\
&- 8\pi\sqrt{-g}\delta(g^{[\mu\nu]}F_{\mu\nu})^2, \qquad \text{(A8)}
\end{aligned}
$$

where we have also used the standard relation $\delta\sqrt{-g} = -\frac{1}{2}g_{\mu\nu}\delta g^{\mu\nu}$.

The only remaining task is to calculate

$$
\begin{aligned}
&H^{\beta\mu}H_{\gamma\alpha}\delta K_{\beta\mu}{}^{\gamma\alpha} \\
&\quad = H_{\epsilon\tau}H_{\gamma\alpha}\, g^{\epsilon\beta}g^{\tau\mu}(\delta g_{\delta\beta}\, g_{\rho\mu}\, g^{\gamma\delta}g^{\alpha\rho} + g_{\delta\beta}\delta g_{\rho\mu}\, g^{\gamma\delta}g^{\alpha\rho} \\
&\qquad + g_{\delta\beta}\, g_{\rho\mu}\delta g^{\gamma\delta}g^{\alpha\rho} + g_{\delta\beta}\, g_{\rho\mu}\, g^{\gamma\delta}\delta g^{\alpha\rho}) \\
&\quad = 2H_{\epsilon\tau}H_{\gamma\alpha}\, g^{\alpha\tau}\delta g^{\gamma\epsilon} - 2H_{\epsilon\tau}H_{\gamma\alpha}\, g^{\alpha\tau}\, g^{\gamma\delta}g_{\delta\beta}\delta g^{\epsilon\beta}. \qquad \text{(A9)}
\end{aligned}
$$

In Eq. (A9) the identity $\delta(g^{\mu\nu}g_{\mu\rho}) = 0 = g^{\mu\nu}\delta g_{\mu\rho} + \delta g^{\mu\nu}g_{\mu\rho}$ was used. By rearranging dummy indices and using $g^{\gamma\delta} = g^{\delta\gamma} - 2g^{[\delta\gamma]}$ we find that

$$
\begin{aligned}
H^{\beta\mu}H_{\gamma\alpha}\delta K_{\beta\mu}{}^{\gamma\alpha} &= 4H_{\mu\alpha}H_{\nu\tau}g^{[\alpha\tau]}\delta g^{\mu\nu} \\
&\quad - 4H_{\mu\tau}H_{\gamma\alpha}g^{\alpha\tau}g_{\delta\nu}\, g^{[\gamma\delta]}\delta g^{\mu\nu} \\
&= J_{\mu\nu}\delta g^{\mu\nu} \qquad \text{(A10)}
\end{aligned}
$$

Finally we have

$$
\begin{aligned}
\delta L_M &= \{8\pi\sqrt{-g}H_{\beta\nu}H_{\alpha\mu}\, g^{\alpha\beta} - 16\pi\sqrt{-g}(g^{[\alpha\beta]}F_{\alpha\beta})F_{\mu\nu} \\
&\quad - 2\pi\sqrt{-g}g_{\mu\nu}(H^{\alpha\beta}H_{\alpha\beta} - 2(g^{[\alpha\beta]}F_{\alpha\beta})^2) \\
&\quad - 4\pi J_{\mu\nu}\}\delta g^{\mu\nu} + \{8\pi\sqrt{-g}H^{\mu\nu} \\
&\quad - 16\pi\sqrt{-g}g^{[\alpha\beta]}F_{\alpha\beta}g^{[\mu\nu]}\}\delta F_{\mu\nu}, \qquad \text{(A11)}
\end{aligned}
$$

[1] J. W. Moffat, *Proceedings of the VII International School of Gravitation and Cosmology*, Erice, Sicily, edited by V. de Sabbata (World Scientific, Singapore, 1982).
[2] G. Kunstatter and R. Yates, J. Phys. A **14**, 847 (1981).
[3] G. Kunstatter, J. W. Moffat, and J. Malzan, J. Math. Phys. **24**, 886 (1983).
[4] L. Campbell, J. C. McDow, J. W. Moffat, and D. Vincent, Nature **305**, 508 (1983); J. W. Moffat, Phys. Rev. Lett. **50**, 709 (1983); J. C. McDow and J. W. Moffat, Can. J. Phys. **60**, 1545 (1982); J. C. McDow and J. W. Moffat, Can. J. Phys. **60**, 1556 (1982).
[5] J. C. McDow, Ph.D. thesis, University of Toronto, 1983 (unpublished).
[6] D. E. Vincent, Ph.D. thesis, University of Toronto, 1983 (unpublished).
[7] M. W. Kalinowski, J. Math. Phys. **24**, 1835 (1983).
[8] M. W. Kalinowski and G. Kunstatter, J. Math. Phys. **24**, 886 (1984).
[9] Alternatively, one can do a Palatini variation of $g_{\mu\nu}$ and $W^{\lambda}_{\mu\nu}$ which yields Eqs. (2.2) and (2.3) as field equations. See G. Kunstatter, Gen. Rel. Grav. **12**, 373 (1980) and Ph. D. thesis, University of Toronto, 1979 (unpublished).
[10] M. W. Kalinowski and R. B. Mann, University of Toronto preprint, 1983.
[11] A. Papapetrou, Proc. Roy. Ir. Acad. Sec. A **52**, 69 (1948).
[12] G. Kunstatter and J. W. Moffat, Phys. Rev. D **19**, 1084 (1978).
[13] D. N. Pant, Nuovo Cimento B **25**, 175 (1975).
[14] J. W. Moffat, Phys. Rev. D **19**, 3562 (1979).
[15] J. A. Wheeler, Phys. Rev. D **97**, 511 (1955).
[16] M. Born and L. Infeld, Proc. R. Soc. London Ser. A **144**, 425 (1934).
[17] The more general case with $f \neq 0$, $B \neq 0$ is considered by R. B. Mann, Harvard University preprint HUTP-83/A055, 1983.
[18] J. W. Moffat, Acta Phys. Austriaca, Suppl. XXIII, 665 (1981).

Massless fermions and Kaluza–Klein theory with torsion

Yong-Shi Wu and A. Zee
University of Washington, Seattle, Washington 98195

(Received 12 March 1984; accepted for publication 20 April 1984)

A pure Kaluza–Klein theory contains no massless fermion in four-dimensional theory. We investigate the effect of introducing torsion on the internal manifold and find that there are massless fermions. The hope is that given an isometry group the representation to which these fermions belong is fixed, in contrast to the situation in Yang–Mills theory. We show that this is indeed the case, but the representations do not appear to be the ones favored by current theoretical prejudice. The cases with parallelizable torsions on a group manifold as the internal manifold are analyzed in detail.

PACS numbers: 04.50. + h, 14.60. − z, 02.20. + b, 02.40. + m

I. INTRODUCTION

One of the stumbling blocks facing theorists trying to treat Kaluza–Klein theory[1–3] as the theory of the world is the difficulty of obtaining low mass fermions.[3,4]

Consider a Kaluza–Klein theory[5] based on $M^4 \times B$, where B is an n-dimensional compact manifold with an isometry group G. The resulting theory contains Yang–Mills fields transforming according to the gauge group G. The length scale l_{kk} of B is of the order of the Planck length divided by the gauge coupling constant. It is well known by now that if the $(4+n)$-dimensional theory contains a massless fermion field, then the resulting four-dimensional theory contains an infinite spectrum of fermion fields whose masses are determined by the eigenvalues of the internal Dirac operator $i\not{D}^{(\mathrm{int})}$ appropriate to B. Since the natural mass scale is set by the huge Kaluza–Klein mass $M_{kk} \equiv l_{kk}^{-1}$, the observed quarks and leptons must correspond to the zero eigenvalue of $i\not{D}^{(\mathrm{int})}$.

The difficulty is that if B is a homogeneous space $B = G/H$, then $i\not{D}^{(\mathrm{int})}$ has no zero eigenvalue. The reason is as follows.[5] For G/H (with standard metric) one can show that the scalar curvature R is positive. On the other hand, the square of the Dirac operator may be evaluated[5,6] to be $-D^2 + \frac{1}{4}R$ and is therefore the sum of a non-negative operator and a positive operator.

One possible way around this difficulty involves introducing explicit gauge fields not related to the metric. If the ground state of the theory is such that these explicit gauge fields assume a topological configuration on B, then zero modes exist for the internal Dirac operator.[7,8]

In this paper we investigate an alternative possibility, that of introducing torsion on the manifold B. This means that we treat the connection $\omega^a{}_{bc}$ on B as an object unrelated to the Vielbein $e^a{}_i$. (Our convention is that of Ref. 5. See also Appendix A.) Since the internal Dirac operator, which we identify henceforth as the mass operator M, is given by

$$M \equiv i\not{D}^{(\mathrm{int})} = i\gamma^a(e^{ai}\partial_i - (i/4)\omega^a{}_{bc}\sigma^{bc}), \tag{1.1}$$

one might easily imagine that with some choices of ω one can find zero modes of M.

We show below that the introduction of torsion indeed allows the existence of numerous fermion zero modes. Unfortunately, there are an equal number of left- and right-handed zero modes. This is a problem which has cropped up repeatedly in contemporary particle theory.[9] To demonstrate that this is generally the case, one would have to show that (1) the Atiyah–Singer theorem is unaffected by the introduction of torsion and (2) torsion does not change the Pontryagin number of a manifold. (A partial discussion of these points will be given in Appendices B and C.) Thus, our discussions would appear to be irrelevant for the real world unless we suppose that the fundamental interaction at the preon level is left–right symmetric and that the left–right symmetry is broken by some as-yet unknown mechanism. Nevertheless, we feel that the effect of torsion is worth investigating in some detail.

One class of manifolds with torsion consists of the parallelizable manifolds defined by Cartan and Schouten.[10] The Cartan–Schouten program is a particularly restrictive way of introducing torsion on certain manifolds so that the Riemann curvature tensor vanishes. The preceding discussion makes it suggestive that zero curvature might allow M to have zero modes. Compact Lie groups form a wide class of parallelizable manifolds and we will focus on group manifolds in this paper. Not surprisingly, we are led, after some work, to face certain equations endowed with a rather neat algebraic structure which may be of some mathematical interest in themselves.

We have solved these algebraic equations. It turns out that normally fermion zero modes form an even number of "families," but the number of zero modes escalates rapidly as the rank of the group increases. For example, for SU(5), there are four "families" of zero modes in the representation 1024.

After our work was completed, we learned that Orzalesi and collaborators[11] had launched an extensive program of studying torsion in Kaluza–Klein theory. In particular, Destri, Orzalesi, and Rossi (in Ref. 11) were the first to point out the relevance of torsion for the existence of Dirac zero modes and have studied the case of parallelized group manifolds. Their analysis, while employing a slightly different formalism, is essentially the same as ours, but they do not determine the representation in question for a general simple Lie group as explicitly as we do. Also we give a method for reducing

 0022-2488/84/092696-08$02.50

this representation and point out the appearance of the repetitive structure in this reduction. They also studied the dynamical basis for compactification with torsion which we do not do. For an alternative application of parallelizable torsions in Kaluza–Klein theories, see Ref. 12.

In Sec. II, a brief review of the Cartan–Schouten program is given. In Sec. III, we work out the reduction of fermions in Kaluza–Klein theory. Putting together the material from these two sections, we find in Sec. IV that the search for fermion zero modes leads us to some interesting group theory problems which we solve in Secs. V and VI.

II. TORSION ON GROUP MANIFOLDS

A manifold is said to have torsion if the connection 1-form $\omega^a{}_b$ is treated as independent of the Vielbein 1-form e^a. Define the torsion two-form by

$$T^a = de^a + \omega^a{}_b e^b. \tag{2.1}$$

Without torsion, $T^a = 0$ and so $\omega^a{}_b$ is determined in terms of e^a. The Riemann curvature is given in any case by

$$R^a{}_b = d\omega^a{}_b + \omega^a{}_c \omega^c{}_b. \tag{2.2}$$

We focus our attention on compact Lie groups G. The points of the manifold are associated with group elements g. At a given point, one defines the Vielbein by

$$g^{-1}\, dg = \sum_a i(\lambda^a/2) e^a. \tag{2.3}$$

We normalize the generators of the Lie algebra of G by

$$\left[\frac{\lambda^a}{2}, \frac{\lambda^b}{2}\right] = i f^{abc} \frac{\lambda^c}{2}. \tag{2.4}$$

(We have chosen the Cartan metric on the group to be just δ^{ab} so that we need not distinguish between upper and lower group indices.) Differentiating Eq. (2.3), one finds the Cartan–Maurer equation

$$de^a = \tfrac{1}{2} f^{abc} e^b e^c. \tag{2.5}$$

Without torsion, we see by referring to Eq. (2.1) that the Cartan–Maurer equation implies $\omega^{ab} = \frac{1}{2} f^{abc} e^c$. It is natural and pleasing that the connection ω^{abc} is related to the structure constant f^{abc}. We now take a "physicist's" constructive approach to parallelizable torsions. Adopt the ansatz

$$\omega^{abc} = \tfrac{1}{2} K f^{abc}; \tag{2.6}$$

this defines a one-parameter family of possible connections. A simple computation using Eq. (2.2) and Jacobi's identity gives

$$R^{ab} = \tfrac{1}{4} K (1 - \tfrac{1}{2} K) f^{abc} f^{cdg} e^d e^g. \tag{2.7}$$

By definition, a parallelizable connection leads to a vanishing curvature tensor. This yields two solutions:

$$K = 0, \quad \omega^{abc} = 0, \quad \tau = +1, \tag{2.8}$$

$$K = 2, \quad \omega^{abc} = f^{abc}, \quad \tau = -1. \tag{2.9}$$

The quantity τ specifies the corresponding torsion by

$$T^a = \tfrac{1}{2} \tau f^{abc} e^b e^c. \tag{2.10}$$

Referring to Eq. (A5) in Appendix A, we determine the Christoffel symbol to be

$$\Gamma_{ij}{}^k = e_a{}^k \partial_i e_j{}^a \quad \text{in the } \tau = +1 \text{ case}, \tag{2.8'}$$

and

$$\Gamma_{ij}{}^k = -(\partial^k e_i{}^a) e_{aj} \quad \text{in the } \tau = -1 \text{ case}. \tag{2.9'}$$

Note that since $\omega^a{}_b$ transforms as a connection and not as a tensor 1-form, the relations in Eqs. (2.8) and (2.9) are clearly specific to the Vielbein basis defined by the left-invariant one-form in Eq. (2.3).

The preceding is a simple realization for group manifolds of the Cartan and Schouten theory of parallelizable connections. They posed themselves the following problem. Given a Riemannian manifold with a metric g_{ij} and a Christoffel symbol $\overset{(0)}{\Gamma}{}_{ij}{}^k$ constructed from the metric, is it possible to find a tensor $S_{ij}{}^k$ such that the following three conditions hold (defining $\Gamma_{ij}{}^k \equiv \overset{(0)}{\Gamma}{}_{ij}{}^k + S_{ij}{}^k$)?

(1) The geodesics for Γ are the same as those for $\overset{(0)}{\Gamma}$.

(2) The covariant derivative of g_{ij} relative to Γ vanishes.

(3) The curvature tensor constructed from g_{ij} and $\Gamma_{ij}{}^k$ vanishes.

It is easy to see that conditions (1) and (2) imply that S_{kij} is totally antisymmetric. We understand that Cartan and Schouten prove a nontrivial theorem stating parallelizable manifolds (i.e., those allowing parallelizable connections) besides Euclidean spaces are group manifolds and the seven-sphere S^7. To appreciate this last statement consider simply setting the connection one-form $\omega^a{}_b = 0$ on an arbitrary manifold. This certainly insures $R^a{}_b = 0$. However, in the Cartan–Schouten construction, one replaces a torsion-free connection $\overset{(0)}{\omega}{}^a{}_b$ by

$$\omega^a{}_b = \overset{(0)}{\omega}{}^a{}_b + \lambda^a{}_b, \tag{2.11}$$

where $\lambda^a{}_b$ is required to be a tensor one-form. By simply setting $\lambda^a{}_b = -\overset{(0)}{\omega}{}^a{}_b$ one defines, in a coordinate-dependent way, a $\lambda^a{}_b$ (and therefore $\omega^a{}_b$) which is not globally defined (i.e., singular somewhere) on the manifold, except for those named in the Cartan–Schouten theorem.

We return to our discussion of group manifolds. There exists a beautiful theorem that on group manifolds the Vielbein defined in Eq. (2.3), which is obviously left-translation invariant, gives a set of Killing vectors

$$\xi^a_i = e^a_i. \tag{2.12}$$

This can be easily proven in the following way.

Written out in component form, the Cartan–Maurer equation states

$$\partial_i e^a_j - \partial_j e^a_i = f^{abc} e^b_i e^c_j. \tag{2.13}$$

Multiplying by $e^{di} e^{fj} e^{ak}$ we find that the Killing vectors defined in Eq. (2.12) satisfy Lie's equation

$$\xi^{ai} \partial_i \xi^{bj} - \xi^{bi} \partial_i \xi^{aj} = -f^{abc} \xi^{cj}. \tag{2.14}$$

It is now easy to show that the metric $g^{ij} = e^{ai} e^{aj}$ admits the ξ^a_i as Killing vectors. One verifies, by using Lie's equation, that the Lie derivative of the metric along a Killing vector vanishes:

$$\mathcal{L}_{\xi^a} g^{ij} = \xi^{ak}(\partial_k g^{ij}) - (\partial_k \xi^{ai}) g^{kj} - (\partial_k \xi^{aj}) g^{ik} = 0. \tag{2.15}$$

[Incidentally, a group manifold may be thought of as the symmetric space $G \times G/G_D$, where G_D is the diagonal subgroup of $G \times G$. The isometry could be effected either by left or by right multiplication of group elements, corresponding to the two choices of writing $g^{-1}\,dg$ or $dg\,g^{-1}$ in Eq. (2.3). It is well known in mathematics that the $\tau = +1$ connection in the right-invariant Vielbein corresponding to $dg\,g^{-1}$ has the same form as the $\tau = -1$ connection in the left-invariant one, and vice versa. So we need to do the analysis only with the left-invariant Killing vectors.]

III. REDUCTION OF FERMION FIELD

We discuss here the reduction of Dirac fields in Kaluza–Klein theory with torsion. The discussion is for a general internal manifold B with an isometry group G.

Start with the Dirac Lagrangian in $(4+n)$-dimensional theory:

$$\mathcal{L} = \bar{\Psi} i \hat{\gamma}^{\mu}(\partial_{\hat{\mu}} - (i/4)\sigma^{\hat{\alpha}\hat{\beta}}\,\hat{\omega}_{\hat{\alpha}\hat{\beta}\hat{\mu}})\,\hat{\Psi} \equiv \bar{\Psi} i \hat{\gamma}^{\mu} \mathcal{D}_{\hat{\mu}}\,\hat{\Psi}. \tag{3.1}$$

Our notation is the same as in Ref. 5. Briefly, Greek indices, $\mu, \nu, \ldots, \alpha, \beta \ldots$ refer to the "external" four-dimensional space while Latin indices $i, j, \ldots a, b, \ldots$ refer to the "internal" space. The "hat" notation is used when we have to refer to the entire $(4+n)$-dimensional space. We also find it convenient occasionally to refer to the "internal" coordinates x^i collectively as y and to the "external" coordinates as x.

Fortunately, if we are interested only in dimension-4 terms in the four-dimensional theory, we do not have to compute every component of $\hat{\omega}_{\hat{\alpha}\hat{\beta}\hat{\mu}}$. By a dimensional argument we can see that we can effectively set the connection in the Vielbein basis to be

$$\begin{aligned} &\hat{\omega}_{abc} \to \omega^{(\mathrm{int})}_{abc}, \\ &\hat{\omega}_{\alpha\beta\gamma} \to \omega^{(\mathrm{ext})}_{\alpha\beta\gamma}, \\ &\hat{\omega}_{ab\alpha} \to e_{ai}(e^j_b\,\partial_j \xi^{\,i}_{r} - \xi^{\,j}_{r}\partial_j\, e^i_b) A^{\,r}_{\mu}\, e^{\mu}_{\alpha}. \end{aligned} \tag{3.2}$$

Here $\xi^{\,r}_{j}$ denotes the Killing vector corresponding to the generator of G labeled by the index r and $A^{\,r}_{\mu}$ denotes the corresponding gauge potential. We have chosen the Cartan metric to be flat $\propto \delta^{rs}$.

The Dirac field $\hat{\Psi}$ transforms as a spinor under the local group SO$(4+n)$ [or SO$(3+n,1)$]. Recall that the theory of orthogonal groups is such that the spinor index carried by $\hat{\Psi}$ factorizes into a spinor index for SO(4) and a spinor index for SO(n). The gamma matrices factorize accordingly:

$$\gamma^{\alpha} = \gamma^{\alpha} \times \gamma^5, \tag{3.3a}$$

$$\gamma^{a} = 1 \times \gamma^{a}, \tag{3.3b}$$

$$\gamma_5 = \gamma_5 \times \gamma_5. \tag{3.3c}$$

Note that the presence of γ_5 in Eq. (3.3a) but not in (3.3b) is necessary in order for $\{\gamma^{\alpha}, \gamma^{a}\} = 0$. In Eq. (3.3c) the notation is such that the three γ_5 denote the γ_5 matrix for the Clifford algebra corresponding to SO$(4+n)$, SO(4), and SO(n), respectively.

Combining Eqs. (3.2) and (3.3) we find that

$$\gamma^{\hat{\mu}} \mathcal{D}_{\hat{\mu}} \to \gamma^{\mu} \mathcal{D}^{(\mathrm{ext})}_{\mu} + \gamma^{i} \mathcal{D}^{(\mathrm{int})}_{i} - \gamma^{\alpha} e^{\mu}_{\alpha} A^{\,r}_{\mu}\, T_r. \tag{3.4}$$

Here $\mathcal{D}^{(\mathrm{ext})}_{\mu}$ and $\mathcal{D}^{(\mathrm{int})}_{\mu}$ are the covariant derivatives (not including the Yang–Mills potential) constructed out of $\omega^{(\mathrm{ext})}$ and $\omega^{(\mathrm{int})}$, the "external" and "internal" connections, respectively. The eigenvalues of the operator $M \equiv i\gamma^{i} \mathcal{D}^{(\mathrm{int})}_{i}$ determine the fermion mass spectrum in the resulting four-dimensional theory and so M may be identified as the mass operator.

The operators

$$iT_r \equiv \xi^{\,j}_{r}\,\partial_j + (i/4) e_{ai}(e^j_b \partial_j \xi^{\,i}_{r} - \xi^{\,j}_{r}\,\partial_j\, e^i_b) \tag{3.5}$$

are very interesting. We see that if Eq. (3.4) is to describe correctly the coupling of the Yang–Mills potential to fermion fields, we must have

$$[T_r, T_s] = i f_{rst}\, T_t, \tag{3.6}$$

$$[T_r, M] = 0. \tag{3.7}$$

The operators T_r have been discussed previously by Wetterich[13] and by Tanaka.[14] The derivation given by Tanaka is different from the direct approach followed here and offers additional insight into the origin of T_r. We continue this discussion paying special attention to the case with torsion.

Since we want to add torsion only in the internal manifold, among the components of $\hat{\omega}_{\hat{\alpha}\hat{\beta}\hat{\gamma}}$ in Eq. (3.2) none other than $\hat{\omega}_{abc}$ should be changed, and the change of $\hat{\omega}_{abc}$ is simply to add torsion in $\omega^{(\mathrm{int})}_{abc}$. The operators T_r given by Eq. (3.5) are unchanged by addition of torsion only in the internal manifold. However, if we express T_r in terms of the covariant derivative in the internal space, the expression differs in the cases with and without torsion. Recall that in the case without internal torsion, we have

$$iT_r \equiv \xi^{\,i}_{r}(\partial_i - (i/4)\omega_{abi}\sigma^{ab}) + (i/4) e^j_a e^k_b\, \xi^{\,r}_{j;k}\sigma^{ab}, \tag{3.8}$$

where $\xi^{\,r}_{j;k}$ is the usual covariant derivative of $\xi^{\,r}_{j}$. But, if there is internal torsion, using Eqs. (A4) and (A6) of Appendix A we can rewrite T_r as

$$\begin{aligned} iT_r = {} & \xi^{\,i}_{r}\,\partial_i + (i/4)\sigma^{ab} \\ & \times [e^i_a e^j_b (\nabla_j \xi_{ri} - T_{ijk}\,\xi^{\,k}_{r}) - \omega_{abk}\,\xi^{\,k}_{r}], \end{aligned} \tag{3.9}$$

where

$$\nabla_j \xi^{\,i}_{r} = \partial_j \xi^{\,i}_{r} + \Gamma^{i}_{kj} \xi^{\,k}_{r}. \tag{3.10}$$

According to Eqs. (3.6) and (3.7) the eigenmodes of M corresponding to a given eigenvalue furnish a representation of the group G. One defines[14] spinor harmonics $U(y)$ by the equation

$$(T_r)_{NN'} U^{(\sigma\pm)K}_{N'}(y) = (t^{(\sigma)}_{r})^{KK'} U^{(\sigma\pm)K'}_{N}(y). \tag{3.11}$$

Here N denotes a spinor index (of the internal space), σ specifies the representation of the group G, and K labels the components of the representation σ. We write the representation matrix in the representation σ as $t^{(\sigma)}_{r}$. Furthermore, since T_r commutes with γ_5 of the internal space, the harmonics can be chosen to be chiral eigenstates

$$\gamma_5 U^{(\pm)} = \pm U^{(\pm)}. \tag{3.12}$$

On the other hand, M anticommutes with γ_5 and so takes $U^{(+)}$ into $U^{(-)}$:

$$MU^{(\sigma\pm)} = m^{(\sigma)}U^{(\sigma\mp)}. \tag{3.13}$$

We can now expand the field Ψ as a sum of the spinor harmonics over the internal manifold:

$$\Psi^{(+)} = \sum_{\sigma,K} \{ \psi^{(\sigma+)K}(x)U^{(\sigma+)K}(y) + \psi^{(\sigma-)K}(x)U^{(\sigma-)K}(y)\}. \tag{3.14}$$

The coefficients $\psi^{(\sigma\pm)}(x)$ in this expansion are the fermion fields of the four-dimensional theory. The expansion in Eq. (3.14) is for a right-handed field

$$\gamma_5 \Psi^{(+)} = +\Psi^{(+)}. \tag{3.15}$$

The correlation of chirality in Eq. (3.14) is dictated by Eq. (3.3c).

IV. FERMIONS ON GROUP MANIFOLDS

Here we specialize the discussion of the preceding sections to group manifolds. In this case, the operators T simplify to the form

$$T^a = -i(\xi^{aj}\partial_j - (i/4)\sigma^{bc}f^{abc}) \equiv X^a + Y^a. \tag{4.1}$$

(Since the number of group labels is now equal to the dimension of the manifold, we identify the indices $r, s, \ldots$ as $a, b \ldots$, keeping in mind that T^a is independent of the presence of torsion.)

We find it useful to define the operators

$$X^a \equiv -i\xi^{aj}\partial_j \quad \text{and} \quad Y^a \equiv -\tfrac{1}{4}\sigma^{bc}f^{abc},$$

as indicated in Eq. (4.1). The presence of Y^a in this equation reminds us that under a Killing displacement one has to turn the spinor indices on a Dirac field. We note the algebraic structure

$$[X^a, X^b] = if^{abc}X^c, \tag{4.2}$$

$$[X^a, Y^b] = 0, \tag{4.3}$$

$$[Y^a, Y^b] = if^{abc}Y^c, \tag{4.4}$$

which follows from Lie's equation and from Jacobi's identity. This insures the correct commutation relation for T^a.

Clearly, if one were to reduce scalar fields in Kaluza–Klein theory, X^a would play the role of T^a. Scalar harmonics are defined by an equation analogous to Eq. (3.11):

$$X^a V^{(\sigma)K}(y) = (t^{(\sigma)a})^{KK'} V^{(\sigma)K'}(y). \tag{4.5}$$

[For $G = \mathrm{SO}(3)$ the $V(y)$'s are just the standard rotation functions.] Similarly, we can represent the algebra as realized by the Y^a's [Eq. (4.4)]:

$$(Y^a)_{NN'} W_{N'}^{(\sigma)K} = (t^{(\sigma)a})^{KK'} W_N^{(\sigma)K'}. \tag{4.6}$$

We learn from Eqs. (4.1)–(4.4) that we have here a problem analogous to the addition of angular momentum in quantum mechanics. We can decompose

$$U_N^{(\sigma)K}(y) = \sum_{\substack{\sigma_1\sigma_2 \\ K_1K_2}} C(\sigma K;\sigma_1K_1,\sigma_2K_2) \times V^{(\sigma_1)K_1}(y) W_N^{(\sigma_2)K_2}. \tag{4.7}$$

Here C denotes generalized Clebsch–Gordon coefficients. The dependences of U on the spinor index and on the coordinates of the internal manifold separate.

The mass operator

$$M = i\gamma^c(e^{ci}\partial_i - (i/4)\omega_{abc}\sigma^{ab}) \tag{4.8}$$

depends on torsion. Referring to Eqs. (2.8), (2.9), and (2.12), we see that for group manifolds M has the elegant algebraic form

$$M = -\gamma^a(X^a + \tfrac{1}{2}KY^a), \tag{4.9}$$

where K is a real parameter as defined in Eq. (2.6). For the case of no torsion, $K = 1$. For $\tau = +1$, $K = 0$, and for $\tau = -1$, $K = 2$. The fact that M commutes with T^a follows from the algebraic equations, Eqs. (4.1)–(4.4), and from the fact that γ^a transforms in the adjoint representation:

$$[T^a, \gamma^b] = [Y^a, \gamma^b] = if^{abc}\gamma^c. \tag{4.10}$$

We know from a general theorem that for $K = 1$, the torsion-free case, M has no zero mode. As K varies over the real line, zero modes may appear. In particular, consider the two parallelizable cases. The $\tau = +1$ case is easier and will be discussed first.

For $\tau = +1$, $M = i\gamma^a\xi^{aj}\partial_j$, and so it follows immediately that there is one class of zero modes for which $U(y)$ is independent of y. To put it more formally, we set σ_1 in Eq. (4.7) to be the trivial representation so that we simply have to solve Eq. (4.6). This group theoretic problem is treated in the next section. It can be shown that such zero modes are the only ones, because

$$M^2U = (X^a + Y^a)^2U - (Y^a)^2U = 0, \tag{4.11}$$

and because of the property of the Casimir invariant.

V. A GROUP THEORY PROBLEM

We now address the group theoretic problem encountered in the last section. We will phrase the problem in somewhat more general terms in order to clarify the problem.

Given a Lie algebra G, let $A(G) = \mathrm{SO}(N)$, where $N =$ the number of generators of G. [With the choice that the Cartan metric can be just the Kronecker delta, as in Sec. II, the elements of $\mathrm{SO}(N)$ are automorphisms of G.] Let the generators of $\mathrm{SO}(N)$ be represented by the matrices σ_{bc}, $b,c = 1 \ldots N$, in some representation r. Define a set of matrices Y_a by

$$Y_a = -\tfrac{1}{4}f_{abc}\sigma_{bc}. \tag{5.1}$$

Then, one can verify, using the Jacobi identity, that

$$[Y_a, Y_b] = if_{abc}\, Y_c. \tag{5.2}$$

Thus, Y_a furnishes a representation of G. The definition of Y_a [Eq. (5.1)] defines a map from the set of representation of $A(G)$ into the set of representations of G. We will refer to this map as a projection and write $P(r)$ as the representation of G corresponding to a representation r of $A(G)$. In the last section we are specifically interested in the projection of the spinor representation s of $A(G) = \mathrm{SO}(N)$ [or, strictly speaking, of the spin (N) covering of $\mathrm{SO}(N)$]. We want to determine $P(s)$.

[Our knowledge of the mathematical literature is rather limited. However, as far as we can determine by cursory discussions with a couple of mathematicians, our treatment is not in the standard mathematical literature. Partial results

have been given for the special cases SU(2), SU(3), and SU(5) by Destri *et al.*[11]]

One can easily prove a series of fairly obvious theorems.

Theorem 1: The projection of a reducible representation is reducible. The projection of an irreducible representation may or may not be reducible.

Theorem 2: $P(r_1 \times r_2) = P(r_1) \times P(r_2)$.

This theorem allows us to determine the projection of any representation of SO(N) once we know $P(s)$.

Theorem 3: The projection of the vector representation of SO(N) is the adjoint representation of g.

For example, for $G = \mathrm{SU}(3)$, $A(G) = \mathrm{SO}(8)$, and $P(\underline{8}) = \underline{8}$. Theorem 2 allows us to find, for instance, $P(28) = 10 + \overline{10} + \underline{8}$. The reducibility of $P(28)$ illustrates Theorem 1.

Clearly, the dimension of $P(r)$ is the same as the dimension of r. This provides one clue to determining $P(s)$: the dimension of $P(s)$ is equal to an integral power of 2. Incidentally, this proves rather indirectly the nonobvious theorem that any Lie algebra has a representation with dimensions equal to an integral power of 2.

We can exploit the fact that we know the explicit form of $\sigma_{ab} = (i/2)[\gamma_a, \gamma_b]$ for the spinor representation to evaluate the quadratic Casimir invariant for $P(s)$:

$$Y_a Y_a = (i^2/16) f_{abc} f_{ade} \gamma_b \gamma_c \gamma_d \gamma_e = \tfrac{1}{8} f_{abc} f_{abc}. \tag{5.3}$$

We used Jacobi's identity and properties of the gamma matrices. The fact that $Y_a Y_a$ comes out to be proportional to the unit matrix proves another theorem.

Theorem 4: $P(s)$ is either irreducible or a single irreducible representation repeated. (The number of repetitions is a power of 2.)

These considerations allow us to determine $P(s)$. After all, for a given G, there are not many representations of G with dimensions equal to $2^{[N/2]}/2^{k+1}$ with k a non-negative integer. (Here $[N/2]$ is the smallest integer not less than $N/2$.) It turns out that we have to express our solution using the Dynkin language. Our notation is the standard one as may be found in Refs. 15 and 16, for instance.

Recall that a representation is characterized by its highest weight Λ. Label the representation by Λ. The roots of the algebra are denoted by α_i. Let δ be half of the sum of the positive roots:

$$\delta = \frac{1}{2} \sum_{\substack{\text{positive} \\ \text{roots}}} \alpha_i. \tag{5.4}$$

There exists a theorem that[16]

$$2(\delta, \alpha_j) = (\alpha_j, \alpha_j). \tag{5.5}$$

The scalar product between two vectors α and β is given by

$$(\alpha, \beta) = \sum_{i,j} \alpha_i G_{ij} \beta_j, \tag{5.6}$$

where the Dynkin metric G_{ij} is listed in tables.[15] The α_i are the components of α in the Dynkin basis. The number of components is equal to the rank of G. Also, recall the famous Weyl formula for the dimension of a representation Λ:

$$\dim(\Lambda) = \prod_{\substack{\text{positive} \\ \text{roots}}} \frac{(\Lambda + \delta, \alpha_i)}{(\delta, \alpha_i)}. \tag{5.7}$$

The quadratic Casimir invariant of the representation Λ is given by

$$C(\Lambda) = (\Lambda, \Lambda + 2\delta). \tag{5.8}$$

In this language, we can write Eq. (5.3) as

$$C(\Lambda(P(s))) = \tfrac{1}{8} \dim(\mathrm{adj}) C(\mathrm{adj}). \tag{5.3'}$$

Here "adj" refers to the adjoint representation. We have to find a representation Λ which satisfies Eq. (5.3').

We assert the following.

Theorem 5: The highest weight of $P(s)$ is δ.

From Weyl's formula [Eq. (5.7)] we see immediately that

$$\dim(\delta) = 2^{(\text{number of positive roots})}, \tag{5.9}$$

which indeed is a power of 2.

To check Eq. (5.3') we evaluate

$$C(\delta) = 3(\delta, \delta) \tag{5.10}$$

using Eq. (5.8). Now, according to the theorem in Eq. (5.5), δ has the elegant form

$$\delta = (1,1,1,\ldots,1) \tag{5.11}$$

in the Dynkin basis, and so

$$C(\delta) = 3 \sum_{i,j} G_{ij} \tag{5.12}$$

is just the sum of all the entries in the Dynkin metric. Unfortunately, the Dynkin metric G_{ij} differs from Lie algebra to Lie algebra and so we have to evaluate $C(\delta)$ separately for the different cases in Cartan's classification. Furthermore, Λ_{adj} has different forms for different algebras and the evaluation of $C(\mathrm{adj})$ also has to proceed case by case.

Before we go to the general evaluation there are some simple cases for which the preceding formalism is not necessary. For $G = \mathrm{SU}(2)$, $A(G) = \mathrm{SO}(3)$, we can use the explicit form $\sigma_{ab} = -\epsilon_{abc} \tau_c$ for the spinor representation to evaluate $Y_a = (1/2)\tau_a$ so that $P(s) = \underline{2}$. The projection in this case obviously just expresses the local isomorphism between SO(3) and SU(2). For $G = \mathrm{SU}(3)$, $A(G) = \mathrm{SO}(8)$. After some thoughts, one finds $P(s) = \underline{8}$. According to Theorem 3, P (vector) $= \underline{8}$. This is consistent with the famous automorphism of SO(8) in which the two $\underline{8}$'s of spinor and the $\underline{8}$ of vector can be transformed into each other. In fact, we can exploit this automorphism to determine $P(s)$ in the first place. Let σ^s_{ab} and σ^v_{ab} be the generators of SO(8) in the spinor and vector representations, respectively. There exists a similarity transformation

$$\sigma^s_{ab} = \mathscr{S} \sigma^v_{ab} \mathscr{S}^{-1}. \tag{5.13}$$

Multiplying by f_{abc} and summing over a, b we obtain an explicit construction of $P(s)$.

The general evaluation below uses Tables 7 and 8 in Ref. (15):

$A_n(\mathrm{SU}_{n+1})$: $C(\delta) = \frac{1}{4} n(n+1)(n+2)$, $C(\mathrm{adj}) = 2n + 2$, $\dim(\mathrm{adj}) = n(n+2)$;

$B_n(SO_{2n+1})$: $C(\delta) = \frac{1}{4}n(2n-1)(2n+1)$,
$C(\text{adj}) = 2(2n-1)$,
$\dim(\text{adj}) = n(2n+1)$;

$C_n(Sp_{2n})$: $C(\delta) = \frac{1}{4}n(n+1)(2n+1)$,
$C(\text{adj}) = 2n+2$,
$\dim(\text{adj}) = n(2n+1)$;

$D_n(SO_{2n})$: $C(\delta) = \frac{1}{2}n(n-1)(2n-1)$,
$C(\text{adj}) = 4n-4$,
$\dim(\text{adj}) = n(2n-1)$;

G_2: $C(\delta) = 14$, $\dim(\delta) = 64$,
$C(\text{adj}) = 8$, $\dim(\text{adj}) = 14$;

F_4: $C(\delta) = 117$,
$\dim(\delta) = 16\,777\,216$,
$C(\text{adj}) = 52$, $\dim(\text{adj}) = 52$;

E_6: $C(\delta) = 234$,
$C(\text{adj}) = 24$, $\dim(\text{adj}) = 78$;

E_7: $C(\delta) = 1197/2$,
$C(\text{adj}) = 36$, $\dim(\text{adj}) = 133$;

E_8: $C(\delta) = 1860$,
$C(\text{adj}) = 60$, $\dim(\text{adj}) = 248$.

It is amusing to see that for all simple compact Lie groups the irreducible representation δ with the highest weight (1,1,...,1) in the Dynkin basis satisfies the relation $C(\delta) = \frac{1}{8}N(\text{adj})C(\text{adj})$.

The repetition number, mentioned in Theorem 4, of the irreducible representation δ in $P(s)$ is simply the quotient of the dimension of s, the spinor on the group manifold, over that of δ, namely $2^{[N/2]}/\dim(\delta)$. Thus, here see the natural emergence of something like family structure, with the number of families restricted to be a power of 2. For example, for SU(5), the zero modes form four families of the representation $\underline{1024}$.

Thus, except for the lowest ranked groups, we obtain an exceedingly large number of zero modes. For instance, for F_4, there are more than 16 million fermion zero modes! The reason is clearly that the number of zero modes increases exponentially in the number of generators in the group. We find it extremely unlikely that these zero modes could correspond to quarks and leptons. It is perhaps conceivable that at some preon level the gauge group is small, SU(2) say. One could also imagine a Kaluza–Klein theory with the internal manifold $SU(3)\times SU(2)\times SU(1)$. But the fermion representation appears to be incorrect.

We recognize that Eq. (5.3′) is a necessary but not sufficient condition. Thus, strictly speaking, we have only found a candidate solution and we have not proved that our solution is the solution. However, it seems highly unlikely that another representation exists with first the right dimension (a power of 2), and second, the right Casimir invariant [Eq. (5.3′)]. In particular, for those groups in which tables exist,[17] one can easily verify that our solution is unique. We have not bothered to try to complete the proof because unfortunately these zero modes appear to be irrelevant for phenomenology.

VI. SEARCH FOR ZERO MODES

We now return to the other parallelizable case in which $\tau = -1$. The mass operator simplifies to

$$M = i\gamma^a T^a. \tag{6.1}$$

The equation for zero modes $MU = 0$ can then be written as

$$(\gamma^a \otimes t^a)U = 0, \tag{6.2}$$

with γ^a and t^a acting on the spinor and group indices, respectively. (We suppress the index σ.) Alternatively, regard U as a (rectangular) matrix with a spinor index and a group index and write the rather strange matrix equation

$$\gamma^a U \tilde{t}^a = 0 \tag{6.2′}$$

($\tilde{t}$ is the transpose of t.)

A direct approach would involve using the harmonic expansion in Eq. (4.7). The equation $MU = 0$ then gives an equation involving $t^{(\sigma_1)}$, $t^{(\sigma_2)}$, and Clebsch–Gordon coefficients.

Let us apply M to Eq. (6.2) again

$$(\gamma^a\gamma^b \otimes t^a t^b)U = 0 = (1 \otimes t^2 + \tfrac{1}{2}\sigma^{ab} f^{abc} \otimes t^c). \tag{6.3}$$

We recognize the appearance of the operator Y_a so that Eq. (6.3) may be written as

$$(2Y_a \otimes t_a)U = t^2 U. \tag{6.4}$$

Here t^2 is the second Casimir invariant of the representation σ which U transforms as. The discussion of the preceding section on the property of Y_a tells us that for a given group the representation σ_2 appearing in the Clebsch–Gordon decomposition of $U^{(\sigma)}$ in Eq. (4.7) is determined [to be the one with the highest weight $\delta = (1,1,1,...,1)$]. For a given σ, the representation σ must appear in the direct product $\sigma \otimes \sigma_2^*$. Alternatively, one can regard Eq. (6.4) as an eigenvalue problem determining the representation σ to which the zero modes, if any, belong. [Of course, one must still insure that Eq. (6.2) is satisfied.]

For the simplest case SU(2) it is quite easy to prove that there is no nontrivial solution. For SU(2), $Y_a = \tau_a/2$. We "square" Eq. (6.4)

$$(\tau_a \otimes t_a)(\tau_b \otimes t_b)U = (t^2)^2 U = (1 \otimes t^2 - \tau_a \otimes t_a)U. \tag{6.5}$$

Using Eq. (6.4) again we find $t^2 = t^2(1 - t^2)$ which only has the trivial solution $t^2 = 0$. For groups larger than SU(2), a similar, but not so simple, analysis[18] can be made to show that the only solution to Eq. (6.2) is that with $t^2 = 0$. Therefore, in the $\tau = -1$ case, fermion zero modes are singlets.

VII. CONCLUSION

We conclude that with the introduction of torsion, Kaluza–Klein theories can have massless fermions, but not chiral fermions. Given a gauge group the fermion representation is determined. For parallelizable torsion, the representation is enormous for any but the smallest gauge groups. This is evidently related to the high dimension of group manifolds. As one possibility, we may reduce the dimension of the manifold by looking at a coset homogeneous space G/H instead of G itself. For instance, for SO(10), well known to be a leading candidate for a group relevant to the real world, we might look at $S_9 = SO(10)/SO(9)$. But unfortunately, S_9 is not parallelizable in the Cartan–Schouten sense. Modulo this difficulty, theories with spheres as internal manifold look quite promising to us. On S_9, the spinor is 16 dimensional. Thus, the fact that in SO(10) grand unification fermions belong to the $\underline{16}$ may be explained in the Kaluza–Klein con-

text. The geometry of the internal manifold may be reflected in the fermion spectrum.

In general, one can introduce an arbitrary amount of torsion, not necessarily just so as to make the Riemann curvature tensor vanish. By varying the parameter K so that the scalar curvature becomes negative, one may obtain, conceivably, massless fermions belonging to representations favored by current theoretical prejudice. But one would then be hard put to justify choosing that particular value of K.

These and other questions discussed here should be investigated further.

ACKNOWLEDGMENTS

We thank L. Brown, P. G. O. Freund, and E. Witten for helpful conversations and C. A. Orzalesi for correspondences clarifying Ref. 11.

This work was supported in part by the U. S. Department of Energy under Contract DE-AC06-81ER-40048.

APPENDIX A: CONNECTION AND TORSION

We briefly recall some elementary facts about differential geometry with torsion. Our notation is essentially that of Ref. 5 to which the reader unfamiliar with the subject may wish to turn.

One defines orthonormal basis vectors $\mathbf{e}_a$ and coordinate basis vectors $\mathbf{e}_i$ on the manifold. The Vielbein is defined by expanding

$$\mathbf{e}_a = e_a^i \mathbf{e}_i. \tag{A1}$$

The connection and the Christoffel symbol are defined by infinitesimal transport of the basis vectors:

$$\nabla_i \mathbf{e}_a = -\omega_a{}^b{}_i \mathbf{e}_b, \tag{A2}$$

$$\nabla_i \mathbf{e}_j = \Gamma^k_{ji} \mathbf{e}_k. \tag{A3}$$

Combining Eqs. (A1)–(A3) one finds

$$\partial_i e_a^k + \Gamma^k_{ji} e_a^j + \omega_a{}^b{}_i e_b^k = 0, \tag{A4}$$

which can be rewritten as a relation between Γ and ω,

$$\Gamma^k_{ji} = -(\partial_i e_a^k + \omega_a{}^b{}_i e_b^k) e_j^a. \tag{A5}$$

Recalling the torsion one-form is defined by $T^a = de^a + \omega^a{}_b e^b$ we see that

$$\Gamma^k_{ij} - \Gamma^k_{ji} = T^k_{ji}. \tag{A6}$$

The Christoffel symbol Γ^k_{ij} is symmetric in its two lower indices in the absence of torsion.

APPENDIX B: DIRAC OPERATOR WITH TORSION

In this appendix we discuss properties of the Dirac operator with torsion to see more closely how the usual positivity argument for the absence of zero modes breaks down in this case. We will consider only the internal manifold.

The internal Dirac operator is

$$i\not{\mathscr{D}}\psi = i\gamma^j \mathscr{D}_j \psi = i\gamma^a e_a^i (\mathscr{D}_i - (i/4)\sigma^{bc}\omega_{bci})\psi. \tag{B1}$$

Here ω_{bci} is a generic connection with torsion. Using Eq. (A5) in Appendix A we can prove that

$$\mathscr{D}_i(\gamma^j \mathscr{D}_j)\psi = \gamma^j \mathscr{D}_i \mathscr{D}_j \psi,$$

i.e., $[\gamma^j, \mathscr{D}_i] = 0.$ (B2)

Here we note that the proper definitions for the action of $\mathscr{D}_i$ on both sides are not actually identical:

$$\mathscr{D}_i(\gamma^j \mathscr{D}_j \psi) = (\partial_i - (i/4)\sigma^{ab}\omega_{abi})(\gamma^c e_c^j \mathscr{D}_j \psi), \tag{B3}$$

$$\mathscr{D}_i(\mathscr{D}_j \psi) = (\partial_i \delta^k_j - \Gamma^k_{ji} - (i/4)\sigma^{ab}\omega_{abi}\,\delta^k_j)(\mathscr{D}_k \psi) \tag{B3'}$$

because of the difference in transformation property of $(\gamma^j \times \mathscr{D}_j \psi)$ and $\mathscr{D}_j \psi$. Here Γ^k_{ji} is the connection with torsion in the coordinate basis.

From Eq. (B3') it follows that

$$[\mathscr{D}_i, \mathscr{D}_j] = -(i/4)\,\sigma^{ab}R_{abij} - T^k_{ij}\,\mathscr{D}_k, \tag{B4}$$

where $T^k_{ij} = \Gamma^k_{ji} - \Gamma^k_{ij}$. Therefore by using Eqs. (B2) and (B4) it is easy to obtain

$$(i\not{\mathscr{D}})^2\psi = \{ -g^{ij}\mathscr{D}_i\mathscr{D}_j + \tfrac{1}{8}\sigma^{ij}\sigma^{ab}R_{abij} - (i/2)\sigma^{ij}T^k_{ij}\mathscr{D}_k \}\psi. \tag{B5}$$

The first term on the right side is still a non-negative operator even if $\mathscr{D}_i$ has torsion in it,

$$\int d^n y \sqrt{g}\, \psi^\dagger g^{ij}\mathscr{D}_i\mathscr{D}_j \psi$$

$$= -\int d^n y \sqrt{g}\{ g^{ij}(\mathscr{D}_i \psi)^\dagger(\mathscr{D}_j \psi) + (\mathscr{D}_i g^{ij})(\mathscr{D}_j \psi) + \tfrac{1}{2} g^{ij}T^k_{ji}(\mathscr{D}_k \psi)\}, \tag{B6}$$

since the second and third terms in this equation vanish identically. However, both the second and third terms in Eq. (B5) are not necessarily positive definite. Therefore, the addition of torsion in the Dirac operator may lead to the appearance of zero modes.

Incidentally, if there is no torsion, then the second term in Eq. (B5) collapses to

$$\tfrac{1}{8}\sigma^{ij}\sigma^{ab}R_{abij} = \tfrac{1}{8}\sigma^{ab}\sigma^{cd}R_{abcd} = \tfrac{1}{4}R, \tag{B7}$$

because of

$$R_{abcd} = R_{cdab}, \quad R_{abcd} + R_{acdb} + R_{adbc} = 0. \tag{B8}$$

However, when there is torsion, the two properties of R_{abij} in Eq. (B8) are no longer true, and so neither is Eq. (B7).

APPENDIX C: INDEX THEOREM AND PONTRYAGIN NUMBER IN THE PRESENCE OF TORSION

The fact that we obtain equal number of left- and right-handed zero modes is perhaps not surprising. A heuristic argument[19] goes roughly as follows.

Torsion can be switched on continuously. One can write the ω_t, connection one-form, as

$$(\omega_t)^a{}_b = \overset{(0)}{\omega}{}^a{}_b + t\lambda^a{}_b \tag{C1}$$

so that as t goes from 0 to 1 the connection goes from $\overset{(0)}{\omega}$ to $\omega = \overset{(0)}{\omega} + \lambda$. The difference between the number of left- and right-handed zero modes $(n_L - n_R)$ is zero for $t = 0$ and so by continuity it should not jump to an integer value as t varies.

This argument is basically correct. However, there are a few technical gaps which we need to fill in to turn it into a proof. Here, in fact, we need more than continuity since $(n_L + n_R)$ does jump discontinuously as t varies. The crucial

point is that unlike $(n_L + n_R)$, the quantity $(n_L - n_R)$ is related by the Atiyah–Singer index theorem to a topological quantity, which is constant against continuous deformation of connections.

Therefore, to turn the above argument into an explicit proof we have to show that (1) the index theorem and (2) the Pontryagin numbers are not affected by the presence of torsion.

It should be straightforward to check (1). Here we only check (2).

Recall that in the index theorem[20]

$$n_L - n_R = -\int_M \left[1 - \tfrac{1}{24} P_1 + \tfrac{1}{5760}(7P_1^2 - 4P_2) + \cdots\right], \tag{C2}$$

the right-hand side involves only Pontryagin numbers ($P_k \propto \operatorname{tr} R^{2k}$). We want to show that Pontryagin numbers are not changed by torsion. Let R and $\overset{(0)}{R}$ be the curvature two-form constructed out of the connection one-forms ω and $\overset{(0)}{\omega}$, respectively. We now prove that

$$\int \operatorname{tr} R^{2k} = \int \operatorname{tr} \overset{(0)}{R}{}^{2k} \tag{C3}$$

(where the integrals are over a compact $4k$-dimensional manifold) if $\omega - \overset{(0)}{\omega} = \lambda$ is a tensor.

It suffices to show that

$$\operatorname{tr} R^{2k} - \operatorname{tr} \overset{(0)}{R}{}^{2k} = dX, \tag{C4}$$

where X transforms covariantly. We emphasize that we know $\operatorname{tr} R^{2k}$ is locally exact. Indeed, in a previous work we have derived the representation [Eq. (3.15) of Ref. 21]

$$\operatorname{tr} R^{2k} = d\left[2k \int_0^1 dt \operatorname{Str}\{\omega, (t\,d\omega + t^2\omega^2)^{2k-1}\}\right]. \tag{C5}$$

The properties of the symmetric trace Str may be found in Appendix B of Ref. 20. The point is that the quantity in the square bracket does not transform covariantly and so $\operatorname{tr} R^{2k}$ is not globally exact and its integral over a compact manifold does not necessarily vanish. The claim in Eq. (C4) is that $\operatorname{tr} R^{2k} - \operatorname{tr} \overset{(0)}{R}{}^{2k}$ is globally exact.

For k small, one can verify Eq. (C4) by explicit computation using Eq. (C5). For arbitrary k this approach becomes unwieldy. Instead, define ω_t as in Eq. (C1) and define

$$R_t \equiv d\omega_t + \omega_t^2 = \overset{(0)}{R} + t\overset{(0)}{D}\lambda + t^2\lambda^2. \tag{C6}$$

Here $\overset{(0)}{D}$ is the covariant derivative with the connection $\overset{(0)}{\omega}$. Then we find

$$\begin{aligned}\frac{d}{dt}\operatorname{tr}(R_t)^{2k} &= 2k \operatorname{tr}\left(\overset{(0)}{D}\lambda + 2t\lambda^2\right) R_t^{2k-1} \\ &= 2k \operatorname{tr}(D_t\lambda) R_t^{2k-1} \\ &= 2k \operatorname{Str}(D_t\lambda, R_t^{2k-1}) \\ &= 2k\, d\left[\operatorname{Str}(\lambda, R_t^{2k-1})\right].\end{aligned} \tag{C7}$$

Here D_t is the covariant derivative with the connection ω_t. The last step in Eq. (C7) follows from a property of Str [see Eq. (B13) of Ref. 20] and from the Bianchi identity $D_t R_t = 0$. The square bracket in Eq. (C7) transforms covariantly provided that $\lambda^a{}_b$ is a tensor one-form and so Eq. (C4) follows.

Again, one may be tempted to argue that Eq. (C7) follows merely from continuity. However, one needs the additional input that λ transforms covariantly.

Equation (C4) can be easily generalized to the cases in which the Pontryagin densities are of the form $\operatorname{tr} R^{2k_1} \cdot \operatorname{tr} R^{2k_2} \cdots \operatorname{tr} R^{2k_i}$. So the generic Pontryagin members are also unchanged by addition of torsion.

Incidentally, the discussion here amounts to an indirect proof that the Pontryagin densities for group manifolds vanish. Of course, one can compute them directly by using Eq. (2.7).

[1] Th. Kaluza, Sitzeingsber. Preuss. Akad. Wiss. Berlin, **K1**, 966 (1921).
[2] O. Klein, Z. Phys. **37**, 895 (1926).
[3] References on earlier literature and a nice overview may be found in E. Witten, Nucl. Phys. B **186**, 412 (1981).
[4] L. Palla, in *Proceedings of the 19th International Conference on High Energy Physics* (Tokyo, 1978), edited by S. Homma, M. Kawaguchi and H. Miyazawa (Physical Society of Japan, Tokyo, 1979), p. 629.
[5] Reader unfamiliar with the subject may find a reasonably pedagogical introduction in A. Zee, in *Proceedings of the Fourth Summer Institute on Grand Unified Theories and Related Topics*, edited by M. Konuma and T. Maskawa (World Scientific, Singapore, 1981).
[6] A. Lichnerowicz, in Proceedings of the 1963 Les Houches Summer School, *Relativity, Groups and Topology*, edited by C. DeWitt and B. DeWitt (Gordon and Beach, New York, 1964), p. 849, Eq. (20.5).
[7] Z. Horvath, L. Palla, E. Cremmer, and J. Scherk, Nucl. Phys. B **127**, 57 (1977).
[8] E. Witten (unpublished).
[9] M. Gell-Mann, P. Ramond, and R. Slansky, in *Supergravity*, edited by P. van Nieuwenhuizen and D. Z. Freedman (North-Holland, Amsterdam, 1979); F. Wilczek and A. Zee, Phys. Rev. D **25**, 553 (1982); H. Georgi and E. Witten (unpublished). This problem also appears in supergravity theories.
[10] E. Cartan and J. A. Schouten, Proc. Kon. Ned. Akad. Amsterdam **29**, 803 (1926); see also, S. Goldberg, *Curvature and Homology* (Academic, New York, 1962).
[11] C. A. Orzalesi, Fortschr. Phys. **29**, 413 (1981); C. A. Orzalesi and M. Pauri, Phys. Lett. B **107**, 186 (1981); M. J. Duff and C. A. Orzalesi, Phys. Lett. B **122**, 37 (1983); C. Destri, C. A. Orzalesi, and P. Rossi, Ann. Phys. **147**, 321 (1983). See also, S. Bergia, C. A. Orzalesi, and G. Venturi, Phys. Lett. B **123**, 205 (1983) and CERN TH-3647, 1983; C. A. Orzalesi, CERN TH-3648, 1983.
[12] H. C. Tze, Phys. Lett. B **128**, 160 (1983).
[13] C. Wetterich, Phys. Lett. B **110**, 379 (1982).
[14] S. Tanaka, preprint of Kyoto University, KUNS673 HE(TH) 83/05.
[15] R. Slansky, Phys. Rep. **79**, 1 (1981).
[16] Z. X. Wan, *Lie Algebras* (Pergamon, Oxford, 1975).
[17] W. G. McKay and J. Patera, *Tables of Dimensions, Indices and Branching Rules for Representations of Simple Lie Algebras* (Marcel Dekker, New York, 1981).
[18] N. Turok (private communication).
[19] We heard of this heuristic argument from E. Witten.
[20] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. C **66**, 213 (1980).
[21] B. Zumino, Y. S. Wu, and A. Zee, Seattle preprint 40048-18 P3, 1983.

An infinite chain of inequalities for correlation functions of classical lattice systems

J. Naudts
Physics Department, University of Antwerpen, Universiteitsplein 1, B-2610 Wilrijk, Belgium

(Received 18 January 1984; accepted for publication 13 April 1984)

In quantum-mechanical systems the moments of the symmetrized time-dependent autocorrelation function satisfy an infinite set of chained inequalities. By analogy the same inequalities are derived for correlation functions of classical lattice systems.

PACS numbers: 05.20.Gg, 64.60.Cn

I. INTRODUCTION

Systems in thermodynamical equilibrium satisfy the variational principle which implies a delicate balance between energy and entropy. Perturbations necessarily increase the free energy. Starting from this observation one can derive inequalities for the correlation functions of classical lattice systems.[1,2] It was noted[1] that the energy-entropy inequalities show a formal analogy to quantum-mechanical inequalities.[3,4] In the present paper the analogy is elaborated. As a result new inequalities are obtained. They are applied in a subsequent paper[5] to derive lower bounds for the fluctuations of an order parameter.

The quantum-mechanical inequalities relate moments of time-dependent autocorrelation functions. In fact an infinite chain of inequalities exist[6,7] between moments μ_n of different order n. It can be shown that an upper bound for the moment μ_{-1} implies a lower bound for the moment μ_1. In this way the quantum energy-entropy inequality (in its symmetrized form) is a consequence of the Roepstorff inequality [inequality (10) of Ref. 4, implicitly present in (B18) and (B20) of Ref. 3].

By analogy the same relations hold for the classical inequalities. The energy-entropy inequality of Ref. 1 in its symmetrized form is a consequence of a new inequality which is derived in the present paper. Also the analog of the infinite chain of moment inequalities is derived.

In the next section the notations are fixed. In Sec. III the origin of the analogy is discussed. Section IV introduces the moments of classical correlation functions. The basic relations between consecutive moments are derived. As a consequence bounds on moments always appear as pairs of inequalities. In Sec. V such pairs of inequalities are derived. In Sec. VI a short discussion follows. The Appendix elaborates the relation with the theory of modular automorphisms.

II. NOTATIONS

The configuration space of the classical lattice system is the product space K^{∞} of local configuration spaces $K_i, i \in Z^{\nu}: K^{\infty} = \Pi_i K_i$. The local spaces K_i are all copies of one and the same probability space (K, ρ_0), e.g., a compact group K with Haar measure ρ_0.

The interaction of the system is given by a set of functions

$$\{\phi(X) | X \text{ finite subset of } Z^{\nu}\}.$$

Each element $\phi(X)$ is a continuous function of the configurations in X. Assume that

$$\|\phi\| = \sup_i \sum_{X: i \in X} \|\phi(X)\|_{\infty} < +\infty. \tag{1}$$

The local Hamiltonians are defined by

$$H_{\Lambda} = \sum_{X \subseteq \Lambda} \phi(X). \tag{2}$$

Let us from now on fix a finite subset Λ of Z^{ν}. Consider a Λ-local transformation U of the configuration space, i.e., U is a continuous invertible map of K^{∞} into itself which does not change the configuration outside Λ. It is assumed that the *a priori* probability measure ρ_0 is invariant under U. The change in energy due to the transformation U is measured by the variable

$$P = \frac{1}{2} \sum_X (\phi(X) \circ U^{-1} - \phi(X)). \tag{3}$$

In the latter expression the summation is restricted to the sets X which intersect Λ. One has $P = \lim_{\Lambda \uparrow Z^{\nu}} \frac{1}{2}(H_{\Lambda} \circ U^{-1} - H_{\Lambda})$. From (1) there follows

$$\|P\|_{\infty} \leqslant N(\Lambda) \|\phi\|. \tag{4}$$

[$N(\Lambda)$ denotes the number of points in Λ.]

Hence P belongs to $\mathscr{C}(K^{\infty})$, the algebra of real-valued continuous functions on K^{∞}.

Let us finally fix a normalized regular Borel measure ρ on K^{∞} describing the equilibrium state of the system. Assume that ρ satisfies the DLR equations for the *a priori* probability measure ρ_0 and the interaction ϕ at inverse temperature $\beta = 1$. Then it follows for all A in $\mathscr{C}(K^{\infty})$ that

$$\rho(A \circ U) = \rho(A e^{-2P}). \tag{5}$$

The latter equation is the starting point of the present paper.

III. THE ANALOGY

The formal analogy between the classical and quantum-mechanical inequalities is unraveled. In the next sections the classical inequalities will be derived by use of the analogy, i.e., by translating the quantum-mechanical inequalities and copying the proofs.

In the quantum case the time-dependent correlation function is given by

$$f(t) = \rho(x^* x_t). \tag{6}$$

It satisfied the KMS boundary condition[8] which states that

 0022-2488/84/092704-04$02.50

the function f has an analytical continuation in part of the complex plane, such that

$$f(t+i)=\rho(x_t x^*). \tag{7}$$

One also considers the symmetrized correlation function

$$F(t)=f(t)+f(-t+i)=\rho(x_t x^*+x x^*_t). \tag{8}$$

The latter clearly satisfies the relation

$$F(t+i)=F(-t). \tag{9}$$

The inequalities of interest here[3,4,6,7] are direct consequences of the KMS boundary condition. Hence it is obvious to look for a time-dependent correlation function of the classical system which satisfies an analyticity property similar to the KMS boundary condition. However one cannot expect that the time evolution needed in the classical case coincides with the physical time evolution of the classical system. Indeed a physical time evolution of microscopic nature need not exist (see, e.g., the models with discrete spins). Moreover the classical limit of the KMS boundary condition[9] is no longer expressed as an analyticity condition but either as a static equation or as a differential equation.[10] In some cases one can study the classical limit of the quantum inequalities. But in the limit the outlook of the inequalities changes, and the formal analogy gets lost.

In the quantum inequalities the noncommutativity of the operators x and x^* (the Hermitian conjugate) is important. The classical surrogate for the Hermitian conjugation will be derived from a local transformation U of the phase space, as introduced in the previous section. The desired choice for (the classical analog of) the time-dependent correlation function turns out to be (using the notations of the previous section)

$$f(t)=\rho(Ae^{2iPt}). \tag{10}$$

Because the function P is bounded, an analytic continuation of f exists throughout the complex plane. Using (5) one obtains the boundary condition

$$f(t+i)=\rho((Ae^{2iPt})\circ U). \tag{11}$$

The symmetrized correlation function equals

$$\begin{aligned} F(t)&=f(t)+f(-t+i)\\ &=\rho(A(\cos 2Pt+i\tanh P\sin 2Pt)(1+e^{-2P})). \end{aligned} \tag{12}$$

Note that the classical observable A replaces the operator x^*x. Hence it is obvious that the condition $A\geqslant 0$ will be needed for the inequalities to hold.

The analogy can be carried through to a deeper level in case the transformation U satisfies the condition $U^2=1$. Then U is an involution which corresponds in the quantum context to a basic symmetry between representations and antirepresentations of the algebra of observables. The latter symmetry has been studied in the theory of modular automorphisms.[11] A short discussion on these matters is found in the Appendix.

IV. MOMENTS

The moments of the symmetrized correlation function $F(t)$ are defined in the usual way by

$$\mu_n=(2i)^{-n}\frac{d^n}{dt^n}F(t)|_{t=0},\quad n=0,1,2,..., \tag{13a}$$

and

$$\mu_{-1}=\int_0^1 dt\,F(it). \tag{13b}$$

A straightforward calculation using expression (12) shows that one has

$$\begin{aligned} \mu_n&=\rho(AP^n(1+(-1)^n e^{-2P}))\\ &=\rho(AP^n)+(-1)^n\rho((AP^n)\circ U). \end{aligned} \tag{14}$$

If $U^2=1$ is satisfied then one has $P\circ U=-P$ and there follows

$$\mu_n=\rho(AP^n)+\rho((A\circ U)P^n). \tag{15}$$

Throughout the rest of the paper it is assumed that $A\geqslant 0$, $A\neq 0$, and $P\neq 0$. Because the equilibrium state ρ is faithful it follows that all the moments μ_n are strictly positive.

Theorem 1: Let $n\in N$.

(a) If y is given by

$$\mu_{2n-1}=\mu_{2n}y^{-1}\tanh y, \tag{16a}$$

then one has

$$\mu_{2n+1}\geqslant\mu_{2n}y\tanh y. \tag{16b}$$

(b) If y is given by

$$\mu_{2n+1}=\mu_{2n}y\tanh y, \tag{17a}$$

then one has

$$\mu_{2n-1}\geqslant\mu_{2n}y^{-1}\tanh y. \tag{17b}$$

Proof: It is enough to prove the inequalities for $n=0$. For $n>0$ the result is obtained by substituting A by AP^{2n}. For $n=0$ the theorem gives the classical analog of the Falk and Bruch inequality[3] and corresponds to Theorem II.4 of Ref. 7. The inequalities found below in Theorem 2 correspond to those of Ref. 7 for $n>0$. For completeness the proofs are repeated here.

The function g defined on the unit interval by

$$g(y^{-1}\tanh y)=y\tanh y$$

is well-defined and convex (see Lemma II.3 of Ref. 7). Hence one has, by use of Jensen's inequality,

$$\begin{aligned} \mu_1&=\rho(AP(1-e^{-2P}))\\ &=\rho(A(1+e^{-2P})P\tanh P)\\ &=\rho(A(1+e^{-2P})g(P^{-1}\tanh P))\\ &\geqslant\rho(A(1+e^{-2P}))g\left(\frac{\rho(A(1+e^{-2P})P^{-1}\tanh P)}{\rho(A(1+e^{-2P}))}\right)\\ &=\mu_0 y\tanh y, \end{aligned}$$

with y given by

$$y^{-1}\tanh y=\rho(A(1+e^{-2P})P^{-1}\tanh P)/\rho(A(1+e^{-2P})).$$

Hence (17) follows.

Remark that

$$\mu_{-1}=\rho(AP^{-1}(1-e^{-2P}))\leqslant\rho(A(1+e^{-2P}))=\mu_0.$$

Hence one can decrease y in expression (17b) until one obtains an equality. By doing so (17b) becomes (16a), and (17a)

becomes (17b). In fact the expressions (16) and (17) are equivalent. This ends the proof.

The important consequence of the previous theorem is that any upper bound for the ratio μ_{2n-1}/μ_{2n} gives rise to a lower bound for μ_{2n+1}/μ_{2n}. Similarly any upper bound for μ_{2n+1}/μ_{2n} implies a lower bound for μ_{2n-1}/μ_{2n}.

V. PAIRS OF INEQUALITIES

The previous theorem relates triples of consecutive moments, and gives lower bounds for the odd numbered moments. The next theorem relates pairs of consecutive moments, and gives a lower and an upper bound for the odd numbered moments. The lower bound is less sharp than that of Theorem 1.

Theorem 3 gives an upper bound for the moment μ_{-1} and a lower bound for the moment μ_1. Again the lower bound is less sharp than that of Theorem 1.

Theorem 2: Let $n \geqslant 1$. Let $y = (\mu_{2n}/\mu_0)^{1/2n}$. Then one has

$$\text{(a)}\ \mu_{2n-1} \leqslant \mu_{2n} y^{-1} \tanh y, \tag{18a}$$

and

$$\text{(b)}\ \mu_{2n+1} \geqslant \mu_{2n} y \tanh y. \tag{18b}$$

Proof: The function g defined on the positive axis by

$$g(y^{2n-1} \tanh y) = y^{2n}$$

is well-defined and convex (see Lemma II.3 of Ref. 8). Hence one has, using Jensen's inequality,

$$\begin{aligned}\mu_{2n} &= \rho(A(1+e^{-2P})P^{2n}) \\ &= \rho(A(1+e^{-2P}))g(P^{2n-1}\tanh P)) \\ &\geqslant \rho(A(1+e^{-2P}))g\left(\frac{\rho(A(1+e^{-2P})P^{2n-1}\tanh P)}{\rho(A(1+e^{-2P}))}\right) \\ &= \mu_0 y^{2n}\end{aligned}$$

with y given by $y^{2n-1}\tanh y = \mu_{2n-1}/\mu_0$. Now the function $y \to y^{2n-1}\tanh y$ is monotonically increasing. Hence the foregoing is equivalent to inequality (18a). Inequality (18b) follows from (18a) and Theorem 1. This ends the proof of the theorem.

Theorem 3: One has

$$\text{(a)}\ \mu_{-1} \leqslant 2(\rho(A) - \rho(A\circ U))/\log(\rho(A)/\rho(A\circ U)), \tag{19a}$$

and

$$\text{(b)}\ \mu_1 \geqslant \tfrac{1}{2}(\rho(A) - \rho(A\circ U))\log(\rho(A)/\rho(A\circ U)). \tag{19b}$$

If $\rho(A) = \rho(A\circ U)$ then the inequalities reduce to $\mu_{-1} \leqslant \mu_0$ and $\mu_1 \geqslant 0$.

Proof: The proof uses the argument of Sec. 2 of Ref. 4, taken over in Theorem III.1 of Ref. 12. One has

$$\mu_{-1} = \rho(AP^{-1}(1-e^{-2P})) = \int_0^2 dt\, \rho(Ae^{-Pt}).$$

Let $g(t) = \log \rho(Ae^{-Pt})$. One has

$$\frac{d^2 g}{dt^2} = \rho(Ae^{-Pt})^{-2}(\rho(AP^2 e^{-Pt})\rho(Ae^{-Pt}) - \rho(APe^{-Pt})^2).$$

From Schwarz's inequality it follows that $d^2g/dt^2 \geqslant 0$. Hence the function g is convex. Therefore one has for $0 \leqslant t \leqslant 2$

$$g(t) \leqslant \tfrac{1}{2}(g(2) + g(0)) + \tfrac{1}{2}t(g(2) - g(0)).$$

One obtains

$$\begin{aligned}\mu_{-1} &= \int_0^2 dt \exp g(t) \\ &\leqslant \exp\left(\frac{1}{2}(g(2)+g(0))\right)\int_0^2 dt \exp\left(\frac{1}{2}t(g(2)-g(0))\right) \\ &= 2(g(2)-g(0))^{-1}(e^{g(2)} - e^{g(0)}).\end{aligned}$$

Now one has

$$g(2) - g(0) = \log(\rho(Ae^{-2P})/\rho(A)),$$

and

$$e^{g(2)} - e^{g(0)} = \rho(Ae^{-2P}) - \rho(A).$$

Hence (19a) follows. The inequality (19b) follows from (19a) and Theorem 1. This ends the proof of Theorem 3.

VI. DISCUSSION

By analogy to the quantum case[7] one can expect that if the inequalities (18) of Theorem 2 hold for arbitrary functions $A \geqslant 0$ then the equality (5) should hold. Similarly if one of the inequalities of Theorem 3 holds for arbitrary $A \geqslant 0$ and for a sufficiently large class of transformations U, then one can expect in the light of the proof of Ref. 1 that the DLR equations are satisfied. A confirmation of these expectations would indicate that the inequalities are optimal. The problem is not treated in the present paper.

The energy-entropy inequality of Ref. 1 in the present notations reads

$$2\rho(AP) \geqslant \rho(A)\log(\rho(A)/\rho(A\circ U)). \tag{20}$$

Assume that $U^2 = 1$. Then (20) implies

$$2\rho((A\circ U)P) \geqslant \rho(A\circ U)\log(\rho(A\circ U)/\rho(A)).$$

There follows using (15)

$$\begin{aligned}2\mu_1 &= 2\rho((A + A\circ U)P) \\ &\geqslant (\rho(A) - \rho(A\circ U))\log(\rho(A)/\rho(A\circ U)),\end{aligned}$$

which coincides with inequality (19b). This shows that the inequality (19b) is the symmetrized form of the energy-entropy inequality (20).

An application of the new inequality (19a) is found in a separate paper.[5]

ACKNOWLEDGMENTS

Remarks by Professor H. Falk and Professor A. Verbeure concerning the literature on the quantum inequalities are gratefully acknowledged.

APPENDIX: MODULAR AUTOMORPHISMS

The analysis found here is not of direct relevance for the paper. But the observation made in Sec. III that the local transformation U is related to a kind of KMS boundary condition as expressed in Eqs. (10) and (11) is intriguing enough to justify some further investigation.

Let $\mathcal{A}$ denote the C^*-algebra of continuous complex

functions on K^{∞}. Then the equilibrium measure ρ extends to a faithful state on $\mathscr{A}$. A scalar product is defined on $\mathscr{A}$ in the usual way by

$$(f,f') = \rho(\overline{f'}f).$$

Denote $\mathscr{H}$ the complex Hilbert space obtained by closing $\mathscr{A}$.

Assume that $U^2 = 1$. Then a conjugate linear involution S on $\mathscr{H}$ is defined by

$$S: f \in \mathscr{A} \to \tilde{f} \circ U.$$

The algebra $\mathscr{A}$ equipped with the involution S is an involutive algebra but not a left Hilbert algebra (except if $U = 1$). Nevertheless part of the theory of modular automorphism groups can still be applied. Indeed it has been shown[13] that the KMS boundary condition is concerned with real subspaces of a complex Hilbert space; the presence of von Neumann algebra's is not essential.

Denote $\mathscr{K}_0 = \{ f \in \mathscr{A} \mid f \circ U = \bar{f} \}$ the set of self-adjoint elements for the involution S. One has the following result.

Theorem: There is a unique strongly continuous one-parameter group $(U_t)_t$ of unitaries on $\mathscr{H}$ leaving $\mathscr{K}_0$ invariant and satisfying the KMS boundary condition: for each pair f, f' in $\mathscr{K}_0$ there exists a function g, defined, bounded, and continuous on the strip $0 \leqslant \mathrm{Im}(z) \leqslant \frac{1}{2}$, analytic inside the strip, such that

$$g(t) = (U_t f f'), \quad \text{for all real } t,$$

and

$$g(t + i/2) \text{ is real}, \quad \text{for all real } t.$$

One has for all $f \in \mathscr{A}$ that $U_t f = e^{2iPt} f$.

Proof: (a) Existence. Let $(U_t)_t$ be defined by $U_t f = e^{2iPt} f$, $f \in \mathscr{A}$. It is straightforward to prove that the group $(U_t)_t$ is strongly continuous, leaves $\mathscr{K}_0$ invariant, and satisfies the KMS boundary condition.

(b) Uniqueness. Denote $\mathscr{K}$ the closure of $\mathscr{K}_0$ in $\mathscr{H}$. Remark that any $f \in \mathscr{A}$ has a unique decomposition $f = f_1 + if_2$ with f_1 and $f_2 \in \mathscr{K}_0$. Hence $\mathscr{K}_0 + i\mathscr{K}_0$ and *a fortiori* $\mathscr{K} + i\mathscr{K}$ is dense in $\mathscr{H}$. It then follows from Theorem 3.9 of Ref. 13 that $(U_t)_t$ coincides with the group of modular automorphisms associated with $\mathscr{K}$. Hence uniqueness follows. This ends the proof.

[1] M. Fannes, P. Vanheuverzwijn, and A. Verbeure, J. Stat. Phys. **29**, 545 (1982).
[2] B. Simon and A. D. Sokal, J. Stat. Phys. **25**, 679 (1981).
[3] H. Falk and L. Bruch, Phys. Rev. **180**, 442 (1969).
[4] G. Roepstorff, Commun. Math. Phys. **46**, 253 (1976).
[5] J. Naudts, "*A lower bound for fluctuations in classical lattice systems*," preprint, Universiteit Antwerpen, 1984.
[6] R. Martens, Ph.D.thesis, Leuven, 1979.
[7] M. Fannes, R. Martens, and A. Verbeure, Commun. Math. Phys. **80**, 529 (1981).
[8] R. Haag, N. M. Hugenholtz, and M. Winnink, Commun. Math. Phys. **5**, 215 (1967).
[9] G. Galavotti and E. Verboven, Nuovo Cimento B **28**, 274 (1975); G. Galavotti and M. Pulvirenti, Commun. Math. Phys. **46**, 1 (1976).
[10] J. V. Pulè and A. Verbeure, J. Math. Phys. **18**, 826 (1977).
[11] M. Takesaki, *Tomita's Theory of Modular Hilbert Algebras and its Applications*, Lecture Notes in Mathematics 128 (Springer-Verlag, Berlin, 1970).
[12] J. Naudts and A. Verbeure, J. Math. Phys. **17**, 419 (1976).
[13] M. A. Rieffel and A. Van Daele, Pac. J. Math. **69**, 187 (1976).

Kinetic potentials in quantum mechanics

Richard L. Hall
Department of Mathematics, Concordia University, 1455 de Maisonneuve, Montreal, Quebec, Canada H3G 1M8

(Received 8 February 1984; accepted for publication 13 April 1984)

Suppose that the Hamiltonian $H = -\Delta + vf(r)$ represents the energy of a particle which moves in an attractive central potential and obeys nonrelativistic quantum mechanics. The discrete eigenvalues $E_{nl} = F_{nl}(v)$ of H may be expressed as a Legendre transformation $F_{nl}(v) = \min_{s>0}(s + v\bar{f}_{nl}(s))$, $n = 1,2,3,\ldots$, $l = 0,1,2,\ldots$, where the "kinetic potentials" $\bar{f}_{nl}(s)$ associated with $f(r)$ are defined by $\bar{f}_{nl}(s) = \inf_{D_{nl}} \sup_{\psi \in D_{nl}} \int \psi(\mathbf{r}) f([\psi, -\Delta\psi)/s]^{1/2} r)\psi(\mathbf{r}) d^3\mathbf{r}$, and D_{nl} is an n-dimensional subspace of $L^2(\mathbb{R}^3)$ labeled by $Y_l^m(\theta,\phi)$, $m = 0$, and contained in the domain $\mathscr{D}(H)$ of H. If the potential has the form $f(r) = \Sigma_{i=1}^N g^{(i)}(f^{(i)}(r))$ then in many interesting cases it turns out that the corresponding kinetic potentials can be closely approximated by $\Sigma_{i=1}^N g^{(i)}(\bar{f}_{nl}{}^{(i)}(s))$. This nice behavior of the kinetic potentials leads to a constructive global approximation theory for Schrödinger eigenvalues. As an illustration, detailed recipes are provided for arbitrary linear combinations of power-law potentials and the log potential. For the linear plus Coulomb potential and the quartic anharmonic oscillator the approximate eigenvalues are compared to accurate values found by numerical integration.

PACS numbers: 05.30.Fk, 05.30.Jp, 03.65.Ge

I. INTRODUCTION

The term "kinetic potential" was introduced[1] in 1983 and is a shortened version of the more explanatory name "minimum mean isokinetic potential." This concept arose from our geometrical theory[1] of energy trajectories in quantum mechanics. The main idea is as follows. We consider the Hamiltonian

$$H = -\Delta + vf(r), \quad r = |\mathbf{r}|, \tag{1.1}$$

which represents a particle moving in an attractive central potential $vf(r)$ with positive coupling constant v. We suppose that for v sufficiently large there exist discrete eigenvalues E_{nl} for H in some suitable domain $\mathscr{D}(H) \subset L^2(\mathbb{R}^3)$, where l is the angular-momentum quantum number and n is a radial quantum number: this is guaranteed, for example, when $f(r)$ is monotone increasing on $(0,\infty)$. In this notation for the eigenvalues we have $E_{nl} \leqslant E_{kl}$, $k > n$, $n = 1,2,3,\ldots$, and therefore each of the eigenvalues so labeled has degeneracy precisely $2l + 1$. Hence, for each v sufficiently large, we can write

$$E_{nl} = F_{nl}(v), \tag{1.2}$$

where the graphs $(v, F_{nl}(v))$ are called "energy trajectories" of the potential $f(r)$. Thus the potential with shape f gives rise to a family of trajectory functions $\{F_{nl}\}$.

The kinetic potentials $\bar{f}_{nl}(s)$ emerge when the constraint $(\psi, -\Delta\psi)/(\psi,\psi) = s > 0$ is applied during the optimization process leading to the eigenvalues of H: the subset of $\mathscr{D}(H)$ identified by this side condition is not a subspace of the Hilbert space $L^2(\mathbb{R}^3)$. This fact will become important when we proceed to the higher kinetic potentials in Sec. II below. For the ground state, however, we can immediately define the kinetic potential $\bar{f}(s)$ associated with the potential shape $f(r)$ by the equation

$$\bar{f}(s) = \inf_{\substack{\psi \in \mathscr{D} \\ \|\psi\| = 1 \\ (\psi, -\Delta\psi) = s}} (\psi, f\psi). \tag{1.3}$$

It follows from the variational principle that the ground-state eigenvalue is given by

$$E = F(v) = \min_{s>0} [s + v\bar{f}(s)]. \tag{1.4}$$

In the case that $\bar{f}(s)$ is smooth we obtain the following parametric equations for the energy trajectory $F(v)$:

$$E = F(v) = s + v\bar{f}(s); \quad v^{-1} = -\bar{f}'(s), \quad s > 0. \tag{1.5}$$

Thus the trajectory function F has been reached from the potential shape f in two stages: $f \to \bar{f} \to F$; the second stage which transforms the graph $(s, \bar{f}(s))$ to the graph $(v, F(v))$ is essentially a Legendre transformation.[2] We shall extend definition (1.3) to cover the higher eigenvalues in Sec. II. An elementary illustration is provided by the harmonic oscillator $H = -\Delta + vr^2$ for which we find

$$f(r) = r^2, \quad \bar{f}_{nl}(s) = (2n + l - \tfrac{1}{2})^2 s^{-1},$$
$$F_{nl}(v) = (4n + 2l - 1)v^{1/2}. \tag{1.6}$$

The reason we use kinetic potentials is that they have some very nice properties under certain transformations of the potential, such as convex transformations and linear combinations. This simple behavior on the kinetic energy hypersurface allows us to construct a global approximation theory for Schrödinger operators. A geometrical theory based on these concepts has been developed in a sequence of four articles which we shall hereafter refer to, respectively, as I (Ref. 3), II (Ref. 3), III (Ref. 3), and IV (Ref. 1). The main

 0022-2488/84/092708-08$02.50

thrust of the work to date has been to do with applications to the N-identical particle problem whose energy is related to the single-particle spectrum via the necessary permutation symmetry of the N-particle states. In the present article we focus our attention on the basic problem of a single particle in a central potential.

We present two principal theoretical points and then consider two examples in some detail, the linear plus Coulomb potential and the anharmonic oscillator potential. We are able to offer a sounder foundation for the higher kinetic potentials than we were able to present in IV. It is likely that the theory could eventually be used to support parts of conventional operator theory. However, in this article we continue to make safe assumptions about the potentials and to rely on the standard theory to justify what we do: the emphasis is still on concrete results and examples for which all the details can be worked out. The main novelty of the present paper is our discovery that kinetic potentials are almost additive. This notion is captured by the following relations:

$$f(r) = Af^{(1)}(r) + Bf^{(2)}(r), \tag{1.7}$$

$$\bar{f}_{nl}(s) \sim A\bar{f}^{(1)}_{nl}(s) + B\bar{f}^{(2)}_{nl}(s). \tag{1.8}$$

Although no such simple approximation as (1.8) holds for the energy trajectory functions $F_{nl}(v)$ themselves, they are easily obtained from the kinetic potentials $\bar{f}_{nl}(s)$ by the Legendre transformation (1.5). The relation (1.8) extends to a larger sum of terms and also to integrals. As the vector $(A, B,...)$ of coefficients approaches a vector with only one nonzero entry, "$\sim$" approaches "$=$"; in all cases where $n = 1$, "$\sim$" becomes "$\geqslant$" (leading to a lower energy bound for the bottom of the spectrum in each angular-momentum subspace). In other cases the approximation turns out to be consistently good and therefore very useful, as the examples will show.

Every numerical result in this article can be found by integrating Schrödinger's equation directly with the aid of a computer. In fact we have had to master this art in order to be able to test the quality of our approximations. For the sort of regular problem which we are considering the computer gives very reliable and reasonably fast results. This competition from the machine provides strong selection pressure guiding our analytical work in the direction of simplicity: if our results are not simple and general, it may be more effective to spend our time writing a computer program to take care of the problem. Actually, for this article, we found our energy bounds and other estimates of great help in guiding the computer program in its numerical search for the eigenvalues. Our principal conclusion after our experience with this particular combination of analytical and numerical techniques is that, even in the presence of a friendly computer, there is still nothing quite so nice as a good formula.

II. KINETIC POTENTIALS

We first restrict the shape $f(r)$ of the potentials so that our problems are comfortably within the scope of standard nonrelativistic quantum mechanics. We refer the reader to the textbooks by Prugovečki,[4] Reed and Simon,[5] and Thirring,[6] and in particular to Reed–Simon Chap. XIII. We consider the Hamiltonian

$$H = -\Delta + vf(r), \quad r = |\mathbf{r}|, \quad v > 0, \tag{2.1}$$

with the hypotheses

$$f'(r) > 0, \quad r > 0,$$

$$|r^2 f(r)| \to 0, \quad r \to 0. \tag{2.2}$$

Under these conditions we know that H is essentially self-adjoint on some domain $\mathscr{D}(H) \subset L^2(\mathbb{R}^3)$ and also that discrete eigenvalues always exist for sufficiently large values of the coupling constant v. The implied differentiability of the potential shape $f(r)$ is a convenience which allows us to use calculus.

We now suppose that v is constant and sufficiently large to guarantee at least n discrete eigenvalues at the bottom of the spectrum of H. We work with the min–max characterization of the eigenvalues (see Reed–Simon,[5] Volume IV, p. 75, and Thirring,[6] Vol. 3, Sec. 3.5.21) and show how the kinetic potentials $\bar{f}_n(s)$ emerge from the optimization process. Later, we can repeat the argument inside the angular-momentum subspace of $L^2(\mathbb{R}^3)$ labeled by the spherical harmonic $Y_l^m(\theta,\phi)$ and $m = 0$ to obtain the corresponding kinetic potential $\bar{f}_{nl}(s)$. Once the definition of these objects and some of their elementary properties have been established we shall be able to work constructively and use the known exact trajectory functions $F_{nl}(v)$ to derive the corresponding kinetic potentials $\bar{f}_{nl}(s)$: these will then become the building blocks for our approximation theory.

Suppose $\{\psi_1,\psi_2,...,\psi_n\}$ is a set of n linearly independent vectors in $\mathscr{D}(H)$, then we define

$$D_n = \text{span}\{\psi_1,\psi_2,...,\psi_n\}, \quad D_n \subset \mathscr{D}(H), \tag{2.3}$$

$$(\hat{\sigma}\psi_i)(\mathbf{r}) = \psi_i(\mathbf{r}/\sigma), \quad i = 1,2,...,n, \tag{2.4}$$

$$\hat{\sigma}D_n = \text{span}\{\hat{\sigma}\psi_1,\hat{\sigma}\psi_2,...,\hat{\sigma}\psi_n\}, \tag{2.5}$$

$$\mathscr{D}_n = \bigcup_{\sigma>0} \{\hat{\sigma}D_n\}. \tag{2.6}$$

We note that although $\mathscr{D}_n$ is the union of a large number of n-dimensional linear spaces, it is not itself a linear space. By the min–max (Rayleigh–Ritz) principle we have

$$F_n(v) = \inf_{D_n} \sup_{\substack{\psi\in D_n \\ \|\psi\|=1}} (\psi,H\psi). \tag{2.7}$$

A possible route for the process (2.7) is given by

$$F_n(v) = \inf_{D_n} \inf_{\sigma>0} \sup_{\substack{\psi\in\hat{\sigma}D_n \\ \|\psi\|=1}} (\psi,H\psi), \tag{2.8}$$

i.e.,

$$F_n(v) = \inf_{D_n} \inf_{s>0} \sup_{\substack{\psi\in\mathscr{D}_n \\ \|\psi\|=1 \\ (\psi,-\Delta\psi)=s}} (\psi,H\psi). \tag{2.9}$$

In (2.9) ψ no longer explores a linear space but rather that part of $\mathscr{D}_n$ which satisfies $\|\psi\| = 1$ and also $(\psi, -\Delta\psi) = s$. The danger here is that the sup may fall below the nth eigenvalue; however, this does not happen. We can see that this is so by the following observations: setting $(\psi, -\Delta\psi) = s$ amounts to choosing a scale σ for the functions which are used in the approach to the sup of $(\psi,H\psi)$; this is a very

complicated way to choose a scale, but all scales are included in the union of linear spaces $\mathscr{D}_n$. We now make the following step:

$$F_n(v) = \inf_{s>0} \inf_{D_n} \sup_{\substack{\psi \in \mathscr{D}_n \\ \|\psi\| = 1 \\ (\psi, -\Delta\psi) = s}} (\psi, H\psi). \tag{2.10}$$

The exchange of the inf's from (2.9) to (2.10) is allowed because, in either order, the overall minimization explores the same set of all n-dimensional subspaces of $L^2(\mathbb{R}^3)$ which are contained in $\mathscr{D}(H)$. Hence we can write $F_n(v)$ in the form

$$F_n(v) = \inf_{s>0} (s + v\bar{f}_n(s)), \tag{2.11}$$

where

$$\bar{f}_n(s) = \inf_{D_n} \sup_{\substack{\psi \in \mathscr{D}_n \\ \|\psi\| = 1 \\ (\psi, -\Delta\psi) = s}} (\psi, f\psi). \tag{2.12}$$

An important point about this definition of the kinetic potential is that $\bar{f}_n(s)$ is not labeled by the value of v: the reason for this is that $(\psi, -\Delta\psi)/\|\psi\|^2 = s$ is held constant under the sup's in (2.9) and (2.10); the partial trade-off between kinetic and potential energy concerns only the *shape* of $f(r)$.

An equivalent definition to (2.12), which may prove to be more convenient to work with, is obtained in the following way. We first note that the operators $-\Delta$ and f scale according to the equations

$$(\hat{\sigma}\psi, -\Delta\hat{\sigma}\psi)/(\hat{\sigma}\psi, \hat{\sigma}\psi) = \sigma^{-2}(\psi, -\Delta\psi)/(\psi, \psi), \tag{2.13}$$

and

$$\frac{(\hat{\sigma}\psi, f\hat{\sigma}\psi)}{(\hat{\sigma}\psi, \hat{\sigma}\psi)} = \|\psi\|^{-2} \int \psi(\mathbf{r}) f(\sigma r) \psi(\mathbf{r}) d^3\mathbf{r}. \tag{2.14}$$

Consequently we have from (2.12)

$$\bar{f}_n(s) = \inf_{D_n} \sup_{\substack{\psi \in D_n \\ \|\psi\| = 1}} \int \psi(\mathbf{r}) \times f\left(\left[\frac{(\psi, -\Delta\psi)}{s}\right]^{1/2} r\right) \psi(\mathbf{r}) d^3\mathbf{r}. \tag{2.15}$$

We exploited a form like (2.15) in paper IV [Eq. (3.8)] to obtain ground-state upper bounds by choosing a particular shape for ψ: the subsequent minimization with respect to s in this case is equivalent to a minimization of the energy expectation with respect to the scale of the wave function.

As we mentioned above, we get the kinetic potentials $\bar{f}_{nl}(s)$ by repeating all the above steps inside the subset $\mathscr{D}_l(H)$ of $\mathscr{D}(H)$ which is also contained in the subspace of $L^2(\mathbb{R}^3)$ defined by the projector corresponding to the spherical harmonic $Y_l^m(\theta, \phi)$ with, for example, $m = 0$. For each value of v (sufficiently large) the equation

$$F_{nl}(v) = \min_{s>0} (s + v\bar{f}_{nl}(s)), \tag{2.16}$$

then gives back the eigenvalue $F_{nl}(v)$ which has degeneracy exactly $2l + 1$ because, of course, the eigenvalues obtained by the variational method satisfy $F_{nl}(v) \geqslant F_{nl}(v)$, for $n > k \geqslant 1$. By using calculus we then obtain from (2.16) the Legendre transformations

$$F_{nl}(v) = s + v\bar{f}_{nl}(s), \quad v^{-1} = -\bar{f}'_{nl}(s), \tag{2.17}$$

and

$$\bar{f}_{nl}(s) = F'_{nl}(v), \quad s = F_{nl}(v) - vF'_{nl}(v). \tag{2.18}$$

In the cases where we already know the trajectory functions $F_{nl}(v)$, rather than use the general definition corresponding to (2.12), we instead solve (2.18) to find the kinetic potentials $\bar{f}_{nl}(s)$. For many useful potentials, this can be done exactly.

This completes the foundation work for the higher kinetic potentials which was only briefly glossed over in Sec. 6 of paper IV. In the present article we have used the trajectory functions $F(v)$ rather than the functions $G(u) = uF(1/u)$ which were employed in paper IV to discuss convexity. We refer the reader to the earlier article for a general discussion of ordering, scaling, and convexity properties, and for a table of kinetic potentials.

III. SOME ESTABLISHED PROPERTIES OF KINETIC POTENTIALS

We present here a very brief summary from paper IV of the results which we shall need concerning kinetic potentials. The potentials $f(r)$ will be constructed in various ways but we shall always assume that they satisfy the following conditions:

$$\begin{aligned} &f'(r) > 0, \quad r > 0, \\ &|r^2 f(r)| \to 0, \quad r \to 0. \end{aligned} \tag{3.1}$$

The smoothness and nice behavior at the origin are analytical conveniences. The essential restriction in the present work is that $f(r)$ is monotone increasing.

We first gather in one theorem the principal scaling, convexity, and ordering results. It has also been established that $F(v)$ is concave and that $G(u) = uF(1/u)$ is monotone increasing and concave, but we shall not need to use these facts, nor the various results concerning variational approximations, in the present article.

Theorem 3.1: We suppose that the kinetic potentials corresponding to the potential $f(r)$ are given by $\bar{f}_{nl}(s)$.

(a) $A + Bf(r/a) \to A + B\bar{f}_{nl}(sa^2)$, $B > 0$, $a > 0$.

(b) $\bar{f}_{nl}(s)$ is monotone increasing and convex on $(0, \infty)$.

(c) $f^{(1)} < f^{(2)} \to \bar{f}^{(1)} < \bar{f}^{(2)} \to F^{(1)} < F^{(2)}$.

the ordering result (c) means that variational arguments do yield the results we expect to get as we proceed from potential to kinetic potential to trajectory function: the functional inequalities, of course, are meaningful only over common domains.

The second theorem summarizes the key results which allow us to approximate the spectra of Schrödinger Hamiltonians in which the potential has been composed out of "soluble" potentials by convex transformations and linear combinations. When there is more than one term ($N \geqslant 2$) we have been able to establish a bound (in fact, a lower bound) on the exact kinetic potentials only when they correspond to the bottom ($n = 1$) of the energy spectrum in each given angular-momentum subspace (given l). One of the main objectives of the present article is to attempt to transcend this limitation.

Theorem 3.2: Suppose that

$$f(r) = \sum_{i=1}^{N} g^{(i)}(f^{(i)}(r)),$$

and

$$\Phi_{nl}(s) = \sum_{i=1}^{N} g^{(i)}(\bar{f}_{nl}^{(i)}(s)).$$

(a) If the $\{g^{(i)}\}$ are all convex and if $n = 1$ or $N = 1$ we have $\bar{f}_{nl}(s) \geqslant \Phi_{nl}(s)$.

(b) If $N = 1$ and g is concave, then $\bar{f}_{nl}(s) \leqslant \Phi_{nl}(s)$.

It will be noticed that Theorem 3.2 yields no upper bound in the case of more than one term ($N \geqslant 2$). This is the reason that variational upper bounds are studied in the framework of kinetic potentials in paper IV. However, we shall not discuss this topic in the present article.

The potentials we propose to consider as illustrations of the theory will be composed out of power-law potentials and the log potential; we therefore collect here the established results[1] which we shall need for these potentials. The power-law potentials are defined by

$$f(r) = \mathrm{sgn}(q) r^q, \quad q \geqslant -1, \quad q \neq 0. \tag{3.2}$$

By scaling arguments one finds for these potentials that the eigenvalues of $H = -\Delta + vf(r)$ are given by

$$F_{nl}(v) = v^{2/(2+q)} E(q, n, l), \quad E(q, n, l) = F_{nl}(1). \tag{3.3}$$

In terms of these $v = 1$ energies $E(q, n, l)$, whose meaning is perhaps more immediate than the various coefficients of paper IV, we can write the power-law kinetic potentials in the form

$$\bar{f}_{nl}(s) = (2/q)|qE(q, n, l)/(q+2)|^{(q+2)/2} s^{-q/2}, \quad s > 0. \tag{3.4}$$

Similarly, for the log potential

$$f(r) = \ln(r), \tag{3.5}$$

we have found[1,7] that the eigenvalues of $H = -\Delta + vf(r)$ are given by

$$F_{nl}(v) = vE(L, n, l) - \tfrac{1}{2} v \ln(v), \quad E(L, n, l) = F_{nl}(1), \tag{3.6}$$

and it follows [from Eq. (2.18)] that the kinetic potentials are given by

$$\bar{f}_{nl}(s) = E(L, n, l) - \tfrac{1}{2} \ln(2se), \quad s > 0, \quad e = \exp(1). \tag{3.7}$$

We now have the theorems we shall need, and also the kinetic potentials for the potential components expressed in terms of the elementary eigenvalues $E(q, n, l)$ and $E(L, n, l)$ of the power-law and log potentials; tables of these eigenvalues for $q = 1$ and 4 and for the log potential are given in the Appendix.

Eventually we expect to have an independent theory of kinetic potentials which would start from Eq. (2.15) as a definition. The object which we shall have to study is the functional $\Gamma: L^2(\mathbb{R}^3) \to \mathbb{R}$ which for each fixed f and s is given by

$$\Gamma(\psi) = \int \psi(\mathbf{r}) f([(\psi, -\Delta\psi)/s]^{1/2} r) \psi(\mathbf{r}) d^3\mathbf{r}. \tag{3.8}$$

The kinetic potentials $\bar{f}_{nl}(s)$ are the stationary values of $\Gamma(\psi)$. We make two comments about $\Gamma(\psi)$ with potentials of type (3.1) for future reference: (i) $\Gamma(\psi)$ is invariant under spatial scale changes; and (ii) the spectrum of $\Gamma(\psi)$ appears always to have an infinite number of discrete stationary values. The Hulthén potential $f(r) = -(e^r - 1)^{-1}$ illustrates this very well. Suppose we consider the Hamiltonian

$$H = -\Delta - v(e^r - 1)^{-1}. \tag{3.9}$$

The discrete S-state spectrum of H is well known[8] and is given by

$$F_{n0}(v) = -(v - n)^2/4n^2, \quad v \geqslant n, \quad n = 1,2,3,\ldots. \tag{3.10}$$

We find by using Eq. (2.18) that the corresponding kinetic potentials are given by

$$\bar{f}_{n0}(s) = -\{(1 + 4s)^{1/2} - 1\}/2n, \quad s > 0. \tag{3.11}$$

The critical coupling constants $v_{n0} = n$ are buried in the shapes of the kinetic potentials $\bar{f}_{n0}(s)$ which are defined for all $s > 0$: the quantity $F_{n0}(v) = \min\{s + v\bar{f}_{n0}(s)\}$ only exists for $v \geqslant n$. These considerations indicate that kinetic potentials have even more nice properties which are not revealed by examples involving only the power-law and log interactions.

IV. ALMOST ADDITIVITY AND THE SUM APPROXIMATION

Suppose we can solve the Schrödinger eigenvalue problem for $H = -\Delta + vf^{(i)}(r)$, with $i = 1,2,\ldots,N$. What can we then say about the spectrum of $H = -\Delta + f(r)$, where

$$f(r) = \sum_{i=1}^{N} A^{(i)} f^{(i)}(r), \quad A^{(i)} > 0. \tag{4.1}$$

Our claim is that for the potential (4.1) it is a good approximation simply to add the component kinetic potentials, that is to say, to use the *sum approximation*

$$\bar{f}_{nl}(s) \sim \sum_{i=1}^{N} A^{(i)} \bar{f}_{nl}^{(i)}(s). \tag{4.2}$$

We know from Theorem 3.2 (a) that if we look at the bottom of the spectrum ($n = 1$) in a particular angular-momentum subspace (a given l), then Eq. (4.2) yields a lower bound (i.e., "$\sim$" = "$\geqslant$"). The relation becomes an equality, of course, whenever there is only one term in the sum. The Legendre transformation (2.17) which converts the kinetic potential into an energy trajectory corresponds to a minimization with respect to scale: consequently the approximate energy trajectories which derive from (4.2) obey the same scaling laws as do the (unknown) exact trajectories; specific examples are discussed in Secs. V and VI.

The form of the sum approximation (4.2) is such that if it is good for any two potentials from a certain collection, then it will be good for the sum over any N of these potentials. Let us therefore look first at an example of the case $N = 2$ in which the details can be written down simply and explicitly and analyzed. We study the example

$$f(r) = -A/r + Br^2, \quad A \geqslant 0, \quad B \geqslant 0. \tag{4.3}$$

This potential increases too fast for large r to represent the central part of a quark–quark interaction, but it is a potential of that general type. We already have the general form of the kinetic potentials for the components in Eq. (3.4) above. The eigenvalues $E(-1, n, l) = -1/4(n + l)^2$ and $E(2, n, l) = (4n + 2l - 1)$ are the well-known results for the hydrogen atom and the spherical oscillator. Consequently,

we can immediately write down the details of our approximation as follows:

$$\text{Coulomb component: } -1/r \to -s^{1/2}/(n+l), \quad (4.4)$$

$$\text{Oscillator component: } r^2 \to (2n+l-\tfrac{1}{2})^2 s^{-1}. \quad (4.5)$$

Hence our sum approximation in this example becomes

$$\bar{f}_{nl}(s) \sim -As^{1/2}/(n+l) + B(2n+l-\tfrac{1}{2})^2 s^{-1}. \quad (4.6)$$

Before we try to evaluate the numerical quality of the eigenvalues which result from (4.6) we shall first use our geometrical theory to find analytical bounds on $\bar{f}_{nl}(s)$. We do this by employing the method of potential envelopes introduced in paper I (see Ref. 3). This method has been condensed into the cases $N=1$ of the present summary Theorem 3.2. We first set up dual representations for $f(r)$ as a convex transformation of $(-1/r)$ and as a concave transformation of (r^2). Thus we have

$$f(r) = A(-1/r) + B(-1/r)^{-2}$$
$$= -A/(r^2)^{1/2} + B(r^2). \quad (4.7)$$

If we now apply Theorem 3.2 we find

$$-As^{1/2}/(n+l) + B(n+l)^2 s \leqslant \bar{f}_{nl}(s)$$
$$\leqslant -As^{1/2}/(2n+l-\tfrac{1}{2}) + B(2n+l-\tfrac{1}{2})^2 s. \quad (4.8)$$

The nice form of (4.8) is not accidental: a full explanation may be found in paper IV, Sec. 8A. This result by the envelope method suggests that we consider approximations of the form

$$\bar{f}_{nl}(s) \sim -As^{1/2}/\nu_1 + B(\nu_2 s^{1/2})^2, \quad (4.9)$$

for various ν_1 and ν_2 [including, of course, $\nu_1 = \nu_2 = \nu$, where ν lies between $(n+l)$ and $(2n+l-\tfrac{1}{2})$]. In order to get the correct result whenever A or B is zero, one is led back again in this way to the sum approximation (4.6). We have gone into these details in order to make clear how the different methods interact with the problem. Later in this section we shall establish general formulas which summarize the application of the sum approximation to arbitrary combinations of power-law potentials and the log potential. The envelope method which we have just used as a tool will not be treated in general because it has already been applied to many different situations in papers I, II, and IV.

In order to obtain the approximate eigenvalues from the kinetic potentials (4.6) we must apply the Legendre transformation (2.17). Some numerical results for $A = B = 1$, along with accurate values (in parentheses) obtained by "shooting methods" on a microcomputer, are as follows: $E_{10} = 1.706$ (1.785), $E_{20} = 6.048$ (6.029), $E_{11} = 4.193$ (4.229), $E_{33} = 16.512$ (16.533), and $E_{55} = 28.619$ (28.634). This data is typical of what we have found for the sum approximation: the error is at worst a few percent and decreases rapidly with increasing n and l; when the potential components are less "different," the errors turn out to be much smaller, as we shall see in Secs. V and VI. The bounds (4.8) provided by the envelope method were very useful in automating the computer search for the accurate eigenvalues needed for this comparison.

We now organize the details of the application of the sum approximation (4.2) to a wider class of potentials. More interesting specific applications will then be considered in Secs. V and VI. The potential we consider is given by

$$f(r) = \sum_q A^{(q)} \operatorname{sgn}(q) r^q + B \ln(r),$$
$$q \geqslant -1, \quad A^{(q)} \geqslant 0, \quad B \geqslant 0. \quad (4.10)$$

We substitute Eqs. (3.4) and (3.7) in Eq. (4.2), apply the Legendre transformation (2.17), and simplify the resulting expressions. If we make the convenient change of variables $t = s^{-1/2}$ we find finally that the eigenvalue $E(n,l)$ of the operator $H = -\Delta + vf(r)$ is approximated by

$$Ev^{-1} = \sum_q A^{(q)} \left| \frac{qE^{(q)}}{(q+2)} \right|^{(q+2)/2} \frac{t^q(2+q)}{q}$$
$$+ B(\epsilon + \tfrac{1}{2}\ln(t^2/2)),$$

where

$$v^{-1} = \sum_q A^{(q)} \left| \frac{qt^2 E^{(q)}}{(q+2)} \right|^{(q+2)/2} + \frac{Bt^2}{2}, \quad t>0, \quad (4.11)$$

and we have written for the pure power-law and log potentials

$$E^{(q)} = E(q,n,l) \quad \text{and} \quad \epsilon = E(L,n,l).$$

Tables of the eigenvalues $E(q,n,l)$ for $q=1$ and 4, and $E(L,n,l)$ are given in the Appendix. The simple formulas (4.11) are parametric equations for the energy trajectories $(v, F_{nl}(v))$ in terms of the parameter $t>0$; they are the equations to use in a typical application of the sum approximation to potentials in the class (4.10). Since the errors are usually too small to show on a graph, in the present paper we shall continue to choose values for v and the B and the $\{A^{(q)}\}$, and compare our results with accurate values in tables.

The sum approximation evolved as a result of the interplay between the potential-envelope method of paper I and the linear-combinations method of paper III, along the lines of the illustration discussed at the start of this section. Our original approach to the problem (paper III) was to consider the following identity satisfied by the Hamiltonian $H = -\Delta + f^{(1)}(r) + f^{(2)}(r)$:

$$H = w[-\Delta + w^{-1}f^{(1)}(r)]$$
$$+ (1-w)[-\Delta + (1-w)^{-1}f^{(2)}(r)],$$
$$1 > w > 0. \quad (4.12)$$

For each fixed w the Hamiltonian therefore has the form

$$H = H^{(1)} + H^{(2)}. \quad (4.13)$$

The question of the spectrum of a sum of operators is, of course, very old. Two important contributions to this topic are the articles of Weyl[9] and Fan.[10] A more recent discussion of Weyl's Theorem in the context of Schrödinger operators may be found in the book by Weinstein and Stenger.[11] If the operators $H, H^{(1)}$, and $H^{(2)}$ each have at least n discrete eigenvalues at the bottom of their spectra, then we know from Weyl's theorem that

$$E_n \geqslant E_q^{(1)} + E_r^{(2)}, \quad q+r = n+1. \quad (4.14)$$

If we restrict the problem to an angular-momentum subspace (labeled by l), choose $n = q = r = 1$, and maximize the right-hand side of Eq. (4.14) with respect to the parameter w of Eq. (4.12), then we get the linear-combination method of paper III [this is included in Theorem 3.2 (a) of the present article]. However, if we choose $n>1$, and repeat the same

application of Weyl's Theorem, we find that the resulting lower bounds are very weak. In terms of the present discussion, our sum approximation can be thought of as a procedure which is based on the initial approximation $E_n \sim E_n^{(1)} + E_n^{(2)}$ and the observation that this yields good results provided that it is optimized with respect to the convex-combination parameter w, $1 > w > 0$, of Eq. (4.12).

V. THE COULOMB PLUS LINEAR POTENTIAL

We consider the potential

$$f(r) = -A/r + Br, \quad A \geqslant 0, \quad B \geqslant 0. \tag{5.1}$$

This potential is of practical interest because it represents the approximate shape of the central part of the quark–quark interaction. For the Hamiltonian $H = -\Delta + f(r)$, both the exact energies $E_{nl}(A, B)$ and the approximations given by the sum approximation are easily shown to obey the same scaling laws:

$$E_{nl}(A, B) = \sigma^2 E_{nl}(A/\sigma, B/\sigma^3), \quad \sigma > 0, \tag{5.2}$$

and, in particular with $\sigma = A$, one finds

$$E_{nl}(A, B) = A^2 E_{nl}(1, B/A^3). \tag{5.3}$$

Consequently it is sufficient for us to consider the eigenvalues $E_{nl}(1,\lambda)$ for $\lambda > 0$. Our recipe (4.11) for these eigenvalues becomes in this case

$$E_{nl}(1,\lambda) = -|E^{(-1)}|^{1/2} t^{-1} + 3\lambda\, |E^{(1)}/3|^{3/2} t,$$

where

$$1 = |E^{(-1)}|^{1/2} t + \lambda\, |E^{(1)}/3|^{3/2} t^3, \tag{5.4}$$

$E^{(-1)} = -[4(n + l)^2]^{-1}$, and $E^{(1)} = E(1, n, l)$ are given in Table III in the Appendix. Equation (5.4) is useful for plotting the graph (E,λ) because E and λ are given explicitly in terms of the parameter $t > 0$. For numerical values the simplest procedure is to solve the cubic equation for t by Newton's method and substitute this into the equation for E. Some values we find for $E_{nl}(1,\lambda)$ are shown in Table I together with accurate values (in parentheses) which we have found by numerical integration.

VI. THE QUARTIC ANHARMONIC OSCILLATOR

There has been a long tradition of using the anharmonic oscillator to test approximation methods in quantum mechanics. Consequently the literature on this problem is now vast[12]: for a careful summary of the mathematical situation and a good bibliography we recommend the recent review article by Simon.[13] If we use the potential

$$f(r) = Ar^2 + Br^4, \tag{6.1}$$

then elementary scaling arguments yield the following equation for the eigenvalues $E_{nl}(A, B)$ of the Hamiltonian $H = -\Delta + f(r)$:

$$E_{nl}(A, B) = \sigma E_{nl}(A/\sigma^2, B/\sigma^3), \quad \sigma > 0. \tag{6.2}$$

Our approximate eigenvalues [given by the sum approximation (4.11)] obey the same scaling law. If we choose $\sigma^2 = A$, then we obtain the special case

$$E_{nl}(A, B) = A^{1/2} E_{nl}(1, BA^{-3/2}). \tag{6.3}$$

Because of Eq. (6.3) we need only consider the eigenvalues $E_{nl}(1,\lambda)$ for $\lambda > 0$. The general equations (4.11) for the sum approximation become in the present example

$$E_{nl}(1,\lambda) = 2[\tfrac{1}{2} E^{(2)}]^2 t + 3\lambda\, [2E^{(4)}/3]^3 t^2/2,$$

where

$$1 = [\tfrac{1}{2} E^{(2)}]^2 t^2 + \lambda\, [2E^{(4)}/3]^3 t^3, \quad t > 0, \tag{6.4}$$

$E^{(2)} = (4n + 2l - 1)$, the pure quartic eigenvalues $E^{(4)} = E(4, n, l)$ are given in Table IV in the Appendix, and, for convenience, we have used t in Eq. (6.4) in place of the t^2 of Eq. (4.11). In this problem it is possible to solve Eqs. (6.4) and obtain the following explicit formula for λ in terms of $E = E_{nl}(1,\lambda)$:

$$\lambda = \frac{P^4 3^3 [P^2 - \{2E - (4E^2 - 3P^2)^{1/2}\}^2]}{Q^3 4^3 [2E - \{4E^2 - 3P^2\}^{1/2}]^3}, \tag{6.5}$$

where $E \geqslant P = E(2, n, l) = (4n + 2l - 1)$ and $Q = E(4, n, l)$. This is the nearest we have come to an explicit solution to the quartic anharmonic-oscillator problem. In paper III we proved for the ground state that formula (6.5) with $n = 1$ and $l = 0$ yields (by inversion) $E_{nl}(1,\lambda)$ for a given λ with error always less than 1%. Actually, the present more general result typically has far smaller errors: this fact is demonstrated in Table II.

VII. CONCLUSION

The energies of conservative quantum-mechanical systems may be characterized in terms of the extrema of the Rayleigh quotient $(\psi, H\psi)/(\psi,\psi)$ of the Hamiltonian $H = -\Delta + vf(r)$. In the abstract theory of this problem,[4–6] the potential $vf(r)$ is regarded as a perturbation of the kinetic-energy operator $-\Delta$. What we have tried to do is to look at

TABLE I. Comparison of some of the eigenvalues $E_{nl}(1,\lambda)$ of the Hamiltonian $H = -\Delta - 1/r + \lambda r$ given by the kinetic-potential sum approximation, and accurate values (in parentheses) obtained by numerical integration.

E_{nl}	$\lambda = 0.01$	$\lambda = 1$	$\lambda = 100$
E_{10}	−0.223 (−0.221)	1.361 (1.398)	46.189 (46.402)
E_{30}	0.148 (0.142)	5.059 (5.033)	116.83 (116.74)
E_{50}	0.295 (0.286)	7.615 (7.575)	169.64 (169.50)
E_{12}	0.101 (0.102)	3.842 (3.851)	89.672 (89.850)
E_{32}	0.250 (0.251)	6.563 (6.572)	146.58 (146.61)
E_{52}	0.366 (0.366)	8.827 (8.829)	194.41 (194.43)
E_{14}	0.205 (0.206)	5.513 (5.517)	123.54 (123.53)
E_{34}	0.324 (0.326)	7.880 (7.888)	173.79 (173.78)
E_{54}	0.427 (0.428)	9.960 (9.966)	218.06 (218.06)

TABLE II. Comparison of some of the eigenvalues $E_{nl}(1,\lambda)$ of the Hamiltonian $H = -\Delta + r^2 + \lambda r^4$ given by the kinetic-potential sum approximation, and accurate values (in parentheses) obtained by numerical integration.

E_{nl}	$\lambda = 0.01$	$\lambda = 1$	$\lambda = 100$
E_{10}	3.035 (3.036)	4.640 (4.649)	17.827 (17.830)
E_{30}	11.435 (11.426)	23.323 (23.298)	99.040 (99.033)
E_{50}	20.235 (20.210)	47.007 (46.965)	204.82 (204.80)
E_{12}	7.148 (7.151)	12.474 (12.486)	50.691 (50.693)
E_{32}	15.756 (15.746)	34.229 (34.208)	147.32 (147.31)
E_{52}	24.742 (24.712)	59.836 (59.795)	262.46 (262.40)
E_{14}	11.332 (11.336)	21.583 (21.595)	90.094 (90.097)
E_{34}	20.134 (20.126)	45.733 (45.721)	198.49 (198.48)
E_{54}	29.297 (29.268)	73.064 (73.029)	322.04 (321.89)

this optimization process subject to the constraint $(\psi, -\Delta\psi)/(\psi,\psi) = s = \text{const}$. In order to do this we first have had to consider sets of states which are composed of unions of all "scales" of an n-dimensional subspace of $L^2(\mathbb{R}^3)$ contained in $\mathscr{D}(H)$. For potentials which are monotone increasing this eventually leads to the following two-stage formulation of the eigenvalue problem:

$$\bar{f}_{nl}(s) = \inf_{D_{nl}} \sup_{\substack{\psi \in D_{nl} \\ \|\psi\| = 1}} \int \psi(\mathbf{r}) f\left(\left[\frac{(\psi, -\Delta\psi)}{s}\right]^{1/2} r\right) \times \psi(\mathbf{r}) d^3\mathbf{r}, \tag{7.1}$$

$$F_{nl}(v) = \min_{s>0} (s + v\bar{f}_{nl}(s)), \tag{7.2}$$

where, as in the conventional theory, D_{nl} is an n-dimensional subspace of $P(l,0)L^2(\mathbb{R}^3)$ contained in $\mathscr{D}(H)$ and $P(l,0)$ is the projector corresponding to the spherical harmonic $Y_l^m(\theta,\phi)$ with $m = 0$. Thus the kinetic potentials $\bar{f}_{nl}(s)$ are generated from the extrema of the scale invariant functional $\Gamma(\psi)$, where

$$\Gamma(\psi) = \int \psi(\mathbf{r}) f\left(\left[\frac{(\psi, -\Delta\psi)}{s}\right]^{1/2} r\right) \psi(\mathbf{r}) d^3\mathbf{r}. \tag{7.3}$$

The reason that we go to the trouble of reformulating what is by now a classical problem of mathematical physics is that the kinetic potentials $\bar{f}_{nl}(s)$ appear to behave very nicely under transformations of the potential $f(r)$. A way of summarizing all the key constructive results of both paper IV and the present article is to write the relation

$$\sum_{i=1}^{N} g^{(i)}(f^{(i)}(r)) \to \sim \sum_{i=1}^{N} g^{(i)}(\bar{f}_{nl}{}^{(i)}(s)). \tag{7.4}$$

What we have done is to discover under what circumstances "$\sim$" represents a bound or a "good approximation." In all cases the approximate eigenvalue is obtained from the kinetic potential by the final minimization (7.2): this latter step can usually be carried out very simply and yields parametric equations for the energy trajectory whose graph is $(v, F_{nl}(v))$, $v > 0$.

Although we have so far adopted a constructive approach in this work and we have strived to produce detailed recipe solutions for large families of specific problems, there does appear to be good reason to believe that some aspects of abstract operator theory may benefit from an approach of the type we have used. The most interesting object for further study is probably the functional (7.3). Here $\Gamma(\psi)$ is scale invariant and captures the trade-off between the kinetic energy and the *shape* of the potential $vf(r)$. Meanwhile, the kinetic potentials $\bar{f}_{nl}(s)$ which derive from the "spectrum" of $\Gamma(\psi)$ are not labeled by the coupling constant v and yet the shapes of the $\bar{f}_{nl}(s)$ determine the energy trajectories $F_{nl}(v)$ via the Legendre transformation (7.2). When v is small H may have only a few or no discrete eigenvalues whereas there always appear to be an infinite number of discrete extrema at the bottom of the spectrum of $\Gamma(\psi)$. An example of this was discussed at the end of Sec. III above. These general observations along with the large number of interesting specific results so far obtained suggest that there may be more still to gain by the further study of this geometrical theory and in particular of kinetic potentials.

TABLE III. Eigenvalues $E(1, n, l)$ of the Hamiltonian $H = -\Delta + r$ obtained by numerical integration.

l	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0	2.338	4.088	5.521	6.787	7.944
1	3.361	4.884	6.208	7.406	8.515
2	4.248	5.630	6.869	8.010	9.077
3	5.051	6.332	7.505	8.597	9.627
4	5.794	6.999	8.117	9.168	10.166
5	6.493	7.637	8.709	9.724	10.692

TABLE IV. Eigenvalues $E(4, n, l)$ of the Hamiltonian $H = -\Delta + r^4$ obtained by numerical integration.

l	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0	3.800	11.645	21.238	32.099	43.981
1	7.108	16.033	26.350	37.774	50.127
2	10.842	20.643	31.615	43.568	56.370
3	14.923	25.472	37.036	49.485	62.716
4	19.301	30.506	42.614	55.528	69.166
5	23.941	35.734	48.344	61.695	75.723

ACKNOWLEDGMENT

The author would like to thank the Natural Sciences and Engineering Research Council Canada for partial financial support of this work by Grant No. A3438.

APPENDIX: SOME COMPONENT EIGENVALUES AND INEQUALITIES

Tables III, IV, and V exhibit the eigenvalues $E(q, n, l)$ of the Hamiltonian $H = -\Delta + r^q$ for the pure linear potential $q = 1$ and the quartic potential $q = 4$; and also the eigenvalues $E(L, n, l)$ for the Hamiltonian $H = -\Delta + \ln(r)$. These

TABLE V. Eigenvalues $E(L, n, l)$ of the Hamiltonian $H = -\Delta + \ln(r)$ obtained by numerical integration.

l	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0	1.044	1.847	2.290	2.596	2.830
1	1.641	2.151	2.491	2.746	2.949
2	2.013	2.387	2.663	2.880	3.059
3	2.284	2.580	2.810	2.999	3.159
4	2.497	2.742	2.940	3.107	3.251
5	2.673	2.881	3.055	3.205	3.335

numbers are needed by the sum approximation (4.11) whenever the potential contains the corresponding components. With the numerical solution of Schrödinger's equation it is very helpful if the approximate location of the eigenvalue is known. By using the method of potential envelopes [particularly paper IV, Eq. (8.4)] it is straightforward to establish the following inequalities which we have used to guide the computer in its search for the eigenvalues:

$$3[\tfrac{1}{2}(n+l)]^{2/3} \leqslant E(1,n,l) \leqslant 3\left[n+\tfrac{1}{2}l-\tfrac{1}{4}\right]^{2/3}, \tag{A1}$$

$$3[2n+l-\tfrac{1}{2}]^{4/3}2^{-2/3} \leqslant E(4,n,l), \tag{A2}$$

$$\ln(n+l)+\tfrac{1}{2}(1+\ln(2)) \leqslant E(L,n,l) \leqslant \ln(2n+l-\tfrac{1}{2}) + \tfrac{1}{2}(1+\ln(2)). \tag{A3}$$

For the Coulomb potential ($q=-1$) and the harmonic oscillator ($q=2$) we have the following well-known exact results:

$$E(-1,n,l) = -[2(n+l)]^{-2}, \tag{A4}$$

$$E(2,n,l) = [4n+2l-1]. \tag{A5}$$

[1] R. L. Hall, J. Math. Phys. **24**, 324 (1983) (paper IV).
[2] I. M. Gel'fand and S. V. Fomin, *Calculus of Variations* (Prentice–Hall, Englewood Cliffs, NJ, 1963).
[3] R. L. Hall, Phys. Rev. D **22**, 2062 (1980) (paper I); R. L. Hall and M. Satpathy, J. Phys. A: Math. Gen. **14**, 2645 (1981) (paper II); R. L. Hall, Phys. Rev. D **23**, 1421 (1981) (paper III).
[4] E. Prugovečki, *Quantum Mechanics in Hilbert Space* (Academic, New York, 1981).
[5] M. Reed and B. Simon, *Methods of Modern Mathematical Physics IV: Analysis of Operators* (Academic, New York, 1978).
[6] W. Thirring, *A Course in Mathematical Physics 3: Quantum Mechanics of Atoms and Molecules* (Springer, New York, 1981).
[7] R. L. Hall and C. Kalman, Phys. Lett. B **83**, 80 (1979).
[8] S. Flügge, *Practical Quantum Mechanics* (Springer, New York, 1974).
[9] H. Weyl, Math. Ann. **71**, 441 (1911).
[10] Ky Fan, Proc. Nat. Acad. Sci. (U.S.) **35**, 652 (1949).
[11] A. Weinstein and W. Stenger, *Methods of Intermediate Problems for Eigenvalues* (Academic, New York, 1972); Weyl's Theorem is discussed on p. 163.
[12] C. M. Bender and T. T. Wu, Phys. Rev. **184**, 1234 (1969); S. N. Biswas, K. Datta, R. P. Saxena, P. K. Srivastava, and V. S. Varma, J. Math. Phys. **14**, 1190 (1973); F. T. Hioe, D. MacMillen, and E. W. Montroll, J. Math. Phys. **17**, 1320 (1976); W. E. Caswell, Ann. Phys. **123**, 153 (1979); J. Killingbeck, J. Phys. A: Math. Gen. **13**, 49 (1980).
[13] B. Simon, Int. J. Quantum Chem. **21**, 3 (1982).

Sensitivity analysis of stochastic kinetic models

D. K. Dacol and H. Rabitz
Department of Chemistry, Princeton University, Princeton, New Jersey 08544

(Received 22 November 1983; accepted for publication 10 February 1984)

A formalism for sensitivity analysis of stochastic models describing fluctuation phenomena in chemically reacting systems is developed. The method is not restricted to chemical kinetics and can be used to analyze any model of a physical system whose state variables obey stochastic differential equations with white noise. Expressions for the sensitivity coefficients and densities are obtained. These expressions are suitable for direct evaluation by means of a stochastic simulation in a computer. The relationship between these quantities and the response functions studied in statistical mechanics is discussed.

PACS numbers: 05.40. + j, 82.20.Fd, 02.50.Fz

I. INTRODUCTION

Stochastic differential equations form the basis of widely used phenomenological models that describe the nonequilibrium behavior of physical systems. In most investigations in this area the emphasis is on obtaining either exact or approximate analytical results, which reduces considerably the range of systems that can be analyzed. However, in recent times there has been a growing interest in the study of more complex systems where a numerical approach is necessary in view of the difficulty in obtaining reliable analytical approximations or exact results. In such cases, the problem of how to study systematically the dependence of the quantities of interest on the parameters that define the model becomes highly nontrivial. In applied mathematics and statistics, techniques developed to address this problem are grouped under the name sensitivity analysis. It is the purpose of this work to develop a formalism for the sensitivity analysis of models described by stochastic differential equations. In order to facilitate this development a specific class of physical systems will be studied in detail. Thus, in the remainder of this paper, the discussion will be focused on a model that describes concentration fluctuations in chemically reacting systems.

The concentration of participating species in a chemically reacting system are fluctuating quantities due to the aleatory nature of the intermolecular processes. The magnitude of the concentration fluctuations usually is of the order of the inverse of the volume of the system. Thus, for macroscopic systems, the fluctuations are often negligible and the deterministic kinetic equations provide an accurate description of the behavior of the concentrations. However, there are situations in which, even for macroscopic systems, fluctuations are important. A typical case occurs when chemical instabilities develop in the system. In this case large-scale fluctuations, spanning macroscopic volumes comparable to the volume of the system, which normally are very rapidly damped, can be amplified and cause a transition to a state distinct from the initial one.[1] Thus in the modelling of kinetic processes allowance should be made for the inclusion of fluctuation effects.

A convenient framework for the description of fluctuations in chemical kinetics involves the use of stochastic differential equations.[2] These equations are similar in structure to the deterministic ones, except for the addition of noise terms. The noise terms attempt to describe the changes in the concentrations that happen on a much faster time scale than the relaxation times for the chemical processes under study. These fast changes are caused by reactive molecular collisions. The strength of the noise depends on the instantaneous values of the concentrations since the probability of a given reaction occurring in the system depends on the concentrations of the species involved. The parameters that enter in those equations are the same that enter in the deterministic discussion: rate constants, initial values of concentrations, etc. The quantities of interest here are not the solutions of the equations themselves but the average values, variances and correlations of the concentrations. Here, as in the deterministic case, the question of the sensitivity of these quantities with respect to changes in the parameters controlling the chemical processes naturally arises. In mathematical modelling the study of such questions is known as sensitivity analysis and it plays an important role in the understanding and optimization of models.[3] Consider the deterministic case, where the concentrations obey differential equations obtained from the mass action law. In its simplest form the sensitivity analysis of this system involves the computation of the derivatives of the concentrations with respect to the parameters that define the system.[4] These derivatives are evaluated assuming that the concentrations are known by solving the kinetic equations for a set of reference values of the parameters. Thus, the sensitivity coefficients, defined as first-order derivatives of the concentration with respect to the parameters, are just the gradients of the concentration evaluated at a reference point in parameter space.

A detailed discussion of possible applications of information obtained from sensitivity analysis in deterministic chemical kinetics and in scattering theory has been presented elsewhere.[5,6]

The main objective of this paper is to develop a methodology for the sensitivity analysis of stochastic chemical kinetics. However, the resulting formalism is not restricted to kinetic problems. It provides a framework for sensitivity studies of any system whose state variables obey stochastic differential equations with either additive or multiplicative white noise. Only minor modifications are required to include more general stochastic equations. The sensitivity analysis problem in stochastic chemical kinetics can be stat-

 0022-2488/84/092716-12$02.50

ed as follows. Suppose that for a certain set of values of the parameters, the reference values, all the expectation values of products of the concentrations can be determined. Such expectation values will be called correlation functions in this discussion. They are obtained by solving the stochastic differential equations of chemical kinetics and averaging appropriate products of solutions. It must be noted that the correlation functions are the observable quantities in stochastic chemical kinetics. If one is interested in the sensitivity of a certain correlation function with respect to variations in the parameters of the system then one should begin by studying the sensitivity coefficients of this particular correlation function, that is, its gradient in parameter space. In order to compute these derivatives, their expressions in terms of concentration correlation functions must be obtained. This task is accomplished in this work. The generalization to the case of time-dependent parameters, involving the concept of sensitivity densities, is also studied.

The use of stochastic differential equations in chemical kinetics is discussed in Sec. II. The sensitivity analysis for stochastic kinetics is developed in Sec. III. In Sec. IV the sensitivity analysis for stochastic kinetics is further investigated by considering the Fokker–Planck and related distributions associated with the stochastic differential equations of chemical kinetics. Section V considers the sensitivity analysis of the quasilinear approximation to stochastic kinetics. Finally, a simple application of the sensitivity formulas derived in Secs. III and IV is presented in Sec. VI.

II. STOCHASTIC DIFFERENTIAL EQUATIONS IN CHEMICAL KINETICS

In a chemically reacting system concentration fluctuations are always present because of the randomness that characterizes molecular processes. Despite the fact that normally the magnitude of these fluctuations is inversely proportional to the size of the system there are situations where fluctuation phenomena are important. This is the case when chemical instabilities develop in the system, for example, when initially the system is close to an unstable steady state.[2] If the system possesses multiple steady states then the transition to one of these states will be driven by the fluctuations (i.e., the decay of the unstable steady state is a fluctuation phenomenon). A particularly interesting case is the one of a system in which a stable steady state becomes unstable as a parameter is varied. A common situation is the critical bifurcation, where there is only one stable steady state S_0 when a parameter p is smaller than a critical value p_c. When p is increased beyond p_c, S_0 becomes unstable and two stable steady states S_1 and S_2 appear.[2] Obviously stable and unstable steady states here refer to regimes where fluctuations are negligible, as the system approaches an unstable state the fluctuations are magnified and lead the system towards one of the stable states, thus causing decay of a state that, albeit unstable, would be steady in the absence of fluctuations.

These considerations indicate the need for an adequate framework for describing the fluctuations that yield the deterministic limit in a simple way. Such a framework is provided by using stochastic differential equations to describe concentration fluctuations in chemical kinetics. For a well-stirred mixture of N species the equations for the N concentrations $\rho_i(t)$ are

$$\frac{d\rho_i}{dt} = R_i(\boldsymbol{\rho}, \boldsymbol{\alpha}) + \sum_{j=1}^{N} P_{ij}(\boldsymbol{\rho}, \boldsymbol{\alpha})\xi_j(t), \quad \rho_i(0) = \rho_i^0, \quad (2.1)$$

where $\{\xi_i(t)\}$ is a delta correlated Gaussian stochastic process[7] (i.e., white noise), with

$$\langle \xi_i(t) \rangle = 0, \quad \langle \xi_i(t)\xi_j(t') \rangle = \delta_{ij}\delta(t - t'). \quad (2.2)$$

It is well known that a stochastic differential equation like Eq. (2.1) is meaningless without an additional interpretation rule.[8] The required rule, compatible with the physics of this problem, is that Eq. (2.1) be interpreted as a stochastic differential equation in the sense of Stratonovich.[9] From now on all stochastic differential equations appearing in this work will be interpreted in this sense unless accompanied by an explicit statement to the contrary.

The symbol $\boldsymbol{\alpha}$ in Eq. (2.1) denotes the parameters that enter in the definition of the chemical rate vector R. The matrix $P(\boldsymbol{\rho}, \boldsymbol{\alpha})$ can be found from the fluctuation-dissipation analysis of Keizer[10a] and Grossmann.[10b] Thus

$$\sum_{k=1}^{N} P_{ik}(\boldsymbol{\rho}, \boldsymbol{\alpha})P_{jk}(\boldsymbol{\rho}, \boldsymbol{\alpha}) = Q_{ij}(\boldsymbol{\rho}, \boldsymbol{\alpha}), \quad (2.3)$$

where Q is a symmetric, positive semidefinite matrix that can be written in terms of the forward and backward rates for the elementary reactions included in the rate vector R in Eq. (2.1). Assuming that the concentrations are expressed in (number of molecules)/(unit volume) and that the forward and backward rates $\vec{\pi}_k$ and $\overleftarrow{\pi}_k$ for the k th elementary reaction, are measured in (molecules/unit volume)/(unit time) the rate vector R, and the variance matrix Q can be written as[10]

$$R_i(\boldsymbol{\rho}, \boldsymbol{\alpha}) = \sum_{k=1}^{M} \nu_{ik}(\vec{\pi}_k(\boldsymbol{\rho}, \boldsymbol{\alpha}) - \overleftarrow{\pi}_k(\boldsymbol{\rho}, \boldsymbol{\alpha})), \quad (2.4)$$

$$Q_{ij}(\boldsymbol{\rho}, \boldsymbol{\alpha}) = \frac{1}{V}\sum_{k=1}^{M} \nu_{ik}\nu_{jk}(\vec{\pi}_k(\boldsymbol{\rho}, \boldsymbol{\alpha}) + \overleftarrow{\pi}_k(\boldsymbol{\alpha}, \boldsymbol{\rho})), \quad (2.5)$$

where M is the number of elementary reactions, ν_{ik} is the stoichiometric coefficient of species j in the reaction k, and V is the volume containing the chemically reacting system under study.

From a mathematical point of view the stochastic process $\xi_i(t)$ induces a new stochastic process $\rho_i(t)$ through the mediation of Eq. (2.1). The statistical properties of $\xi_i(t)$ plus the dynamical properties incorporated in Eq. (2.1) (including initial conditions) uniquely determine the statistical properties of $\rho_i(t)$, which allows for the computation of average values, variances and correlation of the concentrations. Physically Eq. (2.1) assumes that the intermolecular processes (reactive collisions) that give rise to the fluctuations occur on a time scale much faster than the relaxation time predicted by the mass balance equation when fluctuations are neglected [Eq. (2.1) without the stochastic driving term]. The model defined by Eqs. (2.1)–(2.5) constitutes an independent, self-consistent formalism for the description of concentration fluctuations. In addition, a rigorous link has been established between the stochastic model adopted here and another popular phenomenological description of fluctuation, the birth and death formalism.[11] Essentially, it can be shown

that Eq. (2.1) constitutes a diffusion approximation to a corresponding master equation in the birth and death formalism.[12] The larger the system is, the better is the approximation.

The question of whether stochastic differential equations provide an appropriate description of fluctuation phenomena in chemical kinetics has been the subject of considerable debate. Recent developments[13] reinforce the author's belief that stochastic differential equations of the type studied in this work do provide an adequate description of fluctuations in chemical kinetics even though the precise prescriptions used to obtain R_i and F_{ij} [see Eq. (2.1)] might have to be modified. In this connection it must be emphasized that the main results of this work, namely the general expressions for sensitivity coefficients and densities, are independent of the specific dependence of R_i and P_{ij} on the concentrations. The recent developments mentioned above refer to an extension of Onsager's thermodynamic fluctuation theory to nonlinear systems that has been obtained on the basis of statistical mechanical considerations by Grabert *et al.*[13a,b] In this approach fluctuations are described by a Fokker–Planck equation which implies an underlying stochastic description using nonlinear stochastic differential equations of the same form as Eq. (2.1). An application of this formalism to the study of fluctuations in reversible chemical reactions can be found in a very recent publication of H. Grabert *et al.*[13c]

III. SENSITIVITY ANALYSIS OF STOCHASTIC KINETICS

When modelling physical systems one is confronted with the question of how sensitive the behavior of the system is to changes in the values of the parameters that define the model. Sensitivity analysis is a very general methodology developed to deal with these questions.[3] In its simplest version the essence of the method consists in extracting from the sensitivity coefficients all possible information about the response of the system to variations in the input parameters. These coefficients are, by definition, the derivatives of the state functions that describe the behavior of the system with respect to the input parameters. In the case of chemically reacting systems the state functions that describe the kinetic behavior are average values of functionals of the stochastic variable $\boldsymbol{\rho}(t)$, which represents the instantaneous (and fluctuating) values of the concentrations of the chemical species. The purpose of this section is to obtain expressions for the sensitivity coefficients of such quantities. The generalization of these results to the case of time-dependent parameters is also discussed. This involves the computation of sensitivity densities,[14] which here are defined as functional derivatives of average values of functionals of $\boldsymbol{\rho}(t)$ with respect to the time-dependent parameters.

Previous work on sensitivity analysis of chemical kinetics focused on the deterministic situation when the concentrations satisfy rate equations like Eq. (2.1) with the noise terms absent. In these works the fundamental quantities from the point of view of sensitivity analysis are the sensitivity coefficients of the concentrations evaluated at a given set of reference values of the parameters. They are expressed as functionals of the concentrations which are solutions of the rate equations with the parameters set at their reference values.[4] The quantities of interest in the stochastic case are not the solutions to Eq. (2.1) themselves but averages of those solutions like the average value and the variance of $\rho_i(t)$,

$$\overline{\rho_i(t)} = \langle \rho_i(t) \rangle, \quad \sigma_i(t) = (\langle \rho_i^2(t) \rangle - \langle \rho_i(t) \rangle^2)^{1/2}. \tag{3.1}$$

The sensitivity coefficients of the average values of functionals of $\boldsymbol{\rho}(t)$ are themselves average values of related functionals of $\boldsymbol{\rho}(t)$. Let $F[\boldsymbol{\rho}]$ be an arbitrary functional of $\boldsymbol{\rho}(t)$ with $\langle F[\boldsymbol{\rho}] \rangle$ its average value. The parameters that enter in Eq. (2.1) are denoted by α_p, $p = 1, ..., P$, where in general $P > N$ (N is the number of chemical species). For future convenience it is assumed that $\alpha_{i+P-N} = \rho_i^0$, $i = 1, ..., N$, where ρ_i^0 is the initial value of the concentration $\rho_i(t)$. The sensitivity coefficient of $\langle F[\boldsymbol{\rho}] \rangle$ with respect to a parameter α_p is defined as

$$S_p^F = \frac{\partial}{\partial \alpha_p} \langle F[\boldsymbol{\rho}] \rangle. \tag{3.2}$$

Besides an implicit dependence on the parameters through $\boldsymbol{\rho}$, $F[\boldsymbol{\rho}]$ can also explicitly depend on $\boldsymbol{\alpha}$, thus

$$\frac{\partial}{\partial \alpha_p} \langle F[\boldsymbol{\rho}] \rangle = \left\langle \frac{\partial F}{\partial \alpha_p}[\boldsymbol{\rho}] \right\rangle + \sum_{i=1}^{N} \int \left\langle \frac{\delta F[\boldsymbol{\rho}]}{\delta \rho_i(t)} \frac{\partial \rho_i(t)}{\partial \alpha_p} \right\rangle. \tag{3.3}$$

If $F[\boldsymbol{\rho}]$ depends on α_p only through the concentrations then the first term in the right-hand side of the above equation is zero. An expression for $\partial \rho_i(t)/\partial \alpha_p$ can be obtained by means of the Green's function method introduced in the sensitivity analysis of deterministic kinetics.[4] Thus,

$$\frac{\partial \rho_i(t)}{\partial \alpha_p} = \sum_{j=1}^{N} G_{ij}(t, 0) \lambda_{jp} + \sum_{j=1}^{N} \int ds\, G_{ij}(t, s) \times \left(\frac{\partial R_j}{\partial \alpha_p}(\boldsymbol{\rho}(s), \boldsymbol{\alpha}) + \sum_{k=1}^{N} \frac{\partial P_{jk}}{\partial \alpha_p}(\boldsymbol{\rho}(s), \boldsymbol{\alpha}) \xi_k(s) \right). \tag{3.4}$$

Above $\lambda_{jp} = \partial \rho_j^0 / \partial \alpha_p$, $\lambda_{jp} = 0$ if α_p is not an initial value (i.e., if $p \leqslant P - N$) and $\lambda_{jp} = \delta_{ij}$ if $p = P - N + i$, that is, if α_p is an initial value. The Green's function satisfies

$$\frac{\partial}{\partial t} G_{ij}(t, s) = \sum_{k} \left(\frac{\partial R_i}{\partial \rho_k}(\boldsymbol{\rho}(t), \boldsymbol{\alpha}) + \sum_{k'} \frac{\partial P_{ik'}}{\partial \rho_k}(\boldsymbol{\rho}(t), \boldsymbol{\alpha}) \xi_{k'}(t) \right) \times G_{kj}(t, s) + \delta_{ij} \delta(t - s). \tag{3.5}$$

From Eq. (3.5) it follows that

$$G_{ij}(t, s) = \theta(t - s) \sum_{k} U_{ik}(t) U_{kj}^{-1}(s), \tag{3.6}$$

where $U_{ij}(t)$ obeys an homogeneous differential equation like Eq. (3.5) without the delta function term and with initial condition $U_{ij}(0) = \delta_{ij}$. In Eq. (3.6) $U^{-1}(s)$ is the inverse of the matrix $U(s)$ whose elements are the $U_{ij}(s)$.

In order to combine Eqs. (3.3) and (3.4) it will prove useful to relate the Green's function $G_{ij}(t, s)$ to the functional derivative of the concentration $\rho_i(t)$ with respect to the noise $\xi_k(s)$. This relation is obtained by proceeding in a manner analogous to the one used to obtain Eq. (3.4). Thus, computing the functional derivatives of both sides of Eq. (2.1) and solving the resulting linear differential equation for $\delta \rho_i(t)/\delta \xi_k(s)$ with the help of the Green's function it follows

that

$$\frac{\delta \rho_i(t)}{\delta \xi_k(s)} = \sum_j G_{ij}(t, s) P_{jk}(\rho(s), \alpha). \tag{3.7}$$

By solving Eq. (3.7) for $G_{ij}(t, s)$ and substituting into Eq. (3.4) the right-hand side of Eq. (3.3) can be written as

$$\sum_i \int ds \left\langle \frac{\delta F[\rho]}{\delta \xi_i(s)} h_{ip}(\rho(s), \xi(s), \alpha) \right\rangle, \tag{3.8}$$

where

$$h_{ip}(\rho(s), \xi(s), \alpha) = \sum_j P^{-1}_{ij}(\rho(s), \alpha) \left(\lambda_{jp} \delta(s) + \frac{\partial R_j}{\partial \alpha_p}(\rho(s), \alpha) + \sum_k \frac{\partial P_{jk}}{\partial \alpha_p}(\rho(s), \alpha) \xi_k(s) \right). \tag{3.9}$$

Further manipulations can cast Eq. (3.8) into a form that does not involve $\delta F[\rho]/\delta \xi_i(s)$. It is convenient to eliminate this functional derivative since its evaluation demands obtaining $G_{ij}(t, s)$ which satisfies another stochastic differential equation, Eq. (3.5), independent (although related) to Eq. (2.1).

Suppose $H[\xi]$ is an arbitrary functional of the Gaussian stochastic process $\xi(t)$, then the following identity (proved in Appendix A) holds:

$$\langle \xi_k(s) H[\xi] \rangle = \left\langle \frac{\delta H[\xi]}{\delta \xi_k(s)} \right\rangle. \tag{3.10}$$

Using the above identity Eq. (3.8) can be written as

$$\sum_i \int ds \left\langle \frac{\delta F[\rho]}{\delta \xi_i(s)} h_{ip}(\rho(s), \xi(s), \alpha) \right\rangle = \lim_{\epsilon \to 0^+} \sum_i \int ds \langle F[\rho] \xi_i(s + \epsilon) h_{ip}(\rho(s), \xi(s), \alpha) \rangle. \tag{3.11}$$

Equation (3.11) is established by applying Eq. (3.10) to the expectation value in its right-hand side and noticing that the functional derivative of $h_{ip}(\rho(s), \xi(s), \alpha)$ with respect to $\xi_i(s + \epsilon)$, $\epsilon > 0$, is zero since $\xi(s)$ and $\xi(s + \epsilon)$ are independent variables and $\rho(s)$ depends on $\xi(t)$ only for $t \leqslant s$ [a consequence of the fact that Eq. (2.1) respects causality].

The final expression for S^F_p is

$$S^F_p = \left\langle \frac{\partial F}{\partial \alpha_p}[\rho] \right\rangle + \sum_i \int ds \langle F[\rho] \xi_i(s^+) h_{ip}(\rho(s), \xi(s), \alpha) \rangle, \tag{3.12}$$

where the notation $\xi_i(s^+)$ indicates the limit prescribed in Eq. (3.11). It must be pointed out that the evaluation of the sensitivity coefficient S^F_p [Eq. (3.12)] does not require the solution of any other differential equation besides Eq. (2.1) for the concentrations. In particular, it is not necessary to obtain the Green's function $G_{ij}(t, s)$. This is to be contrasted with the deterministic case[4] where the evaluation of the sensitivity coefficients requires not only the concentrations but also the deterministic analog of the Green's function $G_{ij}(t, s)$.

If the parameters are allowed to be functions of time then the quantities of interest in sensitivity analysis are the sensitivity densities. These are functional derivatives of functionals of the concentrations with respect to the time-dependent parameters. For simplicity in this work it will be assumed that $R_i(t)$ and $P_{jk}(t)$ are functions of $\rho(t)$ and $\alpha(t)$. However the following discussion can be extended to more general cases, the only restriction being that both $R_i(t)$ and $P_{jk}(t)$ must be causal (or nonanticipating) functions [i.e., neither $R_i(t)$ or $P_{jk}(t)$ depend on $\rho(s)$ for $s > t$]. The sensitivity densities are defined as

$$D^F_p(t) = \left\langle \frac{\delta F[\rho]}{\delta \alpha_p(t)} \right\rangle. \tag{3.13}$$

Proceeding as before one obtains

$$D^F_p(t) = \left\langle \frac{\delta F[\rho]}{\delta \alpha_p(t)} \right| \Big\rangle_\rho + \sum_i \langle F[\rho] \xi_i(t^+) h_{ip}(\rho(t) \xi(t), \alpha(t)) \rangle. \tag{3.14}$$

The assumption that R_i and P_{jk} are functions of ρ and α was used above. This assumption implies

$$\left. \frac{\delta R_j}{\delta \alpha_p(t)}(\rho(s), \alpha(s)) \right|_\rho = \frac{\partial R_j}{\partial \alpha_p}(\rho(s), \alpha(s)) \delta(s - t), \tag{3.15}$$

$$\left. \frac{\delta P_{jn}}{\delta \alpha_p(t)}(\rho(s), \alpha(s)) \right|_\rho = \frac{\partial P_{jk}}{\partial \alpha_p}(\rho(s), \alpha(s)) \delta(s - t). \tag{3.16}$$

If F is a functional of both ρ and α then $\delta F/\delta \alpha_p(t)|_\rho$ is to be calculated by keeping ρ fixed while allowing α to vary. Thus, unless $F[\rho]$ is explicitly dependent on α, the first term on the right-hand side of Eq. (3.14) vanishes.

In general $F[\rho]$, S^F_p and $D^F_p(t)$ will have to be numerically evaluated. The procedure can be briefly outlined as follows. A particular realization of the stochastic variable $\xi^a(t)$ is called a sample trajectory. It will be denoted $\xi^a(t)$, where a is an index labeling the sample trajectory in question. The sample trajectories $\xi^a(t)$ can be obtained by means of a random number generator. For each sample trajectory $\xi^a(t)$ numerical integration of Eq. (2.1) yields a sample trajectory $\rho^a(t)$ which is a particular realization of the stochastic variable $\rho(t)$. Thus, if $E[\rho, \xi]$ is a functional of the stochastic variables ρ and ξ then

$$\langle E[\rho, \xi] \rangle = \lim_{M \to \infty} \frac{1}{M} \sum_{a=1}^{M} E[\rho^a, \xi^a]. \tag{3.17}$$

Therefore, after generating the $\xi^a(t)$ and obtaining the $\rho^a(t)$ from Eq. (2.1) $\langle F[\rho] \rangle$, $S^F_p(t)$ and $D^F_p(t)$ can be obtained by means of Eq. (3.17).

Before concluding this section some further remarks on the Green's function $G_{ij}(t, s)$ are in order. The average value of this function is the concentration response function. It gives the linear response of the concentration of species i to the presence of a weak source of species j. In order to see this, consider Eq. (2.1) with species sources added to the right-hand side

$$\frac{d \rho^J_i}{dt} = R_i(\rho^J, \alpha) + \sum_{j=1}^{N} P_{ij}(\rho^J, \alpha) \xi_j(t) + J_i(t). \tag{3.18}$$

By ρ^J_i it is meant the concentration of species i when sources are present. The concentrations in the absence of sources will be just ρ_i. If the sources are sufficiently weak then

$$\rho^J_i(t) = \rho_i(t) + \sum_{j=1}^{N} \int ds \left(\frac{\delta \rho^J_i(t)}{\delta J_j(s)} \Big|_{J=0} \right) J_j(s) + \mathcal{O}(J^2). \tag{3.19}$$

Now, it follows from Eq. (3.18) that $\delta \rho^J_i(t)/\delta J_j(s)$ obeys a differential equation that is identical to the one obeyed by

$G_{ij}(t, s)$ if in the latter $\rho_i(t)$ is replaced by $\rho_i^J(t)$ wherever it occurs. Therefore, it is clear that

$$G_{ij}(t, s) = \frac{\delta \rho_i^J(t)}{\delta J_j(s)}\bigg|_{J=0}. \tag{3.20}$$

From Eqs. (3.19) and (3.20), to first order in J,

$$\langle \rho_i^J(t)\rangle = \langle \rho_i(t)\rangle + \sum_{j=1}^{N} \int ds \langle G_{ij}(t, s)\rangle J_j(s) + \mathcal{O}(J^2). \tag{3.21}$$

Equation (3.21) allows the identification of $\langle G_{ij}(t, s)\rangle$ as a concentration response function. Thus, for example, if a small amount of species j is introduced at $t = t_0$, that is, if $J_i(t) = \lambda \delta_{ij} \delta(t - t_0)$, then the change in the average concentrations, to first order in λ, will be

$$\delta \langle \rho_i(t)\rangle = \lambda \langle G_{ij}(t, t_0)\rangle. \tag{3.22}$$

In fact $\langle G_{ij}(t, s)\rangle$ is the equivalent in this problem of a response function introduced by Martin, Siggia, and Rose in their study of the dynamics of classical stochastic fields.[15] In the deterministic limit, when fluctuations are neglected [formally obtained by letting $V \to \infty$ in Eq. (2.5) which leads to $P_{ij} = 0$] the Green's function $G_{ij}(t, s)$ becomes equal to the Green's function introduced in a previous work on the sensitivity analysis of deterministic kinetics.[4]

From an analysis similar to the one carried out above for the Green's function it can be concluded that the sensitivity densities $D_p^F(t)$ are generalized response functions.

IV. SENSITIVITY ANALYSIS AND THE FOKKER–PLANCK EQUATION

A useful approach to the study of a stochastic differential equation like Eq. (2.1) is to try to determine the probability density $P_1(\mathbf{y}, t)$ such that the joint probability that the $\rho_i(t)$, $i = 1, \ldots, N$, assume values in the intervals $[\, y_i, y_i + dy_i\,]$ is given by $P_1(\mathbf{y}, t)d\mathbf{y}$. In the case of Eq. (2.1), $P_1(\mathbf{y}, t)$ is the solution of a Fokker–Planck equation. The knowledge of $P_1(\mathbf{y}, t)$ allows the computation of quantities like $\langle F(\boldsymbol{\rho}(t))\rangle$, where $F(\boldsymbol{\rho}(t))$ is a function of the concentration vector $\boldsymbol{\rho}(t)$, thus

$$\langle F(\boldsymbol{\rho}(t))\rangle = \int d\mathbf{y}\, P_1(\mathbf{y}, t) F(\mathbf{y}). \tag{4.1}$$

An equation for $P_1(\mathbf{y}, t)$ can be obtained by first noticing that[16]

$$P_1(\mathbf{y}, t) = \langle \delta(y_1 - \rho_1(t)) \cdots \delta(y_N - \rho_N(t))\rangle. \tag{4.2}$$

Substitution of the above expression for $P_1(\mathbf{y}, t)$ into the right-hand side of Eq. (4.1) leads to an identity. Let $\delta(\mathbf{y} - \boldsymbol{\rho}(t)) = \delta(y_1 - \rho_1(t)) \cdots \delta(y_N - \rho_N(t))$. Then

$$\frac{\partial}{\partial t} P_1(\mathbf{y}, t) = -\sum_i \frac{\partial}{\partial y_i} \left\langle \delta(\mathbf{y} - \boldsymbol{\rho}(t)) \frac{d\rho_i(t)}{dt} \right\rangle. \tag{4.3}$$

Now using Eq. (2.1) to eliminate $d\rho_i(t)/dt$ and employing Eq. (3.10) to deal with the term proportional to $\xi(t)$ it follows that

$$\frac{\partial}{\partial t} P_1(\mathbf{y}, t) = -\sum_i \frac{\partial}{\partial y_i} (R_i(\mathbf{y}, \boldsymbol{\alpha}(t)) P_1(\mathbf{y}, t)) + \frac{1}{2} \sum_{ijk} \frac{\partial}{\partial y_i} (P_{ij}(\mathbf{y}, \boldsymbol{\alpha}(t))) \times \frac{\partial}{\partial y_k} (P_{kj}(\mathbf{y}, \boldsymbol{\alpha}(t)) P_1(\mathbf{y}, t)). \tag{4.4}$$

In obtaining Eq. (4.4) the following identity was used

$$\frac{\delta \rho_i(t)}{\delta \xi_k(t)} = \lim_{\epsilon \to 0^+} \frac{1}{2} \left\{ \frac{\delta \rho_i(t+\epsilon)}{\delta \xi_k(t)} + \frac{\delta \rho_i(t-\epsilon)}{\delta \xi_k(t)} \right\} = \frac{1}{2} P_{ik}(\boldsymbol{\rho}(t), \boldsymbol{\alpha}(t)). \tag{4.5}$$

The first equality above is the definition of the equal time limit of $\delta \rho_i(t)/\delta \xi_k(s)$ according to the Stratonovich calculus, and the second equality is a consequence of Eq. (3.7). Equation (4.5) is an important identity which will be used many times in this section.

The initial condition for Eq. (4.4) is just

$$P_1(\mathbf{y}, 0) = \delta(\mathbf{y} - \boldsymbol{\rho}^0). \tag{4.6}$$

The boundary condition necessary to solve Eq. (4.4) can be obtained by demanding that $P_1(\mathbf{y}, t)$ be normalizable, that is

$$\int dy\, P_1(\mathbf{y}, t) = 1. \tag{4.7}$$

In the same way that $P_1(\mathbf{y}, t)$ was introduced in order to evaluate $\langle F(\boldsymbol{\rho}(t))\rangle$ it is possible to introduce a multiple time probability distribution $P_n(\mathbf{y}_1, t_1; \cdots \mathbf{y}_n, t_n)$ such that[17]

$$\langle F_1(\boldsymbol{\rho}(t_1)) \cdots F_n(\boldsymbol{\rho}(t_n))\rangle = \int d\mathbf{y}_1 \cdots \int d\mathbf{y}_n\, \mathrm{F}_1(\mathbf{y}_1) \cdots F_n(\mathbf{y}_n) P_n(y_1, t_1; \ldots; \mathbf{y}_n, t_n). \tag{4.8}$$

Therefore, $P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) d\mathbf{y}_1 \cdots d\mathbf{y}_n$ is the joint probability that the $\rho_i(t_m)$ assume values in the intervals $[\, y_{im}, y_{im} + dy_{im}\,]$. Like $P_1(\mathbf{y}, t)$ the multiple time distribution $P_n(\mathbf{y}_1 t_1, \ldots, \mathbf{y}_n t_n)$ can also be written as an average value of delta functions

$$P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) = \langle \delta(\mathbf{y}_1 - \boldsymbol{\rho}(t_1)) \cdots \delta(\mathbf{y}_n - \boldsymbol{\rho}(t_n))\rangle. \tag{4.9}$$

In order to derive a relation from which explicit expressions for the $P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n)$ can be obtained it is useful to notice that it can be assumed without loss of generality that $t_1 > t_2 > \cdots > t_n$. This is an obvious consequence of Eq. (4.9). For example, if in the left-hand side some of the time variables are equal then the right-hand side will be reduced to a product of deltas multiplied by a multiple time distribution that depends only on the distinct time variables. For example, suppose that $t_1 = t_2 = t$ with $t > t_3 > \cdots > t_n$, then

$$P_n(\mathbf{y}_1, t; \mathbf{y}_2, t; \ldots; \mathbf{y}_n, t_n) = \delta(\mathbf{y}_1 - \mathbf{y}_2) P_{n-1}(\mathbf{y}_2, t; \mathbf{y}_3, t_3; \ldots; \mathbf{y}_n, t_n). \tag{4.10}$$

In Appendix B, it is shown that

$$P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) = G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2) P_{n-1}(\mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n). \tag{4.11}$$

In Eq. (4.11), $G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2)$ is the Green's function associated with the Fokker–Planck operator [see Eqs. (B2), (B8), and

(B9)]. Therefore, it is clear that once $P_1(\mathbf{y}, t)$ and $G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2)$ have been found, all the multiple time distributions can be obtained from Eq. (4.11).

The sensitivity analysis of stochastic chemical kinetics may also be studied in this framework. Thus in this section the sensitivity coefficients and densities of $\langle F_1(\boldsymbol{\rho}(t_1))\cdots F_n(\boldsymbol{\rho}(t_n))\rangle$ will be studied using this alternative formalism. It will be assumed that $F_i(\boldsymbol{\rho}(t_i))$ depends on $\boldsymbol{\alpha}$ only through $\boldsymbol{\rho}(t_i)$ and that $t_1 > t_2 > \cdots > t_n$.

The sensitivity coefficients and densities, as defined in Sec. III [Eqs. (3.2) and (3.13) with $F[\boldsymbol{\rho}] = F_1(\boldsymbol{\rho}(t_1))\cdots F_n(\boldsymbol{\rho}(t_n))]$, are then

$$S_p^F = \int d\mathbf{y}_1 \cdots d\mathbf{y}_n \, F_1(\mathbf{y}_1)\cdots F_n(\mathbf{y}_n) S_p^n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n), \quad (4.12)$$

$$D_p^F(t) = \int d\mathbf{y}_1 \cdots d\mathbf{y}_n \, F_1(\mathbf{y}_1)\cdots F_n(\mathbf{y}_n) D_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n). \quad (4.13)$$

S_p^n and D_p^n are the sensitivity coefficients and densities of the multiple time distribution P_n,

$$S_p^n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) = \frac{\partial}{\partial \alpha_p} P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n), \quad (4.14)$$

$$D_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) = \frac{\delta}{\delta \alpha_p(t)} P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n). \quad (4.15)$$

Equation (4.11) implies a recursion relation for the S_p^n and the D_p^n,

$$\begin{aligned} &S_p^n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) \\ &\quad = G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2) S_p^{n-1}(\mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n) \\ &\qquad + \left(\frac{\partial}{\partial \alpha_p} G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2)\right) P_{n-1}(\mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n), \end{aligned} \quad (4.16)$$

$$\begin{aligned} &D_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) \\ &\quad = G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2) D_p^{n-1}(t\,|\mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n) \\ &\qquad + \left(\frac{\delta}{\delta \alpha_p(t)} G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2)\right) P_{n-1}(\mathbf{y}_2, t_2; \ldots; y_n, t_n). \end{aligned} \quad (4.17)$$

Thus once the sensitivity coefficients and densities of $P_1(\mathbf{y}, t)$ and $G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2)$ have been obtained all the others can be determined through the recursion relations.

Both $P_1(\mathbf{y}, t)$ and $G(\mathbf{y}_1, t; \mathbf{y}_2, t_2)$ satisfy linear diffusion equations. Thus the sensitivity densities and coefficients of P_1 and G can be obtained by using a version of the Green's function method[4] recently developed to deal with reaction-diffusion equations.[14,18] The basic idea is to determine the sensitivity coefficients by computing appropriate derivatives (functional derivatives) of both sides of the equations satisfied by P_1 and G thereby obtaining a linear differential equation for the sensitivity coefficients (densities). This equation can be integrated with the help of a Green's function which in this case is just G, the Green's function defined in Eqs. (B8) and (B9). Thus

$$S_p^1(\mathbf{y}, t) = \sum_i \lambda_{ip} \frac{\partial}{\partial x_i} (G(\mathbf{y}, t; \mathbf{x}, 0))\Big|_{\mathbf{x}=\boldsymbol{\rho}^0} + \int dt' \int d\mathbf{x}\, G(\mathbf{y}, t; \mathbf{x}, t') L(\mathbf{x}, t') P_1(\mathbf{x}, t'), \quad (4.18)$$

$$D_p^1(s|\mathbf{y}, t) = \delta(s) \sum_i \lambda_{ip} \frac{\partial}{\partial x_i} (G(\mathbf{y}, t; \mathbf{x}, 0))\Big|_{\mathbf{x}=\boldsymbol{\rho}^0} + \int d\mathbf{x}\, G(\mathbf{y}, t; \mathbf{x}, s) L(\mathbf{x}, s) P_1(\mathbf{x}, s). \quad (4.19)$$

For the sensitivity coefficients and densities of the Green's function one obtains

$$\frac{\partial}{\partial \alpha_p} G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2) = \int dt \int d\mathbf{x}\, G(\mathbf{y}_1, t_1; \mathbf{x}, t) L(\mathbf{x}, t) G(\mathbf{x}, t; \mathbf{y}_2, t_2), \quad (4.20)$$

$$\frac{\delta}{\delta \alpha_p(t)} G(\mathbf{y}_1, f_1; \mathbf{y}_2, f_2) = \int d\mathbf{x}\, G(\mathbf{y}_1, t_1; \mathbf{x}, t) L(\mathbf{x}, t) G(\mathbf{x}, t; \mathbf{y}_2, t_2). \quad (4.21)$$

The differential operator $L(\mathbf{x}, t)$ introduced above is just $(\partial / \partial \alpha_p) D(\mathbf{x}, t)$, where $D(\mathbf{x}, t)$ is the Fokker–Planck differential operator defined by Eq. (B2), its explicit expression is

$$\begin{aligned} L(\mathbf{x}, t) f(\mathbf{x}, t) = &- \sum_i \frac{\partial}{\partial x_i}\left(f(\mathbf{x}, t) \frac{\partial}{\partial \alpha_p} R_i(\mathbf{x}, \boldsymbol{\alpha}(t))\right) \\ &+ \frac{1}{2} \sum_{ijk} \frac{\partial}{\partial x_i}\left[\left(\frac{\partial}{\partial \alpha_p} P_{ij}(\mathbf{x}, \boldsymbol{\alpha}(t))\right)\right. \\ &\times \frac{\partial}{\partial x_k} (f(\mathbf{x}, t) P_{kj}(\mathbf{x}, \boldsymbol{\alpha}(t))) + P_{ij}(\mathbf{x}, \boldsymbol{\alpha}(t)) \\ &\left. \times \frac{\partial}{\partial x_k}\left(f(\mathbf{x}, t) \frac{\partial}{\partial \alpha_p} P_{kj}(\mathbf{x}, \boldsymbol{\alpha}(t))\right)\right]. \end{aligned} \quad (4.22)$$

Equations (4.12)–(4.22) provide the basic algorithm for carrying out a sensitivity analysis of systems described by stochastic differential equations in the Fokker–Planck equation approach.

At this point it is necessary to demonstrate the equivalence between Eq. (4.13) and Eq. (3.14) for the sensitivity density $D_p^F(t)$. Equation (3.14) with $F[\boldsymbol{\rho}] = F_1(\boldsymbol{\rho}(t_1))\cdots F_n(\boldsymbol{\rho}(t_n))$ can be written as

$$D_p^F(t) = \int d\mathbf{y}_1 \cdots d\mathbf{y}_n \, F_1(\mathbf{y}_1)\cdots F_n(\mathbf{y}_n) I_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n), \quad (4.23)$$

with

$$I_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) = \sum_i \langle \xi_i(t^+) h_{ip}(\boldsymbol{\rho}(t), \boldsymbol{\xi}(t), \boldsymbol{\alpha}(t)) \times \delta(\mathbf{y}_1 - \boldsymbol{\rho}(t_1))\cdots\delta(\mathbf{y}_n - \boldsymbol{\rho}(t_n))\rangle. \quad (4.24)$$

Equation (B10) with $\mathbf{x} = \mathbf{y}_1$, $t = t_1$, and $s = t_2$ leads to

$$I_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n)$$
$$= G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2) I_p^{n-1}(t\,|\mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n)$$
$$+ \int_{t_2}^{\infty} dt' \int d\mathbf{x}\, G(\mathbf{y}_1, t_1; \mathbf{x}, t') \langle R(\mathbf{x}, \boldsymbol{\rho}(t'), \boldsymbol{\xi}(t'))$$
$$\times \sum_i \xi_i(t^+) h_{ip}(\boldsymbol{\rho}(t), \boldsymbol{\xi}(t), \boldsymbol{\alpha}(t))$$
$$\times \delta(\mathbf{y}_2 - \boldsymbol{\rho}(t_2)) \cdots \delta(\mathbf{y}_n - \boldsymbol{\rho}(t_n)) \rangle. \tag{4.25}$$

From Eq. (4.25) one obtains

$$I_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n)$$
$$= G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2) I_p^{n-1}(t\,|\mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n)$$
$$+ \theta(t - t_2) \int d\mathbf{x}\, G(\mathbf{y}_1, t_1; \mathbf{x}, t)$$
$$\times \left(-\delta(t) \sum_j \lambda_{jp} \frac{\partial}{\partial x_j} + L(\mathbf{x}, t) \right)$$
$$\times P_n(\mathbf{x}, t; \mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n). \tag{4.26}$$

In Eq. (4.26), $\theta(t - t_2)$ is the Heaviside step function, $\theta(t - t_2) = 1$ if $t > t_2$ and $\theta(t - t_2) = 0$ if $t \leqslant t_2$. Equation (4.26) shows that $I_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n)$ is equal to $D_p^n(t\,|\mathbf{y}_2, t_1; \ldots; \mathbf{y}_n, t_n)$ defined by Eq. (4.17). First notice that both I_p^n and D_p^n are zero if $t > t_1$. For $t_1 > t > t_2 \geqslant 0$, Eqs. (4.17) and (4.21) again show that $D_p^n = I_p^n$. If $t_2 > t$ then

$$I_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n)$$
$$= G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2) I_p^{n-1}(t\,|\mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n), \tag{4.27}$$

$$D_p^n(t\,|\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n)$$
$$= G(\mathbf{y}_1, t_1; \mathbf{y}_2, t_2) D_p^{n-1}(t\,|\mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n). \tag{4.28}$$

Now if $t_2 > t > t_3 \geqslant 0$ one has $I_p^{n-1} = D_1^{n-1}$. Proceeding in this way it remains only to prove that $I_p^1 = D_p^1$ to establish that $I_p^n = D_p^n$. From Eq. (4.24) and Eq. (B10) with $\mathbf{x} = \mathbf{y}_1$ and $s = t$ (assuming $t_1 > t$) one gets

$$I_p^1(t\,|\mathbf{y}_1, t_1) = \int d\mathbf{x}\, G(\mathbf{y}_1, t_1; \mathbf{x}, t)$$
$$\times \left(-\delta(t) \sum_j \lambda_{jp} \frac{\partial}{\partial x_j} + L(\mathbf{x}, t) \right) P_1(\mathbf{x}, t). \tag{4.29}$$

Integrating by parts the term proportional to $\delta(t)$ in the right hand of Eq. (4.29) and comparing with Eq. (4.19) shows that $I_p^1(t\,|\mathbf{y}_1, t_1) = D_p^1(t\,|\mathbf{y}_1, t_1)$. Thus these two distinct methods for computing $D_p^F(t)$ yield the same result therefore establishing the equivalence of the two approaches. The method developed in this section seems better suited for analytical investigations while that discussed in Sec. III is more appropriate for numerical studies.

V. QUASILINEAR APPROXIMATION

In chemical systems the magnitude of the fluctuations is inversely proportional to the volume containing the chemically reacting system. This is reflected by the factor $1/V$ in the right-hand side of the equation that defines $Q_{ij}(\boldsymbol{\rho}, \boldsymbol{\alpha})$ [Eq. (2.5)]. Usually $1/V$ is very small when compared with typical concentrations (in this work concentrations are measured in number of molecules/unit volume and so the fluctuations are of small magnitude). For large systems the fluctuations will be important only in the neighborhood of a chemical instability.[1,2] Thus away from a chemical instability an expansion in powers of $1/V$ should provide a reasonably good description of the fluctuations. The zero-order solutions of Eq. (2.1) are just the deterministic concentrations which satisfy the usual rate equations [Eq. (2.1) with the noise term absent]. To lowest order in $1/V$ this quasilinear approximation coincides with Keizer's treatment of concentration fluctuations based on generalized fluctuation-dissipation assumptions.[10]

The quasilinear approximation should be useful from a practical point of view because it provides a relatively simple way of obtaining estimates of the magnitude of fluctuations. Since it predicts its own failure as a chemical instability is approached, it also functions as an early warning system for detection of regions in parameter space where fluctuations are potentially important. Thus a detailed sensitivity analysis of the quasilinear approximation is called for and will be developed here.

The first step in obtaining the quasilinear approximation consists in decomposing $\boldsymbol{\rho}$ into deterministic (or systematic) and fluctuating parts,

$$\rho_i(t) = u_i(t) + f_i(t), \tag{5.1}$$

where

$$\frac{du_i}{dt} = R_i(\mathbf{u}, \boldsymbol{\alpha}), \quad u_i(0) = \rho_i^0. \tag{5.2}$$

From Eq. (2.1) it follows that

$$\frac{df_i}{dt} = R_i(\mathbf{u} + \mathbf{f}, \boldsymbol{\alpha}) - R_i(\mathbf{u}, \boldsymbol{\alpha})$$
$$+ \sum_j P_{ij}(\mathbf{u} + \mathbf{f}, \boldsymbol{\alpha}) \xi_j(t), \quad f_i(0) = 0. \tag{5.3}$$

In order to obtain the quasilinear approximation to $\langle f_i(t) \rangle$, one averages both sides of Eq. (5.3) [using Eqs. (3.10) and (4.5)] to obtain

$$\frac{d}{dt} \langle f_i(t) \rangle = \langle R_i(\mathbf{u} + \mathbf{f}, \boldsymbol{\alpha}) - R_i(\mathbf{u}, \boldsymbol{\alpha}) \rangle$$
$$+ \langle T_i(\mathbf{u} + \mathbf{f}, \boldsymbol{\alpha}) \rangle, \tag{5.4}$$

where

$$T_i(\boldsymbol{\rho}(t), \boldsymbol{\alpha}) = \frac{1}{2} \sum_{jk} \frac{\partial P_{ij}}{\partial \rho_k}(\boldsymbol{\rho}, \boldsymbol{\alpha}) P_{kj}(\boldsymbol{\rho}, \boldsymbol{\alpha}). \tag{5.5}$$

Collecting terms up to first order in $1/V$ in Eq. (5.4) one gets

$$\frac{d}{dt} \langle f_i(t) \rangle = \sum_j \frac{\partial R_i}{\partial u_j}(\mathbf{u}, \boldsymbol{\alpha}) \langle f_j(t) \rangle$$
$$+ T_i(\mathbf{u}, \boldsymbol{\alpha}), \quad \langle f_i(0) \rangle = 0. \tag{5.6}$$

Thus, to first order in $1/V$,

$$\langle f_i(t) \rangle = \sum_j \int ds\, K_{ij}(t, s) T_j(\mathbf{u}(s), \boldsymbol{\alpha}(s)). \tag{5.7}$$

The Green's function $K_{ij}(t, s)$ satisfies

$$\frac{d}{dt} K_{ij}(t, s) = \sum_k \frac{\partial R_i}{\partial u_k}(\mathbf{u}(t), \boldsymbol{\alpha}(t)) K_{kj}(t, s) + \delta_{ij} \delta(t - s). \tag{5.8}$$

This Green's function plays a crucial role in the sensitivity analysis of deterministic chemical kinetics[4] and it turns out to play an important role in this approximate description of fluctuation phenomena.

The correlation function $C_{ij}(t, s) = \langle f_i(t) f_j(s) \rangle$ can also be obtained in a similar way. First $C_{ij}(t, t)$ should be obtained. From Eq. (5.3) it follows that to lowest order in $1/V$

$$\frac{d}{dt} C_{ij}(t, t) = \sum_k \left(\frac{\partial R_i}{\partial u_k}(\mathbf{u}, \boldsymbol{\alpha}) C_{kj}(t, t) + \frac{\partial R_j}{\partial u_k}(\mathbf{u}, \boldsymbol{\alpha}) C_{ki}(t, t) \right) + Q_{ij}(\mathbf{u}, \boldsymbol{\alpha}), \quad C_{ij}(0, 0) = 0. \tag{5.9}$$

In Eq. (5.9) Q_{ij} is the symmetric matrix defined in Eq. (2.5). With the help of the Green's function K, Eq. (5.9) yields

$$C_{ij}(t, t) = \sum_{kl} \int ds\, K_{ik}(t, s) Q_{kl}(\mathbf{u}(s), \boldsymbol{\alpha}(s)) K_{jl}(t, s). \tag{5.10}$$

For $t > s$ one has (again to lowest order in $1/V$)

$$\frac{d}{dt} C_{ij}(t, s) = \sum_k \frac{\partial R_i}{\partial u_k}(\mathbf{u}(t), \boldsymbol{\alpha}(t)) C_{kj}(t, s). \tag{5.11}$$

Equation (5.11) is to be solved subject to the condition that $C_{ij}(s, s)$ be given by Eq. (5.10), thus

$$C_{ij}(t, s) = \sum_k K_{ik}(t, s) C_{kj}(s, s). \tag{5.12}$$

The group property of the Green's function can be used to rewrite Eq. (5.12). This property implies the following identity,[4] $(t > s > r)$

$$\sum_k K_{ik}(t, s) K_{kj}(s, r) = K_{ij}(t, r). \tag{5.13}$$

The final expression for $C_{ij}(t, s)$ is

$$C_{ij}(t, s) = \sum_{kl} \int dr\, K_{ik}(t, r) Q_{kl}(\mathbf{u}(r), \boldsymbol{\alpha}(r)) K_{jl}(s, r). \tag{5.14}$$

Since $C_{ij}(t, s) = C_{ji}(s, t)$ it is clear that Eq. (5.14) is in fact valid for any t and s.

The breakdown of the quasilinear approximation near a chemical instability is directly related to the behavior of the Green's function $K_{ij}(t, s)$. This Green's function provides information on the stability of the solutions to Eq. (5.2). In the Lyapounov linearized stability analysis of kinetic equations like Eq. (5.2), one usually studies the eigenvalues of the matrix $(\partial R_i/\partial u_j)(\mathbf{u}, \boldsymbol{\alpha})$ evaluated at a particular steady state solution. The sign of the real part of the eigenvalues is the basic stability criterion.[1] Thus through Lyapounov's stability analysis a direct link is established between the asymptotic stability properties of $\mathbf{u}(t)$ and $K_{ij}(t, s)$. This connection between sensitivity analysis and linearized stability theory was investigated in detail in a study of limit cycles in chemical oscillators.[19]

The remainder of this section will be dedicated to obtaining expressions for the sensitivity coefficients and densities of $\langle f_i(t) \rangle$ and $C_{ij}(t, s)$ [Eqs. (5.7) and (5.14)]. Although the expressions derived in Secs. III and IV could be used it is simpler to take advantage of the explicit expressions for $\langle f_i \rangle$ and C_{ij} in terms of u and K. The procedure relies on the fact that the sensitivity coefficients and densities of u and K have been obtained in previous studies of the sensitivity analysis of deterministic chemical kinetics.[4] For completeness the expressions for these coefficients and densities in terms of the u and K themselves are transcribed here. They are obtained by taking appropriate derivatives of Eqs. (5.2) and (5.8) and solving the resulting linear equations with the help of the Green's function K. Thus,

$$S_{ip}(t) = \frac{\partial u_i(t)}{\partial \alpha_p} = \sum_j \lambda_{jp} K_{ij}(t, 0) + \sum_j \int ds\, K_{ij}(t, s) \frac{\partial R_j}{\partial \alpha_p}(\mathbf{u}(s), \boldsymbol{\alpha}), \tag{5.15}$$

$$S_{ip}(t, t') = \frac{\delta u_i(t)}{\delta \alpha_p(t')} = \sum_j (\lambda_{jp} \delta(t')) + \frac{\partial R_j}{\partial \alpha_p}(\mathbf{u}(t'), \boldsymbol{\alpha}(t')) K_{ij}(t, t'), \tag{5.16}$$

$$\frac{\partial}{\partial \alpha_p} K_{ij}(t, s) = \sum_{kl} \int ds'\, K_{ik}(t, s') K_{lj}(s', s) \times \left(A_p(s') \frac{\partial R_k}{\partial u_l}(\mathbf{u}(s'), \boldsymbol{\alpha}) \right), \tag{5.17}$$

$$\frac{\delta}{\delta \alpha_p(t')} K_{ij}(t, s) = \sum_{kl} \int ds'\, K_{ik}(t, s') K_{lj}(s', s) \times \left(A_p(s', t') \frac{\partial R_k}{\partial u_l}(\mathbf{u}(s'), \boldsymbol{\alpha}(s')) \right). \tag{5.18}$$

In Eqs. (5.17) and (5.18) $A_p(s')$ and $A_p(s', t')$ are differential operators that act on functions of $\mathbf{u}$ and $\boldsymbol{\alpha}$:

$$A_p(t) = \frac{\partial}{\partial \alpha_p} + \sum_k S_{kp}(t) \frac{\partial}{\partial u_k}, \tag{5.19}$$

$$A_p(t, t') = \delta(t - t') \frac{\partial}{\partial \alpha_p} + \sum_k S_{kp}(t, t') \frac{\partial}{\partial u_k}. \tag{5.20}$$

The sensitivity coefficients of interest here are

$$S_p^f = \frac{\partial}{\partial \alpha_p} \langle f_i(t) \rangle, \quad D_p^f(t') = \frac{\delta}{\delta \alpha_p(t')} \langle f_i(t) \rangle, \tag{5.21}$$

$$S_p^c = \frac{\partial}{\partial \alpha_p} C_{ij}(t, s), \quad D_p^c(t') = \frac{\delta}{\delta \alpha_p(t')} C_{ij}(t, s). \tag{5.22}$$

From Eq. (5.7) it follows, by differentiation, that

$$S_p^f = \sum_j \int ds\, K_{ij}(t, s) A_p(s) (T_j(\mathbf{u}(s), \boldsymbol{\alpha})) + \sum_j \int ds \left(\frac{\partial}{\partial \alpha_p} K_{ij}(t, s) \right) T_j(\mathbf{u}(s), \boldsymbol{\alpha}). \tag{5.23}$$

For $D_p^f(t')$ one obtains

$$D_p^f(t') = \sum_j \int ds\, K_{ij}(t, s) A_p(s, t') T_j(\mathbf{u}(s), \boldsymbol{\alpha}(s)) + \sum \int ds \left(\frac{\delta}{\delta \alpha_p(t')} K_{ij}(t, s) \right) T_j(\mathbf{u}(s), \boldsymbol{\alpha}(s)). \tag{5.24}$$

In an analogous manner the following expressions for S_p^c and $D_p^c(t')$ are obtained from Eq. (5.14),

$$S_p^c = \sum_{kl} \int ds'\, K_{ik}(t,s')K_{jl}(s,s')(A_p(s')Q_{kl}(\mathbf{u}(s'),\alpha)) + \sum_{kl} \int ds'\, Q_{kl}(\mathbf{u}(s'),\alpha)\Big(K_{ik}(t,s')\frac{\partial}{\partial\alpha_p}K_{jl}(s,s') + K_{jl}(s,s')\frac{\partial}{\partial\alpha_p}K_{ik}(t,s')\Big), \tag{5.25}$$

$$D_p^c(t') = \sum_{kl} \int ds'\, K_{ik}(t,s')K_{jl}(s,s')(A_p(s',t')Q_{kl}(\mathbf{u}(s'),\alpha(s'))) + \sum_{kl} \int ds'\, Q_{kl}(\mathbf{u}(s'),\alpha(s'))\Big(K_{ik}(t,s')\frac{\delta}{\delta\alpha_p(t')}K_{jl}(s,s') + K_{jl}(s,s')\frac{\delta}{\delta\alpha_p(t')}K_{ik}(t,s')\Big). \tag{5.26}$$

Thus in the quasilinear approximation it is possible to obtain closed expressions for $\langle \rho_i(t)\rangle$ and $\langle \rho_i(t)\rho_j(s)\rangle$ as well as for their sensitivity coefficients and densities in terms of the deterministic concentrations u and of the deterministic concentration response function K. Since powerful numerical methods are available to obtain both u and K (see Ref. 20) the results obtained in this section should prove useful in the analysis of chemical kinetic models. They allow for carrying out simultaneously both a sensitivity analysis and an investigation of fluctuation phenomena without the need of stochastic simulation in a computer.

VI. AN ILLUSTRATIVE APPLICATION

In this section a simple chemically reacting system is studied. The sensitivity coefficients of the average concentrations are computed using the expressions developed in Sec. III. Since the average concentrations can be obtained as explicit functions of time and of the system's parameters it is possible to compare the resulting sensitivity coefficients with their true expressions found by a direct calculation. This comparison provides a concrete example illustrating the validity of the results of Sec. III.

The reaction scheme to be studied here is

$$A \underset{k_2}{\overset{k_1}{\rightleftarrows}} B. \tag{6.1}$$

Let ρ_i be the concentration of A and ρ_2 the concentration of B. Then the matrix $\mathbf{P}(\rho,\alpha)$ is given by

$$\mathbf{P}(\rho,\alpha) = \left[\frac{1}{2V}(k_1\rho_1 + k_2\rho_2)\right]^{1/2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{6.2}$$

The equation for the concentration is

$$\frac{d\rho_i}{dt} = (-1)^i(k_1\rho_1 - k_2\rho_2) + \sum_{j=1}^{2} P_{ij}(\boldsymbol{\rho},\alpha)\xi_j(t), \qquad \rho_1(0) = A_0, \quad \rho_2(0) = B_0. \tag{6.3}$$

From Eq. (6.3) it follows that

$$\rho_1(t) + \rho_2(t) = \rho_1(0) + \rho_2(0). \tag{6.4}$$

Thus it is useful to introduce a new variable

$$\rho(t) = \rho_1(t) - \rho_1(0) = \rho_2(0) - \rho_2(t). \tag{6.5}$$

Then, for $\rho(t)$ one gets the stochastic differential equation

$$\frac{d\rho}{dt} = \alpha_1 - \alpha_2\rho + (\alpha_3 + \alpha_4\rho)^{1/2}\xi(t), \quad \rho(0) = 0. \tag{6.6}$$

In Eq. (6.6), $\xi(t)$ is a zero mean, delta correlated, Gaussian stochastic process (i.e., white noise) which is related to the ξ_j in Eq. (6.3) by

$$\xi(t) = (1/\sqrt{2})(\xi_1(t) - \xi_2(t)). \tag{6.7}$$

The parameters α_p ($p = 1, 2, 3, 4$) can be expressed in terms of the initial conditions and of the rate constants

$$\alpha_1 = k_2B_0 - k_1A_0, \quad \alpha_2 = k_1 + k_2, \quad \alpha_3 = (1/V)(k_1A_0 + k_2B_0), \quad \alpha_4 = (1/V)(k_1 - k_2). \tag{6.8}$$

Let $M_n(t) = \langle \rho^n(t)\rangle$, then from Eq. (6.6), using Eq. (3.10) and Eq. (4.5), one finds ($n \geqslant 1$)

$$\frac{d}{dt}M_n(t) = -n\alpha_2M_n(t) + n\Big(\alpha_1 + \frac{2n-1}{4}\alpha_4\Big)M_{n-1}(t) + \frac{n(n-1)}{2}\alpha_3M_{n-2}(t), \tag{6.9}$$

$M_n(0) = 0.$

Thus a recursion relation for the $M_n(t)$, $n \geqslant 1$, is obtained

$$M_n(t) = n\int_0^t ds\Big[\Big(\alpha_1 + \frac{2n-1}{4}\alpha_4\Big)M_{n-1}(s) + \frac{n(n-1)}{2}\alpha_3M_{n-2}(s)\Big]\exp -n\alpha_2(t-s). \tag{6.10}$$

Starting with $M_0(t) = 1$ all the $M_n(t)$ can be obtained from Eq. (6.10). The quantity of interest for obtaining the average concentrations is

$$M_1(t) = \langle \rho(t)\rangle = (1/\alpha_2)(\alpha_1 + \alpha_4/4)(1 - e^{-t\alpha_2}). \tag{6.11}$$

The general expression for $M_n(t)$ is

$$M_n(t) = \sum_{m=0}^{n} C_m^n e^{-mt\alpha_2}, \tag{6.12}$$

and a recursion relation for the C_m^n can be obtained from Eq. (6.10).

According to Eq. (3.12) the sensitivity coefficients of $\langle \rho(t)\rangle$ with respect to α_p are determined by

$$S_p^\rho = \int_0^\infty ds\Big\langle \rho(s)\xi(s^+)(\alpha_3 + \alpha_4\rho(s))^{-1/2} \times\Big(\delta_{1p} - \delta_{2p}\rho(s) + \xi(s)(\alpha_3 + \alpha_4\rho(s))^{-1/2} \times\frac{(\delta_{3p} + \delta_{4p}\rho(s))}{2}\Big)\Big\rangle. \tag{6.13}$$

In Eq. (6.13), δ_{qp} is the Kronecker delta function ($\delta_{qp} = 1$ if $p = q$, $\delta_{qp} = 0$ if $q \neq p$). Using the techniques discussed in Sec. IV, Eq. (6.13) can be rewritten as

$$S_p^\rho = \int dx \int dy\, x(\alpha_3 + \alpha_4 y)^{-1/2}[(\delta_{1p} - \delta_{2p}y)g_1(x,y,t) + \tfrac{1}{2}(\delta_{3p} + \delta_{4p}y)(\alpha_3 + \alpha_4 y)^{-1/2}g_2(x,y,t)], \tag{6.14}$$

where

$$g_1(x,y,t)=\int ds\langle\xi(s^+)\delta(x-\rho(t))\delta(y-\rho(s))\rangle, \quad (6.15)$$

$$g_2(x,y,t)=\int ds\langle\xi(s^+)\xi(s)\delta(x-\rho(t))\delta(y-\rho(s))\rangle, \quad (6.16)$$

Using Eq. (B1) plus some further manipulations yields

$$\frac{\partial}{\partial t}g_1(x,y,t)$$
$$=-\frac{\partial}{\partial x}\Big[(\alpha_1-\alpha_2 x)g_1(x,y,t)$$
$$-\frac{1}{2}(\alpha_3+\alpha_4 x)^{1/2}\frac{\partial}{\partial x}((\alpha_3+\alpha_4 x)^{1/2}g_1(x,y,t))\Big]$$
$$-\Big(\frac{\partial}{\partial x}\delta(x-y)\Big)(\alpha_3+\alpha_4 y)^{1/2}P_1(y,t),$$
$$g_1(x,y,0)=0, \quad (6.17)$$

$$\frac{\partial}{\partial t}g_2(x,y,t)$$
$$=-\frac{\partial}{\partial x}\Big[(\alpha_1-\alpha_2 x)g_2(x,y,t)$$
$$-\frac{1}{2}(\alpha_3+\alpha_4 x)^{1/2}\frac{\partial}{\partial x}((\alpha_3+\alpha_4 x)^{1/2}g_2(x,y,t))\Big]$$
$$-\frac{\partial}{\partial x}\Big[(\alpha_3+\alpha_4 x)^{1/2}\Big(\frac{\partial}{\partial x}((\alpha_3+\alpha_4 x)^{1/2}\delta(x-y))\Big)$$
$$\times P_1(y,t)+\frac{1}{2}\frac{\partial}{\partial y}((\alpha_3+\alpha_4 y)^{1/2}\delta(x-y)P_1(y,t))\Big],$$
$$g_2(x,y,0)=0. \quad (6.18)$$

As for boundary conditions, both $g_1(x,y,t)$ and $g_2(x,y,t)$ must vanish fast enough as $|x|$ and $|y|$ grow large such that all necessary integrals are finite. This, of course, is insured by the definition of $g_1(x,y,t)$ and $g_2(x,y,t)$, Eqs. (6.15) and (6.16). It is convenient now to introduce the functions $h_1(y,t)$ and $h_2(y,t)$ defined by

$$h_i(y,t)=(\alpha_3+\alpha_4 y)^{-1/2}\int dx(xg_i(x,y,t)). \quad (6.19)$$

From Eqs. (6.17) and (6.18) it follows that

$$\frac{\partial}{\partial t}h_1(y,t)=-\alpha_2 h_1(y,t)+P_1(y,t), \quad h_1(y,0)=0, \quad (6.20)$$

$$\frac{\partial}{\partial t}h_2(y,t)=-\alpha_2 h_2(y,t)$$
$$+\frac{1}{2}(\alpha_3+\alpha_4 y)^{1/2}\frac{\partial}{\partial y}P_1(y,t),$$
$$h_2(y,0)=0. \quad (6.21)$$

Thus

$$h_1(y,t)=\int_0^t ds\,P_1(y,s)\exp-\alpha_2(t-s), \quad (6.22)$$

$$h_2(y,t)=\frac{1}{2}(\alpha_3+\alpha_4 y)^{1/2}\frac{\partial}{\partial y}h_1(y,t). \quad (6.23)$$

Therefore, for S_p^ρ one gets

$$S_p^\rho=\int dy\Big(\delta_{1p}-\delta_{2p}y+\frac{1}{4}(\delta_{3p}+\delta_{4p}y)\frac{\partial}{\partial y}\Big)h_1(y,t). \quad (6.24)$$

The final result for S_p^ρ is

$$S_p^\rho=\frac{1}{\alpha_2}\Big(\delta_{1p}+\frac{1}{4}\delta_{p4}\Big)(1-e^{-\alpha_2 t})$$
$$+\delta_{2p}\frac{1}{\alpha_2}\Big(\alpha_1+\frac{\alpha_4}{4}\Big)te^{-\alpha_2 t}-\frac{1}{\alpha_2}(1-e^{-\alpha_2 t}). \quad (6.25)$$

Since by definition $S_p^\rho=(\partial/\partial\alpha_p)\langle\rho(t)\rangle$ it is easy to verify by differentiation of Eq. (6.11) that Eq. (6.25) is correct. It is straightforward to verify that Eq. (3.12) also correctly yields the sensitivity coefficients of higher correlation functions like $\langle\rho(t)\rho(t')\rangle$ for example. However, since the algebraic work necessary for a more complete verification is rather involved it will not be presented here.

VII. CONCLUDING REMARKS

The main result of this work is the development of a sensitivity analysis formalism to study kinetic models described by stochastic differential equations with multiplicative and additive white noise. The formulas for the sensitivity coefficients and densities can be generalized to be applicable to the case of colored noise provided that the colored noise itself can be obtained from a stochastic differential equation driven by white noise. This is the case for two commonly used colored noises, the Ornstein–Uhlenbeck process[21] and the Brownian oscillator process.[22]

Phenomenological stochastics models of the type used here to describe concentration fluctuations in chemically reacting systems are usually obtained by starting from a deterministic description of the system and then demanding that certain fluctuation-dissipation relations be satisfied in order to obtain the correct form for the stochastic terms.[10] In the case of chemical kinetics the deterministic description is provided by the rate equations derived from the mass action law.[1] Thus the results obtained from any stochastic model of chemical kinetics must coincide with those obtained from the deterministic model whenever the latter provides a valid description of the behavior of the chemically reacting system. Therefore, one might argue not to bother with the study and sensitivity analysis of stochastic models since it appears sufficient to study the simpler deterministic model. The fallacy of this statement is, of course, connected to the fact that the deterministic model based on the mass action rate equations is not always valid. For example, if the chemical reactions involve a small number of molecules inside a small volume, as is the case for reactions in micelles, then fluctuations become very important. But even in macroscopic systems there are reaction mechanisms that allow for the existence of chemical instabilities leading to the amplification of fluctuations.[1,2] Moreover, due to the randomness of the molecular events responsible for the chemical reactions, it is clear that the concentrations of the chemical species must be represented by stochastic variables. The moments of a finite set of stochastic variables in general do not satisfy a finite, closed set of equations. Thus the interpretation of the mass action rate equations as a closed system of equations for the

average concentrations shows clearly that this model neglects fluctuations and cannot be expected to provide an adequate description of the behavior of systems where fluctuations are expected to be important.

The sensitivity analysis technique developed in this work should enhance the stock of mathematical modelling tools available in chemical kinetics. In particular the quasilinear approximation should be specially useful. It provides a unified framework in which estimates of the magnitude of the fluctuations can be obtained jointly with a sensitivity analysis of both the deterministic rate equations and the fluctuations themselves. In this approximation the fluctuation induced corrections to the concentration values predicted by the deterministic rate equations as well as concentration correlation functions can be obtained, even for complex systems, without carrying out time consuming stochastic simulations.

The results for the sensitivity coefficients were generalized to the case of time-dependent parameters. This demanded the introduction of sensitivity densities which can also be interpreted as generalized response functions. This interpretation sheds light on the physical meaning of sensitivity analysis and shows that the sensitivity coefficients and densities belong to a broad class of mathematical objects (i.e., response functions) of considerable interest in statistical mechanics.

ACKNOWLEDGMENTS

The authors acknowledge support for this research from the Office of Naval Research.

APPENDIX A

A simple derivation of Novikov's theorem can be obtained by using path integral techniques (see, for example, the Appendix of Ref. 16). For the benefit of readers not familiar with the path integral formalism we provide here a derivation of this important result that uses only the functional derivative concept.

Let $Z[\mathbf{J}]$ be the characteristic functional associated with the Gaussian stochastic process ξ. By definition,[7]

$$Z[\mathbf{J}] = \left\langle \exp\left(i \sum_j \int dt\, J_j(t) \xi_j(t) \right) \right\rangle = \exp\left(-\frac{1}{2} \sum_j \int dt\, J_j^2(t) \right). \tag{A1}$$

The first equality above is just the definition of $Z[\mathbf{J}]$, the second is restricted to the case of white noise (i.e., a δ-correlated Gaussian stochastic process). For an arbitrary functional $H[\xi]$ it follows that

$$\langle H[\xi] \rangle = H\left[-i \frac{\delta}{\delta \mathbf{J}} \right] Z[\mathbf{J}] \Big|_{\mathbf{J}=0}, \tag{A2}$$

that is, $H[\xi]$ can be obtained by taking appropriate functional derivatives of $Z[\mathbf{J}]$ and then setting $\mathbf{J} = 0$. Thus,

$$\langle \xi_k(s) H[\xi] \rangle = \left(-i \frac{\delta}{\delta J_k(s)} \right) H\left[-i \frac{\delta}{\delta \mathbf{J}} \right] Z[\mathbf{J}] \Big|_{\mathbf{J}=0}. \tag{A3}$$

From Eq. (A1),

$$-i \frac{\delta}{\delta J_k(s)} H\left[-i \frac{\delta}{\delta \mathbf{J}} \right] Z[\mathbf{J}] = iH\left[-i \frac{\delta}{\delta \mathbf{J}} \right] (J_k(s) Z[\mathbf{J}]). \tag{A4}$$

Defining $P_j(t) = -i\,\delta/\delta J_j(t)$ and interpreting $\mathbf{P}$ and $H[\mathbf{P}]$ as operators acting on functionals of $\mathbf{J}$, the right-hand side of Eq. (A4) can be written as

$$iH[\mathbf{P}](J_k(s) Z[\mathbf{J}]) = i[H[\mathbf{P}], J_k(s)] Z[\mathbf{J}] + iJ_k(s) H[\mathbf{P}] Z[\mathbf{J}]. \tag{A5}$$

Now

$$[P_j(t), J_k(s)] = -i\delta_{jk}\delta(t-s), \tag{A6}$$

therefore,

$$[H[\mathbf{P}], J_k(s)] = -i \frac{\delta H[\mathbf{P}]}{\delta P_k(s)}. \tag{A7}$$

One obtains Eq. (A7) from Eq. (A6) in the same way that in quantum mechanics one gets $[f(p), x] = -i\hbar(\partial f(p)/\partial p)$ from the basic commutation relation $[p, x] = -i\hbar$, $p = -i\hbar(\partial/\partial x)$. Thus,

$$\langle \xi_k(s) H[\xi] \rangle = \left\{ \frac{\delta H[\mathbf{P}]}{\delta P_k(s)} \Big|_{\mathbf{P} = -i(\delta/\delta \mathbf{J})} Z[\mathbf{J}] \right\} \Big|_{\mathbf{J}=0} = \left\langle \frac{\delta H[\xi]}{\delta \xi_k(s)} \right\rangle. \tag{A8}$$

This result is known as Novikov's theorem.[23]

APPENDIX B

In order to obtain Eq. (4.11) (with $t_1 > t_2 > \cdots > t_n$) one should notice that the following identity holds[16]:

$$\frac{\partial}{\partial t} \delta(\mathbf{x} - \boldsymbol{\rho}(t)) = D(\mathbf{x}, t) \delta(\mathbf{x} - \boldsymbol{\rho}(t)) + R(x, \boldsymbol{\rho}(t), \boldsymbol{\xi}(t)), \tag{B1}$$

where $D(\mathbf{x}, t)$ is the Fokker–Planck operator, defined by

$$D(\mathbf{x}, t) f(\mathbf{x}, t) = -\sum_i \frac{\partial}{\partial x_i} \left[R_i(\mathbf{x}, \boldsymbol{\alpha}(t)) f(\mathbf{x}, t) - \frac{1}{2} \sum_{jk} P_{ij}(\mathbf{x}, \boldsymbol{\alpha}(t)) \times \frac{\partial}{\partial x_k} (P_{kj}(\mathbf{x}, \boldsymbol{\alpha}(t)) f(\mathbf{x}, t)) \right]. \tag{B2}$$

For the remaining term in the right-hand side of Eq. (B1) one has

$$\begin{aligned} R(\mathbf{x}, \boldsymbol{\rho}(t), \boldsymbol{\xi}(t)) &= -\sum_i \frac{\partial}{\partial x_i} (R_i(\boldsymbol{\rho}(t), \boldsymbol{\alpha}(t)) - R_i(x, \boldsymbol{\alpha}(t))) \delta(\mathbf{x} - \boldsymbol{\rho}(t)) \\ &\quad - \frac{1}{2} \sum_{ijk} \frac{\partial}{\partial x_i} (P_{ij}(\mathbf{x}, \boldsymbol{\alpha}(t))) \\ &\quad \times \frac{\partial}{\partial x_k} (P_{kj}(\mathbf{x}, \boldsymbol{\alpha}(t)) \delta(\mathbf{x} - \boldsymbol{\rho}(t))) \\ &\quad - \sum_{ij} \frac{\partial}{\partial x_i} (P_{ij}(\boldsymbol{\rho}(t), \boldsymbol{\alpha}(t)) \xi_j(t) \delta(\mathbf{x} - \boldsymbol{\rho}(t))). \end{aligned} \tag{B3}$$

Despite its rather involved expression $R(\mathbf{x}, \boldsymbol{\rho}(t), \boldsymbol{\xi}(t))$ has some simple and useful properties, for example

$$\langle R(\mathbf{x}, \boldsymbol{\rho}(t), \boldsymbol{\xi}(t))\rangle = 0. \tag{B4}$$

Another important property is the following. Suppose that $F[\boldsymbol{\rho}]$ is a functional of $\boldsymbol{\rho}$ that depends only on the $\rho_i(s)$ with $s < t$, then

$$\langle R(\mathbf{x}, \boldsymbol{\rho}(t), \boldsymbol{\xi}(t))F[\boldsymbol{\rho}]\rangle = 0. \tag{B5}$$

The contribution from the first term in the right-hand side of Eq. (B3) is trivial since, upon averaging, the factor in front of $\delta(\mathbf{x} - \boldsymbol{\rho}(t))$ is identically zero. Thus,

$$\begin{aligned}\langle R(\mathbf{x}, \boldsymbol{\rho}(t)), \boldsymbol{\xi}(t)F[\boldsymbol{\rho}]\rangle &= -\frac{1}{2}\sum_{ijk}\frac{\partial}{\partial x_i}\Big(P_{ij}(\mathbf{x}, \boldsymbol{\alpha}(t)) \\ &\times \frac{\partial}{\partial x_k}P_{kj}(\mathbf{x}, \boldsymbol{\alpha}(t))\langle\delta(\mathbf{x} - \boldsymbol{\rho}(t))F[\boldsymbol{\rho}]\rangle\Big) \\ &- \sum_{ij}\frac{\partial}{\partial x_i}(P_{ij}(\mathbf{x}, \boldsymbol{\alpha}(t))\langle\xi_j(t)\delta(\mathbf{x} - \boldsymbol{\rho}(t))F[\boldsymbol{\rho}]\rangle).\end{aligned} \tag{B6}$$

Now, using Eqs. (3.10) and (4.5)

$$\begin{aligned}\langle\xi_j(t)\delta(\mathbf{x} - \boldsymbol{\rho}(t))F[\rho]\rangle &= -\frac{1}{2}\sum_k\frac{\partial}{\partial x_k}(P_{kj}(\mathbf{x}, \boldsymbol{\alpha}(t))\langle\delta(\mathbf{x} - \boldsymbol{\rho}(t))F[\boldsymbol{\rho}]\rangle) \\ &+ \sum_k\int ds\left\langle\delta(\mathbf{x} - \boldsymbol{\rho}(t))\frac{\delta F[\boldsymbol{\rho}]}{\delta\rho_k(s)}\frac{\delta\rho_k(s)}{\delta\xi_j(t)}\right\rangle.\end{aligned} \tag{B7}$$

The definition of $F[\boldsymbol{\rho}]$ implies that $\delta F[\boldsymbol{\rho}]/\delta\rho_k(s) = 0$ for $s \geqslant t$ and Eq. (3.7) states that $\delta\rho_k(s)/\delta\xi_j(t) = 0$ for $t > s$, therefore, the second term in the right-hand side of Eq. (B7) vanishes. The remaining term in the right-hand side of Eq. (B7) exactly cancels the first term in the right-hand side of Eq. (B6) showing that Eq. (B5) is satisfied.

Introducing the Green's function associated with the Fokker–Planck operator, $G(\mathbf{x}, t; \mathbf{y}, t')$,

$$\frac{\partial}{\partial t}G(\mathbf{x}, t; \mathbf{y}, t') = D(\mathbf{x}, t)G(\mathbf{x}, t; \mathbf{y}, t') + \delta(t - t')\delta(\mathbf{x} - \mathbf{y}), \tag{B8}$$

$$\lim_{\epsilon\to 0^+} G(x, t + \epsilon; \mathbf{y}, t) = \delta(\mathbf{x} - \mathbf{y}), \tag{B9}$$

Eq. (B1) can be rewritten as

$$\delta(\mathbf{x} - \boldsymbol{\rho}(t)) = \int dx'\, G(\mathbf{x}, t; \mathbf{x}', s)\delta(\mathbf{x}' - \boldsymbol{\rho}(s)) + \int_s^\infty dt' \int d\mathbf{x}'\, G(\mathbf{x}, t; x', t')R(x', \boldsymbol{\rho}(t'), \boldsymbol{\xi}(t')). \tag{B10}$$

In Eq. (B10) $t > s$. Consider now $P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n)$ with $t_1 > t_2 > \cdots > t_n$. From Eq. (4.9) one has

$$P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) = \langle\delta(\mathbf{y}_1 - \boldsymbol{\rho}(t_1))F(\boldsymbol{\rho}(t_2), \ldots, \boldsymbol{\rho}(t_n))\rangle, \tag{B11}$$

where

$$F(\boldsymbol{\rho}(t_2), \ldots, \boldsymbol{\rho}(t_n)) = \delta(\mathbf{y}_2 - \boldsymbol{\rho}(t_2))\cdots\delta(\mathbf{y}_n - \boldsymbol{\rho}(t_n)). \tag{B12}$$

Setting $\mathbf{x} = \mathbf{y}_1$, $t = t_1$, and $s = t_2$ in Eq. (B10) and inserting the resulting expression for $\delta(y_1 - \boldsymbol{\rho}(t_1))$ in the right-hand side of Eq. (B11) one obtains

$$P_n(\mathbf{y}_1, t_1; \ldots; \mathbf{y}_n, t_n) = \int d\mathbf{x}'\, G(\mathbf{y}_1, t_1; x', t_2)P_n(\mathbf{x}', t_2; \mathbf{y}_2, t_2; \ldots; \mathbf{y}_n, t_n). \tag{B13}$$

Equation (B5) was used to arrive at Eq. (B13). Notice that the multiple time distribution in the right-hand side of Eq. (B13) has two identical time variables. Thus using Eq. (4.10) one obtains Eq. (4.11).

[1] G. Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium Systems* (Wiley, New York, 1977).
[2] M. Mangel, J. Chem. Phys. **69**, 3697 (1978).
[3] P. M. Franck, *Introduction to System Sensitivity Theory* (Academic, New York, 1978).
[4] J. Hwang, E. Dougherty, S. Rabitz, and H. Rabitz, J. Chem. Phys. **69**, 5180 (1978).
[5] H. Rabitz, Computers and Chemistry **5**, 167 (1981).
[6] J. Tilden, V. Costanza, G. McRae, and J. Seinfeld, in *Modelling of Chemical Reaction Systems*, edited by K. Ebert, D. Denflhard, and W. Jager, *Springer Series in Chemical Physics*, Vol. 18 (Springer-Verlag, New York, 1981).
[7] R. F. Fox, Phys. Rep. **48**, 179 (1978).
[8] N. G. van Kampen, J. Statist. Phys. **24**, 175 (1981).
[9] Z. Schuss, *Theory and Application of Stochastic Differential Equations* (Wiley, New York, 1980).
[10] (a) J. Keizer, J. Chem. Phys. **63**, 5037 (1975); (b) S. Grossman, J. Chem. Phys. **65**, 2007 (1976).
[11] D. McQuarrie, J. Appl. Probab. **4**, 413 (1967).
[12] L. Arnold, in *Dynamics of Synergetic Systems*, edited by H. Haken (Springer-Verlag, New York, 1980).
[13] (a) H. Grabert and M. S. Green, Phys. Rev. A **19**, 1747 (1979); (b) H. Grabert, R. Graham, and M. S. Green, Phys. Rev. A **21**, 2136 (1980); (c) H. Grabert, P. Hänggi, and I. Oppenheim, Physica A **117**, 300 (1983).
[14] M. Demiralp and H. Rabitz, J. Chem. Phys. **74**, 3362 (1981).
[15] P. C. Martin, E. D. Siggia, and H. A. Rose, Phys. Rev. A **8**, 423 (1973).
[16] S. Ma and G. F. Mazenko, Phys. Rev. B **11**, 4077 (1975).
[17] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).
[18] D. K. Dacol and H. Rabitz, J. Chem. Phys. **78**, 4905 (1983).
[19] R. Larter, H. Rabitz, and M. Kramer, J. Chem. Phys. **80**, 4120 (1984).
[20] M. A. Kramer, J. M. Calo, and H. Rabitz, Appl. Math. Modelling **5**, 432 (1981).
[21] L. S. Ornstein and G. E. Uhlenbeck, Phys. Rev. **36**, 823 (1930).
[22] M. C. Wang and G. E. Uhlenbeck, Rev. Mod. Phys. **17**, 323 (1945).
[23] E. A. Novikov, Zh. Eksp. Teor. Fiz. **47**, 1919 (1964) [Sov. Phys. JETP **20**, 1290 (1965)].

The Gauss and the Weingarten equations for extended objects and their spinorization

Y. Takahashi
Theoretical Physics Institute, Department of Physics, University of Alberta, Edmonton, Alberta, T6G 2J1, Canada

(Received 16 November 1983; accepted for publication 23 March 1984)

Gürsey's program of geometrization and spinorization of motion of the classical point particle is generalized to the objects with spatial extension. For this purpose, the Gauss and the Weingarten equations in differential geometry were adopted. Since the spinorized version of the Gauss and the Weingarten equations is much less stringent than the original equations, it is expected that it serves to provide us with a way in which solutions are easily obtained.

PACS numbers: 11.10. − z, 11.90. + t, 02.40. + m

I. INTRODUCTION

The motion of the classical point particle is treated by the Newton equation (or its relativistic version). The fundamental variables are the position vector of the particle and their time evolution is described by the Newton equation. In quantum mechanics, the behavior of a particle is described by the Schrödinger equation (or by the Dirac equation). As has been argued by Gürsey,[1] the conceptual gap between the two different approaches may be narrowed and the comparison between the classical and quantum descriptions may become easier, if one could introduce a spinor in classical mechanics.

Following the above line of reasoning, Gürsey described a classical point particle in terms of the Frenet–Serret equation and showed that it can be represented equivalently by spinor equations. The essence of Gürsey's spinorization lies, however, in the geometrization of the motion of a point particle in terms of a tetrad rather than a mere spinorization of the classical particle.

The close relation between a tetrad (or a triad) and a spinor has been exploited by many authors in different contexts. For instance, a number of Japanese authors recognized the convenience of using spinor variables in constructing a rigid sphere model of high energy particles.[2] As has been discussed by us in previous occasions, the existence of a tetrad (or a triad) implies that of a spinor and vice versa, and the link between the tetrad and the spinor is provided by the Fierz identities.[3]

On the other hand, there has been considerable interest in attempts to treating extended objects such as string models in high energy physics and solitons, and various interesting techniques and results are being obtained.[4,5] In particular, a number of authors investigated the soliton problems from the geometrical point of view to gain an insight into apparently different models.[4]

The purpose of this paper is to present a general geometrical scheme for the description of extended objects in terms of the Gauss and the Weingarten equations and to show that these equations can be replaced by spinor equations. This may be considered, on the one hand, as a generalization of Gürsey's program mentioned earlier, it is hoped, on the other hand, that the general formalism will provide us with a method in which extended objects are accommodated in relativistic field theory in a natural way.

As an elementary orientation, we first review in the next section the Frenet–Serret formula for an orbit of a point particle, and impose on it an evolution equation. The integrability of the equations will be discussed briefly. Having geometrized the motion of a point particle, we show in Sec. III that the Frenet–Serret equation and the evolution equation can be reduced to spinor equations (three-dimensional spinor). It is also shown that the integrability condition of the spinor equations agrees entirely with that of the original Frenet–Serret and the evolution equations. Section IV is devoted to a general geometrical formalism of extended objects. In this formalism the motion of tangent vectors of the extended objects is given by the Gauss equation whereas that of normal vectors is given by the Weingarten equation.[5] Thus, the tetrad formed by the tangents and the normals can be reduced to a four-dimensional spinor. Hence, the Gauss and the Weingarten equations can be replaced by spinor equations, as will be shown in Sec. V. Again, the integrability condition for the spinor equations is compatible with that of the Gauss and the Weingarten equations, though they are not exactly the same.

The general formalism presented in Secs. IV and V will be illustrated in Sec. VI explicitly in the case of a string. We shall discuss in the last section some consequences and further problems associated with our formalism.

II. THE FRENET–SERRET EQUATION

Let us consider a string of an arbitrary shape in three-dimensional space. In an arbitrary point on the string, we set up an orthogonal triad consisting of the unit tangent $\mathbf{a}^{(1)}$, the normal $\mathbf{a}^{(2)}$, and the binormal $\mathbf{a}^{(3)}$. The shape of the string is described by the orthonormal triad satisfying the Frenet–Serret equation

$$\mathbf{a}^{(1)\prime} = \kappa \mathbf{a}^{(2)}, \tag{2.1a}$$

$$\mathbf{a}^{(2)\prime} = -\kappa \mathbf{a}^{(1)} + \tau \mathbf{a}^{(3)}, \tag{2.1b}$$

$$\mathbf{a}^{(3)\prime} = -\tau \mathbf{a}^{(2)}, \tag{2.1c}$$

where κ and τ are the curvature and the torsion of the string, respectively. The prime stands for the derivative with re-

 0022-2488/84/092728-08$02.50

spect to the length of the string measured from a specified point on the string, which is indicated by ξ^1.

We note that the Frenet–Serret equation (2.1) can be written in the form,

$$\frac{\partial \mathbf{a}^{(i)}}{\partial \xi^1} = \omega_1^{ij}\mathbf{a}^{(j)}, \tag{2.2}$$

with

$$\hat{\omega}_1 \equiv [\omega_1^{ij}] = \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix}, \tag{2.3}$$

where in Eq. (2.2) the summation over j from 1 to 3 is understood.

In order to describe the motion of the string, we consider the equation of the form

$$\frac{\partial \mathbf{a}^{(i)}}{\partial \xi^0} = \omega_0^{ij}\mathbf{a}^{(j)}, \tag{2.4}$$

where ξ^0 is the time variable and ω_0^{ij} is a matrix skew symmetric with respect to i and j, so that the norm and the orthogonality of the triad is preserved.

Since the vector $\mathbf{a}^{(i)}$ should be determined uniquely from Eqs. (2.2) and (2.4), the integrability condition

$$\left[\frac{\partial}{\partial \xi^1} - \hat{\omega}_1, \frac{\partial}{\partial \xi^0} - \hat{\omega}_0\right] = 0 \tag{2.5}$$

must be satisfied. The square brackets in Eq. (2.5) stand for commutator. Using the explicit form (2.3), we obtain, from (2.5), the integrabilty condition

$$\frac{\partial \omega_0^{12}}{\partial \xi^1} - \frac{\partial \kappa}{\partial \xi^0} - \tau\omega_0^{13} = 0, \tag{2.6a}$$

$$\frac{\partial \omega_0^{13}}{\partial \xi^1} - \kappa\omega_0^{23} - \tau\omega_0^{21} = 0, \tag{2.6b}$$

$$\frac{\partial \omega_0^{32}}{\partial \xi^1} + \frac{\partial \tau}{\partial \xi^0} + \kappa\omega_0^{31} = 0. \tag{2.6c}$$

The skew symmetric quantity $\hat{\omega}_0 = [\omega_0^{ij}]$ is so far arbitrary, as long as it satisfies Eq. (2.6). If we take[6]

$$\hat{\omega}_0 = \begin{pmatrix} 0 & -\kappa\tau & \kappa' \\ \kappa\tau & 0 & \dfrac{\kappa''}{\kappa} - \tau^2 \\ -\kappa' & \tau^2 - \dfrac{\kappa''}{\kappa} & 0 \end{pmatrix}, \tag{2.7}$$

the integrability condition (2.6) becomes

$$\dot{\kappa} = -2\kappa'\tau - \kappa\tau', \tag{2.8a}$$

$$\dot{\tau} = \left(\frac{\kappa''}{\kappa} - \tau^2\right)' + \kappa\kappa'. \tag{2.8b}$$

It can easily be shown that Eqs. (2.2) and (2.4) with (2.3) and (2.7) are reduced to the equation[4]

$$\dot{\mathbf{a}}^{(1)} = \mathbf{a}^{(1)} \times \mathbf{a}^{(1)\prime}, \tag{2.9}$$

which is the continuum limit of the one-dimensional version of the Heisenberg ferromagnet model with the nearest-neighbor interactions.

We point out that the equation of motion such as (2.9), the Frenet–Serret equation (2.2) and their integrability condition have proven to be a useful way to find a new equation. For instance, in the example given above, the integrability condition (2.8) leads to a nonlinear Schrödinger equation

$$i\dot{\psi} = -\psi'' - \tfrac{1}{2}|\psi|^2\psi. \tag{2.10}$$

To show that Eq. (2.8) is equivalent to (2.10), we put, following Hasimoto,

$$\psi \equiv \kappa \exp\left\{i\int^{\xi^1} d\xi^1\,\tau\right\}. \tag{2.11}$$

Then, we have

$$i\dot{\psi} = \left\{i\frac{\dot{\kappa}}{\kappa} - \int^{\xi^1} d\xi^1\,\dot{\tau}\right\}\psi. \tag{2.12}$$

Upon using (2.8b), Eq. (2.12) becomes

$$i\dot{\psi} = \{-2i\tau\kappa'/\kappa - i\tau' - \kappa''/\kappa + \tau^2 - \tfrac{1}{2}\kappa^2\}\psi. \tag{2.13}$$

On the other hand, from Eq. (2.11)

$$\psi' = (\kappa'/\kappa + i\tau)\psi, \tag{2.14}$$

$$\psi'' = (\kappa'/\kappa + i\tau)'\psi + (\kappa'/\kappa + i\tau)^2\psi,$$

$$= (\kappa''/\kappa + i\tau' + 2i\tau\kappa'/\kappa - \tau^2)\psi. \tag{2.15}$$

Combining (2.15) and (2.13), we arrive at Eq. (2.10).

III. SPINORIZATION AND INTEGRABILITY CONDITION

As has been discussed in previous occasions[7] the triad $\mathbf{a}^{(1)}$, $\mathbf{a}^{(2)}$, and $\mathbf{a}^{(3)}$ implies the existence of a two-component spinor indicated by ϕ. If we define the charge conjugation of ϕ by

$$\phi^c \equiv \sigma_2\phi^*, \tag{3.1}$$

where σ_2 is the second member of the Pauli spin matrices, and * the complex conjugate, then the triad can be represented by

$$a_i^{(1)} \equiv J_i/J, \tag{3.2a}$$

$$a_i^{(2)} \equiv (R_i + R_i^*)/2J, \tag{3.2b}$$

$$a_i^{(3)} \equiv (R_i - R_i^*)/2iJ, \tag{3.2c}$$

with

$$J \equiv (\phi^\dagger\phi), \tag{3.3a}$$

$$J_i \equiv (\phi^\dagger\sigma_i\phi), \tag{3.3b}$$

$$R_i \equiv (\phi^{c\dagger}\sigma_i\phi), \tag{3.3c}$$

$$R_i^* \equiv (\phi^\dagger\sigma_i\phi^c). \tag{3.3d}$$

Due to the Fierz identities

$$(\sigma_A)_{\alpha\beta}(\sigma_B)_{\lambda\rho} = \frac{1}{4}\sum_{C,D}\mathrm{Tr}[\sigma_A\sigma_C\sigma_B\sigma_D](\sigma_C)_{\lambda\beta}(\sigma_D)_{\alpha\rho}, \tag{3.4}$$

where

$$\sigma_A, \sigma_B \text{ etc.} = I, \sigma_1, \sigma_2, \sigma_3, \tag{3.5}$$

the quantities defined in (3.3) satisfy the relations

$$J_iJ_i = \tfrac{1}{2}R_i\bar{R}_i = J^2, \tag{3.6a}$$

$$JR_i = i\epsilon_{ijk}J_jR_k, \tag{3.6b}$$

$$\bar{R}_iR_j + \bar{R}_jR_i = 2(\delta_{ij}J^2 - J_iJ_j), \tag{3.6c}$$

which ensure orthonormality and completeness of the triad (3.2). Noting the relations

$$\phi^{c\dagger}\phi = 0, \tag{3.7a}$$

$$\phi^{c\dagger}\phi^{c} = \phi^{\dagger}\phi, \tag{3.7b}$$

$$\phi^{c\dagger}\sigma_i\phi^{c} = -\phi^{\dagger}\sigma_i\phi, \tag{3.7c}$$

we can spinorize the Frenet–Serret equation (2.2) with (2.3) as

$$\frac{\partial\phi}{\partial\xi^1} = -\frac{i}{2}\tau\phi + \frac{\kappa}{2}\phi^{c}, \tag{3.8a}$$

$$\frac{\partial\phi^{c}}{\partial\xi^1} = \frac{i}{2}\tau\phi^{c} - \frac{\kappa}{2}\phi. \tag{3.8b}$$

The evolution equation (2.4) can be spinorized as

$$\frac{\partial\phi}{\partial\xi^0} = -\frac{i}{2}\omega_0^{23}\phi + \frac{1}{2}(\omega_0^{12} + i\omega_0^{13})\phi^{c}, \tag{3.9a}$$

$$\frac{\partial\phi^{c}}{\partial\xi^0} = \frac{i}{2}\omega_0^{23}\phi^{c} - \frac{1}{2}(\omega_0^{12} - i\omega_0^{13})\phi. \tag{3.9b}$$

To demonstrate that Eq. (3.8) can reproduce the Frenet–Serret equation (2.2), we first calculate, using (3.8a) and its conjugate

$$\frac{\partial}{\partial\xi^1}(\phi^{\dagger}\phi) = \frac{\partial\phi^{\dagger}}{\partial\xi^1}\phi + \phi^{\dagger}\frac{\partial\phi}{\partial\xi^1} = \frac{i}{2}\tau\phi^{\dagger}\phi - \frac{i}{2}\tau\phi^{\dagger}\phi = 0, \tag{3.10}$$

where the relation (3.7a) has been used. Similarly, we obtain

$$\frac{\partial}{\partial\xi^1}(\phi^{\dagger}\sigma_i\phi) = \frac{i}{2}\tau\phi^{\dagger}\sigma_i\phi - \frac{i}{2}\tau\phi^{\dagger}\sigma_i\phi + \frac{1}{2}\kappa\phi^{c\dagger}\sigma_i\phi + \frac{1}{2}\kappa\phi^{\dagger}\sigma_i\phi^{c} = \kappa\{(\phi^{c\dagger}\sigma_i\phi) + (\phi^{\dagger}\sigma_i\phi^{c})\}/2. \tag{3.11}$$

Dividing this equation by J and using Eq. (3.10), we arrive at the first equation (2.1a). The other equations (2.1b) and (2.1c) can be obtained in the same fashion. Equations (3.9) can be shown to reproduce the evolution equation (2.4) in exactly the same way.

To investigate the integrability condition of Eqs. (3.8) and (3.9), we define two 2×2 matrices

$$\hat{\Omega}_1 = \frac{1}{2}\begin{pmatrix} \tau & i\kappa \\ -i\kappa & -\tau \end{pmatrix}, \tag{3.12a}$$

$$\hat{\Omega}_0 = \frac{1}{2}\begin{pmatrix} \omega_0^{23} & i(\omega_0^{12} + i\omega_0^{13}) \\ -i(\omega_0^{12} - i\omega_0^{13}) & -\omega_0^{23} \end{pmatrix}, \tag{3.12b}$$

and rewrite the above equations as

$$i\frac{\partial\Psi}{\partial\xi^1} = \hat{\Omega}_1\Psi, \tag{3.13a}$$

$$i\frac{\partial\Psi}{\partial\xi^0} = \hat{\Omega}_0\Psi, \tag{3.13b}$$

where

$$\Psi = \begin{pmatrix} \phi \\ \phi^{c} \end{pmatrix}. \tag{3.14}$$

The integrability condition of Eqs. (3.13) can be written simply as

$$\left[i\frac{\partial}{\partial\xi^1} - \hat{\Omega}_1,\ i\frac{\partial}{\partial\xi^0} - \hat{\Omega}_0\right] = i\frac{\partial}{\partial\xi^0}\hat{\Omega}_1 - i\frac{\partial}{\partial\xi^1}\hat{\Omega}_0 + [\hat{\Omega}_1,\hat{\Omega}_0] = 0. \tag{3.15}$$

On substituting the explicit form (3.12), we obtain the relation (2.6), which was established for the vector equations (2.2) and (2.4). Thus, we see that the integrability condition of the spinor equations is identical to that of the original vector equations.

IV. THE GAUSS AND THE WEINGARTEN EQUATIONS

We generalize the above consideration to extended objects in the 3 + 1-dimensional Minkowski space, represented by x_μ ($\mu = 1,2,3,4$) with $x_4 = it$. Let us consider an extended object of $m+1$ dimensions in the Minkowski space. The spatial extension of the object is expressed in terms of coordinates ξ^a ($a = 1,...,m$) with $0\leqslant m\leqslant 3$, whereas the temporal extension is represented by ξ^0. A point y_μ on the extended object is then a function of ξ^a, i.e.,

$$y_\mu = y_\mu(\xi^a) \quad (a = 0,1,..,m). \tag{4.1}$$

We define the vectors

$$B_{\mu a} \equiv \frac{\partial y_\mu}{\partial\xi^a}, \tag{4.2}$$

which are tangents of the extended object and satisfy

$$B_{\mu a}B_{\mu b} = \frac{\partial y_\mu}{\partial\xi^a}\frac{\partial y_\mu}{\partial\xi^b} \equiv g_{ab}. \tag{4.3}$$

Assuming ξ^0 to be a timelike variable, we have

$$g_{00} < 0. \tag{4.4}$$

We also consider the normals $N_{\mu P}$ ($P = m+1,...,3$) at the point y_μ satisfying

$$B_{\mu a}N_{\mu P} = 0, \tag{4.5}$$

$$N_{\mu P}N_{\mu Q} = \delta_{PQ}. \tag{4.6}$$

For the tangent vectors, the Gauss equation

$$\frac{\partial B_{\lambda a}}{\partial\xi^b} = \left\{{c \atop ba}\right\}B_{\lambda c} + H_{Pba}N_{\lambda P} \tag{4.7}$$

holds where

$$\left\{{c \atop ba}\right\} \equiv \frac{1}{2}g^{ce}\left\{\frac{\partial g_{ae}}{\partial\xi^b} + \frac{\partial g_{be}}{\partial\xi^a} - \frac{\partial g_{ba}}{\partial\xi^e}\right\} \tag{4.8}$$

is the Christoffel symbol and H_{Pba} is the second fundamental tensor with respect to the normal $N_{\lambda P}$.

On the other hand, the normal $N_{\lambda P}$ satisfies the Weingarten equation

$$\frac{\partial N_{\lambda P}}{\partial\xi^b} = -H_{Pb}{}^{a}B_{\lambda a} + L_{PQb}N_{\lambda Q}, \tag{4.9}$$

where L_{PQb} is skew with respect to P and Q, i.e.,

$$L_{PQb} + L_{QPb} = 0. \tag{4.10}$$

The Gauss equation (4.7) and the Weingarten equation (4.9) are the generalization of the evolution equation when $b = 0$ and of the Frenet–Serret equation when $b \neq 0$. These equations determine the shape of the extended object as long as the coefficients $\left\{{c \atop ba}\right\}$, H_{Pba} and L_{PQb} are given.

Of course, for these equations to be compatible, integrability conditions must be satisfied. They are known as the Gauss, Codazzi, and Ricci equations which we shall not write down here explicitly.

Before we proceed further it is interesting to see some special cases. If we consider a point particle moving in the Minkowski space and take the proper time as ξ^0, the Gauss equation (4.7) can be arranged as

$$\frac{dB_{\lambda 0}}{d\xi^0} = -\kappa_1 N_{\lambda 1}, \tag{4.11}$$

and the Weingarten equation (4.9) as

$$\frac{\partial N_{\lambda 1}}{d\xi^0} = -\kappa_1 B_{\lambda 0} + \kappa_2 N_{\lambda 2}, \tag{4.12a}$$

$$\frac{dN_{\lambda 2}}{d\xi^0} = -\kappa_2 N_{\lambda 1} + \kappa_3 N_{\lambda 3}, \tag{4.12b}$$

$$\frac{dN_{\lambda 3}}{d\xi^0} = -\kappa_3 N_{\lambda 2}. \tag{4.12c}$$

The unit tangent $B_{\lambda 0}$ (timelike) and the three spacelike unit normals $N_{\lambda P}$ ($P = 1,2,3$) satisfy the orthonormalization relation

$$h_\mu^{(\alpha)} h_\mu^{(\beta)} = \delta_{\alpha\beta}, \tag{4.13}$$

where α and β run from 1 to 4, and

$$h_\mu^{(4)} \equiv iB_{\mu 0}, \tag{4.14a}$$

$$h_\mu^{(P)} \equiv N_{\mu P} \quad (P = 1,2,3). \tag{4.14b}$$

Equation (4.11) and (4.12) are nothing but the four-dimensional generalization of the Frenet–Serret formula put forward by Synge.[8] In this case, the problem of integrability does not simply arise, since there is only one independent variable involved. We note that Eqs. (4.11) and (4.12) can be written as

$$\frac{dh_\mu^{(\alpha)}}{d\xi^0} = \omega_0^{\alpha\beta} h_\mu^{(\beta)} \quad (\alpha,\beta = 1,2,3,4), \tag{4.15}$$

with

$$\hat{\omega}_0 \equiv [\omega_0^{\alpha\beta}] = \begin{pmatrix} 0 & \kappa_2 & 0 & i\kappa_1 \\ -\kappa_2 & 0 & \kappa_3 & 0 \\ 0 & -\kappa_3 & 0 & 0 \\ -i\kappa_1 & 0 & 0 & 0 \end{pmatrix}. \tag{4.16}$$

The spinorization of the vector equation of the form given in (4.15) is straightforward. Introducing a four-component spinor ψ, the tetrad $h_\mu^{(\alpha)}$ ($\alpha = 1,2,3,4$) can be constructed as

$$h_\mu^{(1)} = (R_\mu + \bar{R}_\mu)/2(J^2 + J_5^2)^{1/2}, \tag{4.17a}$$

$$h_\mu^{(2)} = (R_\mu - \bar{R}_\mu)/2i(J^2 + J_5^2)^{1/2}, \tag{4.17b}$$

$$h_\mu^{(3)} = J_{5\mu}/(J^2 + J_5^2)^{1/2}, \tag{4.17c}$$

$$h_\mu^{(4)} = ih_\mu^{(0)} = iJ_\mu/(J^2 + J_5^2)^{1/2}, \tag{4.17d}$$

where

$$J = \bar{\psi}\psi, \tag{4.18a}$$

$$J_5 = \bar{\psi} i\gamma_5 \psi, \tag{4.18b}$$

$$J_\mu = \bar{\psi} i\gamma_\mu \psi, \tag{4.18c}$$

$$J_{5\mu} = \bar{\psi} i\gamma_5\gamma_\mu \psi, \tag{4.18d}$$

$$R_\mu = \bar{\psi}^c i\gamma_\mu \psi, \tag{4.18e}$$

$$\bar{R}_\mu = \bar{\psi} i\gamma_\mu \psi^c. \tag{4.18f}$$

The ψ^c is the charge conjugation of ψ defined by

$$\psi^c = C\bar{\psi}^t, \tag{4.19}$$

with

$$C^{-1}\gamma_\mu C = -\gamma_\mu^t, \tag{4.20a}$$

$$C = -C^t. \tag{4.20b}$$

Due to the Fierz constraint conditions, orthogonality and completeness of the four vectors $h_\mu^{(\alpha)}$ hold, as has been discussed previously.[9] Using these relations, the spinor equations

$$i\dot{\psi} = -\tfrac{1}{2}(i\kappa_1 + \kappa_3\gamma_5)\psi^c + \tfrac{1}{2}(\kappa_2 + a\gamma_5)\psi, \tag{4.21a}$$

$$i\dot{\psi}^c = -\tfrac{1}{2}(i\kappa_1 + \kappa_3\gamma_5)\psi - \tfrac{1}{2}(\kappa_2 - a\gamma_5)\psi^c, \tag{4.21b}$$

with an arbitrary real number a, are shown to reproduce the four-dimensional Frenet–Serret equation (4.15) with (4.16). The calculation is almost identical with the nonrelativistic case, hence we shall not repeat it here.

Notice that Eqs. (4.21) agree with those obtained by Gürsey when the real constant a is put equal to zero.[1]

V. SPINORIZATION OF THE GAUSS AND THE WEINGARTEN EQUATIONS

For the purpose of spinorization of the Gauss and the Weingarten equations (4.7) and (4.9), it is convenient to rewrite them first in terms of the orthonormal tetrad $h_\mu^{(\alpha)}$ introduced earlier, and to bring them into the form

$$\frac{\partial h_\mu^{(\alpha)}}{\partial \xi^b} = \omega_b^{\alpha\beta} h_\mu^{(\beta)}, \tag{5.1}$$

with

$$\omega_b^{\alpha\beta} + \omega_b^{\beta\alpha} = 0, \tag{5.2}$$

which ensures that the condition

$$h_\mu^{(\alpha)} h_\mu^{(\beta)} = \delta_{\alpha\beta} \tag{5.3}$$

is always preserved. The actual calculation to bring the Gauss and the Weingarten equations into the form (5.1) is tedious, but can be carried out in a straightforward manner by using the Gram–Schmidt procedure. We shall perform an explicit calculation later in the case of a string moving in the Minkowski space.

The spinorization of the equation of the form (5.1) is quite straightforward. Consider a four-component spinor ψ and its charge conjugation ψ^c, and set up their equations as

$$i\frac{\partial \psi}{\partial \xi^a} = \frac{1}{2}\{\omega_a^{12} + (\omega_a^{34} + \eta_a)\gamma_5\}\psi - \tfrac{1}{2}\{\omega_a^{14} + i\omega_a^{24}) + (\omega_a^{23} + i\omega_a^{31})\gamma_5\}\psi^c \tag{5.4a}$$

and

$$i\frac{\partial \psi^c}{\partial \xi^b} = -\frac{1}{2}\{(\omega_b^{14} - i\omega_b^{24}) + (\omega_b^{23} - i\omega_b^{31})\gamma_5\}\psi - \tfrac{1}{2}\{\omega_b^{12} + (\omega_b^{34} - \eta_b)\gamma_5\}\psi^c. \tag{5.4b}$$

The Pauli conjugate equation of these can be obtained if we recall that the index 4 always carries the imaginary unit

$i \equiv \sqrt{-1}$. The quantity η_b is an arbitrary real number.

To show that Eqs. (5.4) agree with (5.1) with the identification (4.17), we first calculate, using (5.4a) and its conjugate,

$$\frac{\partial J}{\partial \xi^a} = -\eta_a J_5. \tag{5.5a}$$

If we use (5.4b) and its conjugate we obtain

$$\frac{\partial J_5}{\partial \xi^a} = \eta_a J, \tag{5.5b}$$

which, together with (5.5a), gives

$$\frac{\partial}{\partial \xi^a}(J^2 + J_5^2) = 0. \tag{5.6}$$

Again combining (5.4a) and its conjugate, we easily establish

$$\begin{aligned} i\frac{\partial}{\partial \xi^a} J_\mu = {} & \omega_a^{41}(R_\mu + \bar{R}_\mu)/2 \\ & + \omega_a^{42}(R_\mu - \bar{R}_\mu)/2i \\ & + \omega_a^{43} J_{5\mu}. \end{aligned} \tag{5.7a}$$

Similarly,

$$\begin{aligned} i\frac{\partial}{\partial \xi^a} R_\mu = {} & \omega_a^{12} R_\mu - (\omega_a^{14} + i\omega_a^{24}) J_\mu \\ & - (\omega_a^{23} + i\omega_a^{31}) J_{5\mu}, \end{aligned} \tag{5.7b}$$

$$\begin{aligned} i\frac{\partial}{\partial \xi^a} \bar{R}_\mu = {} & -\omega_a^{12} \bar{R}_\mu - (\omega_a^{14} - i\omega_a^{24}) J_\mu \\ & + (\omega_a^{23} - i\omega_a^{31}) J_{5\mu}. \end{aligned} \tag{5.7c}$$

Thus, Eqs. (5.1) can be reproduced out of the spinor equations (5.4).

The integrability condition of the original equation (5.1) can be written as

$$\frac{\partial}{\partial \xi^b}\omega_a^{\alpha\beta} - \frac{\partial}{\partial \xi^a}\omega_b^{\alpha\beta} + \omega_a^{\alpha\epsilon}\omega_b^{\epsilon\beta} - \omega_b^{\alpha\epsilon}\omega_a^{\epsilon\beta} = 0. \tag{5.8}$$

In order to show that the integrability condition of the spinor equation (5.4) is compatible with (5.8), we first define the 2×2 matrices

$$\Omega^a = \frac{1}{2}\begin{pmatrix} \omega_a^{12} + (\omega_a^{34} + \eta_a)\gamma_5 & -(\omega_a^{14} + i\omega_a^{24}) - (\omega_a^{23} + i\omega_a^{31})\gamma_5 \\ -(\omega_a^{14} - i\omega_a^{24}) - (\omega_a^{23} - i\omega_a^{31})\gamma_5 & -\omega_a^{12} - (\omega_a^{34} - \eta_a)\gamma_5 \end{pmatrix}. \tag{5.9}$$

Then, Eqs. (5.4) can be put into the form

$$\left[i\frac{\partial}{\partial \xi^a} - \Omega_a\right]\Psi = 0, \tag{5.10a}$$

where

$$\Psi = \begin{pmatrix} \psi \\ \psi^c \end{pmatrix}. \tag{5.10b}$$

Hence, the integrability condition of Eq. (5.10) reads as

$$i\frac{\partial}{\partial \xi^b}\Omega_a - i\frac{\partial}{\partial \xi^a}\Omega_b + [\Omega_a, \Omega_b] = 0. \tag{5.11}$$

Observe that the matrix (5.9) can be written in the form

$$\Omega_a = A_a^{(0)} I + \sum_i A_a^{(i)} \tau_i, \tag{5.12}$$

where τ_i are 2×2 Pauli matrices and

$$A_a^{(0)} = \eta_a \gamma_5, \tag{5.13a}$$

$$A_a^{(1)} = -\omega_a^{14} - \omega_a^{23}\gamma_5, \tag{5.13b}$$

$$A_a^{(2)} = \omega_a^{24} + \omega_a^{31}\gamma_5, \tag{5.13c}$$

$$A_a^{(3)} = \omega_a^{12} + \omega_a^{34}\gamma_5. \tag{5.13d}$$

The integrability condition (5.11) is then reduced to

$$\frac{\partial A_a^{(0)}}{\partial \xi^b} - \frac{\partial A_b^{(0)}}{\partial \xi^a} = 0, \tag{5.14a}$$

$$\frac{\partial}{\partial \xi^b} A_a^{(i)} - \frac{\partial}{\partial \xi^a} A_b^{(i)} + \epsilon_{ijk} A_a^{(j)} A_b^{(k)} = 0. \tag{5.14b}$$

It is not difficult to see that the condition (5.14b) is identical to Eq. (5.8). The additional condition (5.14a) is, on account of (5.13a),

$$\frac{\partial \eta_a}{\partial \xi^b} - \frac{\partial \eta_b}{\partial \xi^a} = 0, \tag{5.15}$$

which implies that the function η_a is arbitrary but must be of the form

$$\eta_a = \frac{\partial \eta}{\partial \xi^a}. \tag{5.16}$$

Hence, the η terms in the spinor equations (5.4) can be eliminated by the local chiral gauge transformation.

VI. STRING IN MINKOWSKI SPACE

The general scheme to accommodate extended objects into the Minkowski space presented in the previous section will now be worked out explicitly in the case of a string. A point on a world sheet formed by the motion of the string will be denoted by (ξ^0, ξ^1). Let the coordinates of the point with respect to the Minkowski space be y_μ. Then,

$$y_\mu = y_\mu(\xi^0, \xi^1). \tag{6.1}$$

The tangents at the point are defined by

$$B_{\mu a} \equiv \frac{\partial y_\mu}{\partial \xi^a} \quad (a = 0,1), \tag{6.2}$$

which are normalized as

$$B_{\mu a} B_{\mu b} = \frac{\partial y_\mu}{\partial \xi^a}\frac{\partial y_\mu}{\partial \xi^b} \equiv g_{ab} = g_{ba}. \tag{6.3}$$

Since one of the tangents $B_{\mu 0}$, say, is of timelike and $B_{\mu 1}$ of spacelike, we have

$$g_{00} < 0, \tag{6.4a}$$

$$g_{11} > 0. \tag{6.4b}$$

Hence, we have

$$g \equiv g_{10}^2 - g_{00} g_{11} > 0. \tag{6.5}$$

There are two unit normals $N_{\mu p}$ $(p = 2.3)$ at the point, satisfying

$$B_{\mu a}N_{\mu P}=0, \tag{6.6a}$$

$$N_{\mu P}N_{\mu Q}=\delta_{PQ}. \tag{6.6b}$$

The tangents $B_{\mu a}$ $(a=0.1)$ and the normals $N_{\mu P}$ $(P=2.3)$ satisfy, respectively, the Gauss equation (4.7) and the Weingarten equation (4.9).

To describe these equations in terms of the orthonormal tetrad $h_{\mu}^{(\alpha)}$, we take $h_{\mu}^{(0)}$ to be in the direction of $B_{\mu 0}$, and $h_{\mu}^{(1)}$ to be a linear combination of $B_{\mu 0}$ and $B_{\mu 1}$. Thus, we obtain

$$h_{\mu}^{(0)}=B_{\mu 0}/\sqrt{-g_{00}}, \tag{6.7a}$$

$$h_{\mu}^{(1)}=(B_{\mu 1}-(g_{10}/g_{00})B_{\mu 0})\sqrt{-g_{00}/g}, \tag{6.7b}$$

$$h_{\mu}^{(2)}=N_{\mu 2}, \tag{6.7c}$$

$$h_{\mu}^{(3)}=N_{\mu 3}. \tag{6.7d}$$

As is easily checked, Eqs. (6.7) satisfy the orthonormalization condition (5.3) when $B_{\mu a}$ and $N_{\mu P}$ satisfy Eqs. (6.3) and (6.6). Equations (6.7) can be inverted as

$$B_{\mu 0}=h_{\mu}^{(0)}\sqrt{-g_{00}}, \tag{6.8a}$$

$$B_{\mu 1}=\sqrt{-g/g_{00}}h_{\mu}^{(1)}-(g_{10}/\sqrt{-g_{00}})h_{\mu}^{(0)}, \tag{6.8b}$$

$$N_{\mu 2}=h_{\mu}^{(2)}, \tag{6.8c}$$

$$N_{\mu 3}=h_{\mu}^{(3)}. \tag{6.8d}$$

If we differentiate (6.7) and use the Gauss and the Weingarten equations, we obtain, after tedious calculations,

$$\frac{\partial h_{\mu}^{(1)}}{\partial \xi^{a}}=i\frac{\sqrt{g}}{g_{00}}\left\{{1 \atop 0a}\right\}h_{\mu}^{(4)}+\sqrt{-\frac{g}{g_{00}}}\{H_{2}{}^{1}{}_{a}h_{\mu}^{(2)}+H_{3}{}^{1}{}_{a}h_{\mu}^{(3)}\}, \tag{6.9a}$$

$$\frac{\partial h_{\mu}^{(2)}}{\partial \xi^{a}}=-i\frac{g}{\sqrt{-g_{00}}}H_{20a}h_{\mu}^{(4)}-\sqrt{-\frac{g}{g_{00}}}\{H_{2}{}^{1}{}_{a}h_{\mu}^{(1)}+L_{23a}h_{\mu}^{(3)}\}, \tag{6.9b}$$

$$\frac{\partial h_{\mu}^{(3)}}{\partial \xi^{a}}=-i\frac{g}{\sqrt{-g_{00}}}H_{30a}h_{\mu}^{(4)}-\sqrt{-\frac{g}{g_{00}}}H_{3}{}^{1}{}_{a}h_{\mu}^{(1)}-L_{23a}h_{\mu}^{(2)}, \tag{6.9c}$$

$$\frac{\partial h_{\mu}^{(4)}}{\partial \xi^{a}}=-i\frac{\sqrt{g}}{g_{00}}\left\{{1 \atop 0a}\right\}h_{\mu}^{(1)}+i\frac{1}{\sqrt{-g_{00}}}\{H_{20a}h_{\mu}^{(2)}+H_{30a}h_{\mu}^{(3)}\}, \tag{6.9d}$$

where

$$h_{\mu}^{(4)}=ih_{\mu}^{(0)} \tag{6.10}$$

and the relations

$$g^{10}=g^{01}=g_{10}/g, \tag{6.11a}$$

$$g^{11}=-g_{00}/g, \tag{6.11b}$$

$$g^{00}=-g_{11}/g \tag{6.11c}$$

have been used. Hence, we have

$$\hat{\omega}_{a}\equiv[\omega_{a}^{\alpha\beta}]=\begin{pmatrix} 0 & H_{2}{}^{1}{}_{a}/\sqrt{g^{\mathrm{II}}} & H_{3}{}^{1}{}_{a}/\sqrt{g^{\mathrm{II}}} & i\left\{{1 \atop 0a}\right\}\sqrt{g}/g_{00} \\ -H_{2}{}^{1}{}_{a}/\sqrt{g^{\mathrm{II}}} & 0 & L_{23a} & -iH_{20a}/\sqrt{-g_{00}} \\ -H_{3}{}^{1}{}_{a}/\sqrt{g^{\mathrm{II}}} & -L_{23a} & 0 & -iH_{30a}/\sqrt{-g_{00}} \\ -i\left\{{1 \atop 0a}\right\}\sqrt{g}/g_{00} & iH_{20b}/\sqrt{-g_{00}} & iH_{30b}/\sqrt{-g_{00}} & 0 \end{pmatrix}. \tag{6.12}$$

It is instructive if the general treatment above is put into the following form. Let us consider the timelike surface given by

$$y_{\mu}=\xi^{0}e_{\mu}^{(0)}+\xi^{1}e_{\mu}^{(1)}+\phi(\xi^{1},\xi^{0})e_{\mu}^{(2)}, \tag{6.13}$$

with

$$e_{\mu}^{(0)}=(0,0,0,i), \tag{6.14a}$$

$$e_{\mu}^{(1)}=(1,0,0,0), \tag{6.14b}$$

$$e_{\mu}^{(2)}=(0,1,0,0), \tag{6.14c}$$

$$e_{\mu}^{(3)}=(0,0,1,0). \tag{6.14d}$$

Note that Eq. (6.13) is equivalent to

$$y_{0}=\xi^{0}, \tag{6.15a}$$

$$y_{1}=\xi^{1}, \tag{6.15b}$$

$$y_{2}=\phi(\xi^{1},\xi^{0}), \tag{6.15c}$$

namely, a string lying on the x_1 axis oscillating in the x_2 direction with the displacement ϕ. If we apply the general theory to (6.13), we obtain

$$B_{\mu 0}=\frac{\partial y_{\mu}}{\partial \xi^{0}}=e_{\mu}^{(0)}+\dot{\phi}e_{\mu}^{(2)} \tag{6.16a}$$

$$B_{\mu 1}=\frac{\partial y_{\mu}}{\partial \xi^{1}}=e_{\mu}^{(1)}+\phi' e_{\mu}^{(2)}, \tag{6.16b}$$

where

$$\dot{\phi}\equiv\frac{\partial \phi}{\partial \xi^{0}}, \tag{6.17a}$$

$$\phi'\equiv\frac{\partial \phi}{\partial \xi^{1}}. \tag{6.17b}$$

And,

$$g_{00}=\dot{\phi}^{2}-1, \tag{6.18a}$$

$$g_{10}=\dot{\phi}\phi', \tag{6.18b}$$

$$g_{11}=\phi'^{2}+1, \tag{6.18c}$$

$$g=g_{10}^{2}-g_{00}g_{11}=1+\phi'^{2}-\dot{\phi}^{2}. \tag{6.18d}$$

The orthonormal tetrad constructed by (6.7) becomes

$$h_{\mu}^{(\alpha)}=f^{\alpha\beta}e_{\mu}^{(\beta)}, \tag{6.19}$$

with

$$\hat{f}\equiv[f^{\alpha\beta}]=\begin{pmatrix} \sqrt{-g_{00}/g} & \phi'/\sqrt{-g_{00}g} & 0 & -i\dot{\phi}\phi'/\sqrt{-g_{00}g} \\ -\phi'/\sqrt{g} & 1/\sqrt{g} & 0 & -i\dot{\phi}/\sqrt{g} \\ 0 & 0 & 1 & 0 \\ 0 & i\dot{\phi}/\sqrt{-g_{00}} & 0 & 1/\sqrt{-g_{00}} \end{pmatrix}. \tag{6.20}$$

Since Eq. (6.19) is an orthogonal transformation, we have

$$\hat{f}^{t}=\hat{f}^{-1}. \tag{6.21}$$

The Gauss and the Weingarten equations can be obtained from (6.19) in a straightforward manner:

$$\dot{h}_{\mu}^{(\alpha)}=\dot{f}^{\alpha\beta}e_{\mu}^{(B)}=(\dot{\hat{f}}\hat{f}^{-1})^{\alpha\beta}h_{\mu}^{(\beta)}\equiv\omega_0^{\alpha\beta}h_{\mu}^{(\beta)}, \tag{6.22a}$$

$$h_{\mu}^{(\alpha)\prime}=f'^{\alpha\beta}e_{\mu}^{(B)}=(\hat{f}'\hat{f}^{-1})^{\alpha\beta}h_{\mu}^{(\beta)}\equiv\omega_1^{\alpha\beta}h_{\mu}^{(\beta)}. \tag{6.22b}$$

Hence from (6.20) and (6.21), we obtain

$$\hat{\omega}_0\equiv[\omega_0^{\alpha\beta}]$$

$$=\begin{pmatrix} 0 & \dfrac{\dot{\phi}\phi'\ddot{\phi}-g_{00}\dot{\phi}'}{g\sqrt{-g_{00}}} & 0 & i\dfrac{\phi'\ddot{\phi}}{g_{00}\sqrt{g}} \\ -\dfrac{\dot{\phi}\phi'\ddot{\phi}-g_{00}\dot{\phi}'}{g\sqrt{-g_{00}}} & 0 & 0 & -i\dfrac{\ddot{\phi}}{\sqrt{-g_{00}g}} \\ 0 & 0 & 0 & 0 \\ -i\dfrac{\phi'\ddot{\phi}}{g_{00}\sqrt{g}} & i\dfrac{\ddot{\phi}}{\sqrt{-g_{00}g}} & 0 & 0 \end{pmatrix} \tag{6.23a}$$

and

$$\hat{\omega}_1\equiv[\omega_1^{\alpha\beta}]$$

$$=\begin{pmatrix} 0 & \dfrac{\dot{\phi}\phi'\dot{\phi}'-g_{00}\phi''}{g\sqrt{-g_{00}}} & 0 & i\dfrac{\phi'\dot{\phi}'}{g_{00}\sqrt{g}} \\ \dfrac{g_{00}\phi''-\dot{\phi}\phi'\dot{\phi}'}{g\sqrt{-g_{00}}} & 0 & 0 & -i\dfrac{\dot{\phi}'}{\sqrt{-g_{00}g}} \\ 0 & 0 & 0 & 0 \\ -i\dfrac{\phi'\dot{\phi}'}{g_{00}\sqrt{g}} & i\dfrac{\dot{\phi}'}{\sqrt{-g_{00}g}} & 0 & 0 \end{pmatrix}. \tag{6.23b}$$

The integrability condition is automatically satisfied in this case as is easily verified. Namely,

$$\left[\frac{\partial}{\partial\xi^0}-\hat{\omega}_0,\frac{\partial}{\partial\xi^1}-\hat{\omega}_1\right]$$
$$=\hat{\omega}_0'-\dot{\hat{\omega}}_1+[\hat{\omega}_0,\hat{\omega}_1]$$
$$=(\dot{\hat{f}}\hat{f}^{-1})'-(\hat{f}'\hat{f}^{-1})^{\cdot}+\dot{\hat{f}}\hat{f}^{-1}\hat{f}'\hat{f}^{-1}-\hat{f}'\hat{f}^{-1}\dot{\hat{f}}\hat{f}^{-1}=0 \tag{6.24}$$

on account of

$$(f^{-1})'=-f^{-1}f'f^{-1}, \tag{6.25a}$$

$$(f^{-1})^{\cdot}=-f^{-1}\dot{f}f^{-1}. \tag{6.25b}$$

The spinor equations can be written down as in Eq. (5.4) using (6.23).[10]

For the complete determination of the behavior of the tetrad, the quantity $\phi(\xi)$ is yet to be given.

If we take the covariance as a guide,[11] the simplest equation of motion can be obtained from the variational principle

$$\delta\int d\xi^0\,d\xi^1\sqrt{g}=0, \tag{6.26}$$

which leads us to covariant equation

$$\frac{\partial}{\partial\xi^a}(\sqrt{g}\,B_{\mu}^{a})=0 \tag{6.27}$$

with

$$B_{\mu}^{a}=g^{ab}B_{\mu b}\ . \tag{6.28}$$

Substituting (6.16) and (6.18) into (6.27) we arrive at

$$\frac{\partial}{\partial\xi^a}(\sqrt{g}\,B_{\mu}^{a})$$
$$=h_{\mu}^{(2)}g^{ab}\frac{\partial^2}{\partial\xi^a\partial\xi^b}\phi(\xi)$$
$$=h_{\mu}^{(2)}\frac{1}{g}\{-(\phi'^2+1)\ddot{\phi}$$
$$+2\dot{\phi}\phi'\dot{\phi}'-(\dot{\phi}^2-1)\phi''\}=0. \tag{6.29}$$

Note that the variational principle (6.26) implies the element of the surface formed by the string to be optimized. Equation (6.29) is a nonlinear equation so that it is difficult to find a general solution. Some of the solutions will be given in the Appendix.

If we give up covariance, we may impose the variational relation such as

$$\delta\int d\xi^0\, d\xi^1\, \mathcal{L}_0 = 0 \tag{6.30}$$

with

$$\mathcal{L}^0 \equiv -\tfrac{1}{2}g \tag{6.31}$$

which gives us the simple wave equation

$$\ddot{\phi} - \phi'' = 0. \tag{6.32}$$

The Liouville field theory can be obtained, again giving up the covariance, if we take, instead of (6.31), the Lagrangian

$$\mathcal{L} = \mathcal{L}_0 - \mu^2 e^{\phi}. \tag{6.33}$$

VII. DISCUSSION

The mathematical scheme given in the text of this paper provides us with a method to deal with a motion of spatially extended physical objects. The Gauss and the Weingarten equations describe the "internal" motion of the objects in terms of the curvature, the torsion and the like. These quantities are treated as given in this paper, but they are something which are to be determined by dynamics of the total system just as in Einstein's gravitational theory, the curvature of the space is determined by the distribution of matter.

The problem of the interaction between physical objects has been left out so far in our formalism. This is because it is difficult to set up at this stage any realistic coupling between the external and internal degrees of freedom, and also the entire formalism is still too primitive to take into account the successful feature of the group theoretical approach in the local field theory. For the latter aspect, the spinorization of the fundamental equation will no doubt play an important role, whereas the former must be carefully investigated in connection with the quantization. In particular, the restriction on the relation between spin and statistics can no longer be taken for granted because the basic postulates to prove the relation are by no means obvious for the internal degree of freedom.

Apart from the fundamental problems mentioned above, there are some practical advantages in geometrization and spinorization of extended objects. As can be observed readily, the spinor version of the fundamental equations is less stringent, since the orthonormalization condition of the tetrad is automatically satisfied. Thus, Inoue and Omote obtained exact solutions of the nonlinear σ-model by transforming the original equation into a spinor form.[12]

The geometrization of one-dimensionally extended objects has been exploited by a number of authors in order to treat soliton solutions and to discover new soliton equations.[4] The method presented in this paper can easily be utilized for the same purpose to treat objects with the higher-dimensional extension.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Science and Engineering Research Council of Canada. We are indebted to Werner Israel and Miki Wadati for valuable suggestions and discussions.

APPENDIX

Equation

$$-(\phi'^2+1)\ddot{\phi} + 2\dot{\phi}\phi'\dot{\phi}' - (\dot{\phi}^2-1)\phi'' = 0. \tag{A1}$$

We have not succeeded in finding all possible solutions, but some of the solutions are as follows:

(i) $$\phi = f(\xi^1 - \xi^0), \tag{A2}$$

where f is an arbitrary function. To verify (A2) is a solution of (A1), we note that for (A2), we have

$$\dot{\phi} = -\phi'. \tag{A3}$$

Substituting (A3) into the left-hand side of (A1), we find that it vanishes identically.

(ii) $$\phi = g(\xi^1 + \xi^0). \tag{A4}$$

Similarly, (A4) is also a solution where g is arbitrary.

(iii) $$\phi = \frac{1}{\lambda}\log\left|\frac{\lambda(\xi^1+\xi^0)+\alpha}{\lambda(\xi^1-\xi^0)+\beta}\right| + \text{const.} \tag{A5}$$

It is interesting to see that although (A2) and (A4) are solutions individually, the sum of them is not in general a solution. If we insist that (A2) and (A4) coexist, the form of f and g is restricted to (A5), where λ, α, and β are arbitrary constants.

(iv) $$\phi = \frac{1}{\lambda}\log\{\sin(\lambda\xi^1+\alpha)\sinh^2(\lambda\xi^0+\beta) + \sqrt{\sin^2(\lambda\xi^1+\alpha)\sinh^2(\lambda\xi^0+\beta)+1}\} + \text{const.} \tag{A6}$$

As can be verified directly (A6) satisfies two equations

$$-(\phi'^2+1)\ddot{\phi} + \tfrac{1}{2}(\phi'^2+1)\dot{\phi} = 0, \tag{A7}$$

$$(1-\dot{\phi}^2)\phi'' - \tfrac{1}{2}(1-\dot{\phi}^2)'\phi' = 0, \tag{A8}$$

the sum of which gives Eq. (A1).[13] The solution (A6) grows with ξ^0 as ξ^0 becomes large.

[1] F. Gürsey, Nuovo Cimento **5**, 784 (1957).

[2] T. Hara and T. Goto, Prog. Theor. Phys. **44**, 1383 (1970); see also T. Takabayasi, Nuovo Cimento **7**, 118 (1958).

[3] Y. Takahashi, Phys. Rev. D **26**, 2169 (1982); Prog. Theor. Phys. **69**, 369 (1983); J. Math. Phys. **24**, 1783 (1983).

[4] H. Hasimoto, J. Fluid Mech. **51**, 477 (1972); M. Lakshmanan, Th. W. Ruijgrok, and C. J. Thompson, Physica **84A**, 577 (1976); G. L. Lamb, Jr., Phys. Rev. Lett. **37**, 235 (1976); see also M. Lakshmanan and K. M. Tamizhmani, Appl. Sci. Res. **37**, 127 (1980); M. Lakshmanan, J. Math. Phys. **20**, 1667 (1979).

[5] M. Kalb and P. Ramond, Phys. Rev. D **9**, 2273 (1973); Y. Nambu, *Proceedings of the 1981 INS Symposium on Quark and Lepton Physics*, Tokyo, June 25–27, 1981 (Institute for Nuclear Study, Univ. of Tokyo, Tokyo, Japan, 1981). T. Goto, Prog. Theor. Phys. **46**, 1560 (1971); *ibid.* **47**, 2090 (1972); T. Takabayasi, *ibid.* **49**, 1724 (1973); H. Matsumoto and H. Umezawa, Phys. Rev. D **14**, 3536 (1976). See also H. Umezawa, H. Matsumoto, and M. Tachiki, *Thermo Field Dynamics and Condensed States* (North-Holland, Amsterdam, 1982), and references therein.

[6] The prime indicates differentiation with respect to ξ^1, whereas the dot implies that with respect to ξ^0.

[7] Y. Takahashi, Prog. Theor. Phys. **69**, 369 (1983).

[8] J. L. Synge, Proc. London Math. Soc. **43**, 376 (1937).

[9] Y. Takahashi, J. Math. Phys. **24**, 1783 (1983).

[10] It is interesting to calculate the Ricci tensor in the above situation. Using the relations (6.18), we obtain the Ricci tensor in a compact form $R_{ab} = (1/g^2)g_{ab}(\phi''\ddot{\phi} - \dot{\phi}'^2)$ $(a,b = 0,1)$.

[11] T. Takabayasi, Prog. Theor. Phys. **49**, 1724 (1973).

[12] K. Inoue and M. Omote, "Spinorization and Exact Solutions of the Discrete Non-linear σ-model," Tsukuba University preprint, 1983.

[13] The solution (A6) is due to N. Papastamatiou. We thank him for this.

Twisted boundary conditions for gauge theories on a torus

Alfred Actor
Department of Physics, Pennsylvania State University, Fogelsville, Pennsylvania 18051

(Received 30 November 1983; accepted for publication 23 March 1984)

Gauge theories on the torus T^2 and higher-dimensional tori are considered. The compactness of these manifolds induces boundary conditions on path integrals, one for each compactified dimension. Gauge fixing affects these boundary conditions, causing them to be nonperiodic in certain gauges, the nonperiodicity being expressed in terms of a gauge transformation or "twist." The eigenvalues of the twist functions can be interpreted as spin variables in an effective spin theory, as is done in temperature field theory. The spin theory action, the effective potential, center symmetry, and the high-temperature limit are discussed. We clarify the connection between this spin theory and t'Hooft's formalism for gauge theories on a torus.

PACS numbers: 11.15 – q, 11.30. – j

I. INTRODUCTION

Gauge theories at finite temperature, as well as theories of the original Kaluza-Klein type, are defined on a cylindrical space-time manifold $M = E^n \times S^1$. The compactness of S^1 is expressed through a periodic boundary condition on the path integrals of the (unfixed) gauge theory. This boundary condition leads to an interesting analogy between the gauge theory, and an effective static "spin" theory defined on E^n. Gauge fixing (for certain gauges) causes this boundary condition to become nonperiodic or "twisted," the twist function $\Omega(\mathbf{x})$ being a static gauge transformation. Alternatively, the eigenvalues of $\Omega(\mathbf{x})$ can be interpreted as the degrees of freedom of a scalar theory defined in terms of the boundary condition—the spin theory. Thermal effects in the gauge theory are closely related to the dynamics of the simpler spin theory. Through the application of universality arguments, which should reveal the critical behavior of the latter, one hopes to gain a better understanding of the original gauge theory.[1]

Our aim in the present article is to extend the gauge theory–spin theory analogy to tori, i.e., to manifolds $E^n \times T^m$. The main features of this generalization are fairly well illustrated by the torus T^2. Moreover, the Euclidean space E^n plays a somewhat passive role. Therefore, we shall suppress it and work mainly with the manifold $M = T^2$.

Initially, it will be unnecessary to specify the gauge theory under consideration. The Euclidean classical action of the theory is denoted by

$$S = S[f], \quad f = A_\mu, \phi, \psi, \ldots .$$

It depends on the gauge potential A_μ and on boson and fermion matter fields ϕ, ψ, respectively.

In Sec. II, we discuss gauge fixing in the general Faddeev–Popov language. This leads (Sec. III) to a general formula for the effective spin theory action in terms of twisted boundary conditions. Section IV is devoted to the definition of center symmetry, and to the breaking of this classical symmetry at high temperature. A discussion of the temporal gauge follows in Sec. V. The connection between 't Hooft's formalism[2] for gauge theories on a torus, and the spin theory formalism used here, is given. We define and investigate the effective potential $V_{\text{eff}}(A_\mu)$ which provides an alternative to the spin theory approach. Some remarks concerning higher dimensions are made in Sec. VI.

Some previous studies of gauge theories on a torus are the following. (i) There is an investigation by 't Hooft[2] in which pure SU(N) gauge theories are defined in a periodic four-dimensional box ($= T^4$). This takes care of infrared regularization. The free energy F of the system with specified electric and magnetic flux lines is expressed in terms of certain path integrals with twisted boundary conditions. F is shown to possess an important electric–magnetic symmetry. It is possible to demonstrate that simultaneous electric and magnetic confinement cannot occur for gauge group SU(3). Indeed, only three phases are possible: an electric confining phase, a (magnetic confining) Higgs phase, and a so-called Georgi–Glashow phase with zero mass gap. The 't Hooft formalism has been extended to supersymmetric SU(N) gauge theories.[3] (ii) Next, there is a series of papers by Richard and Rouet[4,5] which analyze the instanton contribution to the CP^1 model on T^2. (iii) Finally, there is an article[6] showing that the CP^{N-1} models on E^2 and T^2 are related by a conformal transformation.

II. GAUGE FIXING

Before gauge fixing, the generating function for a Euclidean gauge theory on the torus T^2 is

$$Z = \int [df] e^{-S[f]} \tag{1}$$

$$f(\beta, x_1) = \eta f(0, x_1),$$

$$f(x_0, a) = f(x_0, 0),$$

where f represents the fields of the theory and $\eta = +1 (-1)$ for boson (fermion) fields. Boson (fermion) fields are periodic (antiperiodic) in Euclidean time x_0 with period $\beta = T^{-1}$, where T is the temperature. Here all fields are chosen periodic in space, but antiperiodic spatial boundary conditions are also possible.

To fix the gauge we use the Faddeev–Popov formula[7]

$$\Delta[A_\mu] \int \prod_x dU(x) \delta[F(A_\mu^U)] = 1. \tag{2}$$

Here $A_\mu^U = U A_\mu U^{-1} + iU \partial_\mu U^{-1}$ is the gauge-trans-

 0022-2488/84/092736-05$02.50

formed potential obtained from A_μ by the local gauge transformation $U(x)$. Let f^U denote the corresponding gauge transformation of f.

Inserting Eq. (2) into Eq. (1) one finds

$$Z=\int\prod_x dU(x)\int[df]\Delta\,[A_\mu]\delta[F(A_\mu^U)]e^{-S[f]}, \tag{3a}$$

$$f(\beta,x_1)=\eta f(0,x_1),$$

$$f(x_0,a)=f(x_0,0)$$

$$=\int\prod_x dU(x)\int[df^U]\Delta\,[A_\mu^U]\delta[F(A_\mu^U)]e^{-S[f^U]} \tag{3b}$$

$$[f^U]^{U^{-1}}(\beta,x_1)=\eta[f^U]^{U^{-1}}(0,x_1),$$

$$[f^U]^{U^{-1}}(x_0,a)=[f^U]^{U^{-1}}(x_0,0)$$

$$=\int\prod_x dU(x)\int[df]\Delta\,[A_\mu]\delta[F(A_\mu)]e^{-S[f]} \tag{3c}$$

$$f^{U^{-1}}(\beta,x_1)=\eta f^{U^{-1}}(0,x_1),$$

$$f^{U^{-1}}(x_0,a)=f^{U^{-1}}(x_0,0).$$

Here we have used the gauge invariance of $\Delta[A_\mu]$, $S[f]$, and $[df]$ to obtain the second formula, followed by a change of variables $f^U\to f$ to reach the third. These manipulations parallel similar ones in the temporal gauge analysis of $T=0$ gauge theories[8] and $T>0$ gauge theories.[9] The gauge has now been fixed.

III. ANALOG SPIN THEORY

The last equality in Eq. (3) is almost the desired formula for Z. It remains to eliminate the irrelevant integrations in $\Pi_x\,dU(x)$. The integrand for these integrations is a function of the twist variables

$$\Omega_0(x_1)\equiv U^{-1}(\beta,x_1)U(0,x_1),$$

$$\Omega_1(x_0)\equiv U^{-1}(x_0,a)U(x_0,0). \tag{4}$$

Let us give this integrand the name

$$e^{-S_{\rm eff}[\Omega_\mu]}\equiv\int[df]\Delta\,[A_\mu]\delta[F(A_\mu)]e^{-S} \tag{5}$$

$$f(\beta,x_1)=\eta f^{\Omega_0^{-1}}(0,x_1),$$

$$f(x_0,a)=f^{\Omega_1^{-1}}(x_0,0).$$

One then simply discards the integration over $dU(x)$ at points not on the boundary (β,x_1), $(0,x_1)$, (x_0,a), $(x_0,0)$ of the space-time rectangle. Elementary manipulations lead to the final expression

$$Z=\int\prod_{x_1}d\Omega_0(x_1)\prod_{x_0}d\Omega(x_0)e^{-S_{\rm eff}[\Omega_\mu]}. \tag{6}$$

Observe that Z has been expressed in terms of an effective action $S_{\rm eff}[\Omega_\mu]$ whose variables are the twists Ω_μ. The following points deserve attention.

(1) $S_{\rm eff}$ is not an ordinary field theory, in that $\Omega_0(x_1)$ and $\Omega_1(x_0)$ are defined on different manifolds. Of course, a physical system with two sets of variables, depending on position and time, respectively, can be imagined.

(2) Gauge fixing has twisted the boundary conditions of the gauge theory. Naturally, the twist variables are gauge-dependent. There exist gauges in which one or both of these variables is unity, modulo a constant element $z\in C$ of the center C of the gauge group G. (C consists of all elements $z\in G$ which commute with every element $g\in G$.)

(3) The functions Ω_μ are not entirely independent. Self consistency at the corners of the space-time rectangle leads to a constraint[2]

$$\Omega_0(a)\Omega_1(0)=\Omega_1(\beta)\Omega_0(0)z \tag{7}$$

where $z\in C$ is an element of the center.

(4) One must realize that compactification of a particular space-time direction x_α imbues the corresponding component A_α of the gauge potential with more dynamical importance than it possesses in flat space. For example, these particular components A_α cannot be eliminated by gauge transformations from the theory. The twists Ω_α are certain functions of these variables, which differ from gauge to gauge. In gauges with unit twists, one can always define an effective action $S_{\rm eff}[A_\alpha]$ directly in terms of A_α. This action has the same dynamical content as $S_{\rm eff}[\Omega_\alpha]$ in a gauge with variable twists.

IV. CENTER SYMMETRY

Pure non-abelian gauge theories possess a useful classical symmetry under global gauge transformations $U\in C$ belonging to the center C of the gauge group G. Both the gauge potential A_μ and a scalar field in the adjoint representation of G are invariant under the center, i.e., $f^U=f$ for $U\in C$. Thus, for Yang–Mills theories with or without ajoint scalars, the boundary condition in Eq. (5) is explicitly invariant under the replacement $\Omega_\mu\to z_\mu\Omega_\mu$, $z_\mu\in C$. Therefore,

$$S_{\rm eff}[z_\mu\Omega_\mu]=S_{\rm eff}[\Omega_\mu],\quad z_\mu\in C. \tag{8}$$

This is sometimes referred to as center symmetry.[1]

Matter fields in the fundamental representation clash with center symmetry. For such fields the transformation rule is $f^z=zf\neq f$, and the effective action does not have the classical symmetry (8).

Center symmetry provides a useful guide[1] to the dynamics of a given theory. If this symmetry is valid classically, but broken quantum mechanically, then we have a case of dynamical symmetry breakdown.

At high temperature, quantum effects do not tolerate center symmetry. Thus, if the low T theory is sufficiently well understood, one can say whether or not dynamical symmetry breakdown occurs as a function of temperature. A similar comment applies to the parameter a.

Consider the high temperature limit $\beta=T^{-1}\to 0$, where the gauge theory becomes static. In any gauge, $\beta\to 0$ implies $\Omega_0\to z\in C$. This follows the first boundary condition in Eq. (5) and the requirement that all fields $f(0,x_1)$ be defined at $\beta=0$. Consequently, Ω_0 ceases to be a dynamical variable—or, the spin theory exhibits perfect order in Ω_0 at $T=\infty$. We emphasize that Ω_0 approaches a particular (though arbitrary) element $z\in C$. Thus center symmetry will be broken at very high temperature in any gauge theory.

A pure Yang–Mills theory is center symmetric at low T, but not at high T. Hence there occurs dynamical symmetry breakdown.[10] As the broken symmetry is discrete, no

Goldstone bosons are induced, and the high-T phase is implied to be massive for any compact semisimple gauge group. This is true for any dimensionality of configuration space: center symmetry is not tied to any particular dimension. A well-known example is four-dimensional QCD which is massive at finite temperature.[11] In two dimensions, Coleman's theorem,[12] which forbids the breakdown of a continuous symmetry, is not activated as long as C is discrete.

Things are different for abelian gauge groups G, for then the center coincides with the group. If G is continuous, then symmetry breakdown implies the presence of Goldstone particles, and a Coulomb phase for the theory. Actually, there will be no symmetry breakdown in an abelian theory with matter fields, for the simple reason that the latter are always in the fundamental representation. Hence, there is no classical symmetry to be broken, and no reason to expect a massless phase.

A pure abelian gauge theory does have center symmetry, and symmetry breakdown may occur in more than two dimensions. The possibilities for lattice U(1) theories are compiled in Ref. 1.

V. TEMPORAL GAUGE

The temporal gauge $A_0^U = 0$ and the axial gauge $A_1^U = 0$ play symmetric roles on T^2. With obvious changes, the discussion to follow concerning the temporal gauge applies to the axial gauge as well.

A. Analog spin theory

The gauge transformation to the temporal gauge $A_0^U = 0$ from any periodic gauge is

$$U^{-1}(x_0,x_1) = \left\{P \exp i\int_0^{x_0} dx_0\, A_0(x_0,x_1)\right\} U^{-1}(0,x_1), \quad (9)$$

where P denotes path ordering. Observe the factor $U(0,x_1)$ here, which is an arbitrary static gauge transformation expressing the residual gauge freedom in the temporal gauge. We choose $U(0,x_1) = 1$, thereby eliminating this remaining freedom. Then, $U(x_0,a) = U(x_0,0)$ and $\Omega_1 = 1$. The dynamical variables in this gauge are A_1 and the twist

$$\Omega_0(x_1) = P \exp i\int_0^{\beta} dx_0 A_0. \quad (10)$$

An effective action, equivalent to the one in Eq. (5), is thus defined by

$$Z = \int \prod_{x_1} d\Omega_0(x_1) \prod_x dA_1(x) \exp(-S_{\text{eff}}[\Omega_0,A_1]), \quad (11)$$

$$\exp(-S_{\text{eff}}[\Omega_0,A_1]) \equiv \int [df'] e^{-S[A_0=0]}, \quad (12)$$

$$f(\beta,x_1) = \eta f^{\Omega_0^{-1}}(0,x_1),$$

$$f(x_0,a) = f(x_0,0),$$

where $[df']$ is $[df]$ without the factor $\Pi_x dA_\mu(x)$. The temporal gauge is, of course, ghost-free ($\Delta[A_\mu] = 1$). Note that Ω_0 must satisfy the constraint (7), which here reads $\Omega_0(a) = z\Omega_0(0)$.

B. Electric flux on the torus

Next, we would like to relate the effective action S_{eff} to the temporal gauge formalism developed by 't Hooft for the representation of electric and magnetic flux lines in SU(N) gauge theories on a torus. The four-dimensional formalism in Ref. 2 can easily be adapted to other dimensions, and in particular to the present case where only electric flux exists. The twisted functional integrals $W[\mathbf{k},\mathbf{m};\mathbf{a},\beta]$ in Ref. 2 correspond here to the functionals (12) in the following way:

$$W[k;a,\beta] = \exp(-s_{\text{eff}}[\Omega_{0k},A_1]). \quad (13)$$

Note that the integer k is a topological index, labeling different topological sectors defined by

$$\Omega_{0k}(a) = e^{2\pi ik/N}\Omega_{0k}(0). \quad (14)$$

Thus Ω_{0k} is a "twisted twist," a notion that has arisen through the compactification of a second dimension.

Following Ref. 2, the free energy $F(e)$ associated with e flux lines along the x_1 axis is given by

$$e^{-\beta F(e)} = \frac{1}{N}\sum_k e^{-2\pi ike/N} W[k;a,\beta]. \quad (15)$$

Appropriately, $\exp(-\beta F(0))$ is the average of the $W[k;a,\beta]$, while $F(e\neq 0)$ is obtained from a certain combination of these functionals.

The low temperature gauge theory confines electric charge due to the one-dimensional nature of configuration space. For *nonintegral* charge, confinement will persist to any finite T for the same reason. But *integral* charge will be screened at high T, by dynamical charges. (Generally gauge theories become plasmalike for $T\to\infty$. See Refs. 10, 11, and 13 for examples.) Therefore, the flux number e should cease to be defined as $\beta\to 0$. It is interesting to see how this happens.

As $T\to\infty$, Ω_{0k} becomes a constant element of the center Z_N. Together with Eq. (14), this implies that k can only take the value $k = 0$ at $T = \infty$. Nonzero values of k are totally suppressed—the effective action $S_{\text{eff}}[\Omega_{0k},A_1]$ is infinite for $k\neq 0$. The right-hand side of Eq. (15) therefore becomes independent of e, and flux number no longer has a meaning.

This suppression of topological sectors with nonzero k comes about as follows. Consider Eq. (10) for an abelian theory, or for a small non-abelian potential A_0, when it has the form

$$\Omega_{0k}(x_1) = e^{i\beta A_{0s}(x_1)},$$

where $A_{0s}(x_1)$ is the static part of A_0, and is here constrained to satisfy

$$A_{0s}(a) = A_{0s}(0) + 2\pi k/N.$$

This equation defines the k th topological sector for abelian, or for small, A_0. Obviously, nonzero k implies that A_{0s} is not constant. Constant potentials can contribute only to the $k = 0$ sector. Now the essential point is, that A_0 becomes constant as $T\to\infty$, because its quantum mechanical mass goes to infinity (Ref. 9). Thus the $k = 0$ sector is selected as $T\to\infty$, and all other sectors are discarded. Spatial fluctuations cost infinite energy at infinite temperature. This argument can be generalized using Eq. (10), but the basic idea remains the same.

In higher dimensions exactly the same thing happens. One has the twisted functions $W[\mathbf{k},\mathbf{m};\mathbf{a},\beta]$, where the topological integers k_i and m_i label twists in equations like (14) for the planes $(0i)$ and (jk), respectively. The number of magnetic flux lines piercing the (jk) plane is m_i. For electric flux there is an equation analogous to (15) defining integers e_i which specify the number of electric flux lines piercing the $(0i)$ plane. As $T\to\infty$ the integers k_i all must vanish, and electric flux is no longer defined. This is precisely what one expects from the picture[13] of deconfinement as a condensation of electric flux lines filling the vacuum. Note that magnetic flux is not affected.

C. Effective potential

As a further illustration of Eq. (2), let us consider the effective potential $V_{\text{eff}} = S_{\text{eff}}$ for a constant potential A_μ, choosing for simplicity an abelian theory with one-loop action

$$S = \int_{T^2} d^2x \left[\frac{1}{2} F_{01}^2 + \frac{1}{2} \bar{\phi} (- D^2 + M^2) \phi \right]. \tag{16}$$

From Eq. (10) we see that $\Omega_0 = \exp i\beta A_0$, and Eq. (12) becomes

$$e^{-V_{\text{eff}}} = \int [d\bar{\phi}\, d\phi\,] \exp\left(- \frac{1}{2} \int d^2x\, \bar{\phi} (- D^2 + M^2) \phi \right),$$

$$\phi(\beta, x_1) = e^{-i\beta A_0} \phi(0, x_1)$$

$$\phi(x_0, a) = \phi(x_0, 0)$$

$$= [\det_{\Omega_0,1} (- D^2 + M^2)]^{-1}. \tag{17}$$

Here $D_0 = \partial_0$ (because S was evaluated at $A_0 = 0$) and $D_1 = \partial_1 - iA_1$. The subscripts on the determinant refer to the space of functions ϕ on which this determinant is evaluated. These functions satisfy the boundary conditions in Eq. (17), and have the Fourier expansion

$$\phi(x) = \sum_{n,m} \exp\left[i(\omega_n - A_0) x_0 + i k_m x_1 \right] \phi_{nm}, \tag{18}$$

$$\omega_n = 2\pi n/\beta, \quad k_m = 2\pi m/a.$$

The determinant of interest can therefore be written

$$\det_{\Omega_0,1} (- D^2 + M^2)$$

$$= \prod_{n,m} [(\omega_n - A_0)^2 + (k_m - A_1)^2 + M^2]. \tag{19}$$

There are several comments to make concerning this determinant.

(1) It requires normalization. The procedure[14] for one compact dimension cannot be used here. To the author's knowledge the calculation of the determinant (19) remains an open mathematical problem. Until this problem is solved, one cannot give an explicit formula for $V_{\text{eff}}(A_0, A_1)$.

(2) V_{eff} is gauge independent, of course. In the Appendix we sketch the calculation of V_{eff} in the Lorentz gauge.

(3) Fortunately it is possible to understand the structure of V_{eff} without normalization. This potential is a function of A_0^2 and A_1^2. It is periodic in A_0 with period $2\pi/\beta$, and in A_1 with period $2\pi/a$. The minima of V_{eff} are at $A_0 = \omega_n$, $A_1 = k_m$.

(4) Quantum fluctuations in the scalar field lead to the effective potential which breaks center symmetry, i.e., invariance under the replacement $A_\mu \to A_\mu + \lambda_\mu$ for constant λ_μ. In Sec. IV, the functional boundary condition for the theory (16) led us to the same conclusion.

(5) Effective potentials in non-abelian gauge theories for constant matrix potential $A_\mu = \text{diag}(A_{\mu a})$ can be expressed as a sum of abelian effective potentials.[11] Therefore, the elementary example just given can be extended to much more complicated situations.

VI. CONCLUSION

In this article we have described some of the new features encountered in gauge theories when E^2 is compactified to T^2. The extension to tori of higher dimension proceeds in an obvious fashion. For each compactified coordinate $0 \leqslant x_\mu \leqslant a_\mu$ there appears a boundary condition on the component A_μ of the gauge potential. Gauge fixing may or may not twist this boundary condition, depending on the gauge. The effective spin theory will depend on each twist Ω_μ (or on A_μ if $\Omega_\mu = 1$ for this gauge). Consequently, the effective potential depends on A_μ for each compactified coordinate.

When all $d+1$ dimensions have been compactified, the effective potential can be obtained from the determinant

$$\det(- D^2 + M^2)$$

$$= \prod_{n_0,\ldots,n_d} [(\omega_{n_0} - A_0) + \cdots + (\omega_{n_d} - A_d)^2 + M^2], \tag{20}$$

where $\omega_{n_\mu} = 2\pi n_\mu / a_\mu$. For $M = 0$ the logarithm of this determinant is an Epstein zeta function.[15] In a related context,[16] some properties of these functions are given for the case $A_\mu = 0$. Moreover, the extension to $M \neq 0$, $A_\mu = 0$ is discussed in Ref. 16. No study of the general case $M \neq 0$, $A_\mu \neq 0$ is known to the author.

The effective potential is periodic in A_μ and has its minima at $A_\mu = \omega_{n_\mu}$ for all sets of integers $\{n_\mu\}$. When the size of the torus becomes very small in a certain direction, say $a_\beta \to 0$, the corresponding component of A_β approaches one of the values $A_\beta \to \omega_{n_\beta}$. Simultaneously, in the effective spin theory, the twist $\Omega_\beta \to z_\beta \in C$ approaches a constant element of the center.

APPENDIX: LORENTZ GAUGE

The Lorentz gauge $\partial_\mu A_\mu = 0$ is doubly periodic, and therefore $\Omega_\mu = 1$. To see this, consider the infinitesimal gauge transformation $U = 1 + iu$ from an infinitesimal periodic potential A_μ to a neighboring infinitesimal potential $A_\mu^U = A_\mu + \partial_\mu u + i[u, A_\mu]$ which satisfies $\partial_\mu A_\mu^U = 0$. One finds $\Box u = -\partial_\mu A_\mu$, implying that $u(x)$ is doubly periodic.

The dynamical variables in this twistless gauge are A_μ, and one defines the effective action by

$$Z = \int \prod_x dA_\mu(x) e^{-S_{\text{eff}}[A_\mu]},$$

$$e^{-S_{\text{eff}}[A_\mu]} \equiv \int_{\text{periodic}} [df'\, d\bar{c}\, dc] \delta[\partial_\mu A_\mu] e^{-S'},$$

$$S' = S + \int_{T^2} d^2x\, \bar{c} (- \Box + i[A_\mu, \partial_\mu]) c.$$

Here the ghost field term in S' replaces the factor $\Delta[A_\mu]$ in the integrand.

$S_{\rm eff}[A_\mu]$ must contain the same information as $S_{\rm eff}[\Omega_0, A_1]$ in the temporal gauge. Let us demonstrate this using the abelian theory (16) as an example. The ghosts decouple, and for constant A_μ the effective potential is defined by Eq. (17) with strictly periodic boundary conditions,

$$V_{\rm eff} = \ln \det(-D^2 + M^2), \quad D_\mu = \partial_\mu - iA_\mu.$$

This determinant coincides precisely with Eq. (19), demonstrating the gauge invariance of the result.

[1]B. Svetitsky and L. G. Yaffe, Nucl. Phys. B **210** (FS6), 423 (1982).
[2]G. 't Hooft, Nucl. Phys. B **153**, 141 (1979); Acta Phys. Austriaca Suppl. **XXII** (1980).
[3]S. Cecotti and L. Girardello, Nucl. Phys. B **208**, 265 (1982).
[4]J. L. Richard and A. Rouet, Nucl. Phys. B **211**, 447 (1983).
[5]J. L. Richard, Phys. Lett. B **115**, 134 (1982); **128**, 213 (1983).
[6]A. Actor, Lett. Math. Phys. **7**, 9 (1983).
[7]L. D. Faddeev and V. N. Popov, Phys. Lett. B **25**, 29 (1967).
[8]G. C. Rossi and M. Testa, Nucl. Phys. B **163**, 109 (1980).
[9]A. Actor, "Compactification at Finite Temperature in Noncompact Electrodynamics," Ann. Phys. (to appear).
[10]L. D. McLerran and B. Svetitsky, Phys. Rev. D **24**, 450 (1981).
[11]D. J. Gross, R. D. Pisarski, and L. G. Yaffe, Rev. Mod. Phys. **53**, 43 (1981).
[12]S. Coleman, Commun. Math. Phys. **31**, 259 (1973).
[13]T. Banks and E. Rabinovici, Nucl. Phys. B **160**, 349 (1979).
[14]A. Actor, Phys. Rev. D **27**, 2548 (1983).
[15]A. Erdelyi *et al.*, *Higher Transcendental Functions*, Bateman Manuscript Project (McGraw-Hill, New York, 1955), Sec. 17.9.
[16]J. Ambjorn and S. Wolfram, Ann. Phys. **147**, 1 (1983).

Analyticity properties and a convergent expansion for the glueball mass and dispersion curve of strongly coupled Euclidean 2 + 1 lattice gauge theories

Michael O'Carroll and Gastão de Almeida Braga
Departamento de Física do ICEx, Universidade Federal de Minas Gerais, Belo Horizonte, 30000, MG, Brazil

(Received 30 November 1983; accepted for publication 3 February 1984)

For certain strongly coupled ($2/g^2 \equiv \beta > 0$ and small) Euclidean Z^3 lattice gauge theories we show that the glueball mass $m(\beta)$ associated with the truncated plaquette–plaquette correlation function admits the representation $m(\beta) = -4\ln\beta + r(\beta)$. $r(\beta) = \Sigma_{n=0}^{\infty} b_n \beta^n$ is a gauge group representation dependent function, analytic at $\beta = 0$. A finite algorithm is given for determining b_n. b_n depends on a finite number of the $\beta = 0$ Taylor series coefficients of the finite lattice correlation function at a finite number of points, increasing with n, of Z^3.

PACS numbers: 11.15.Ha

I. INTRODUCTION

Recently[1] an analysis of the spectrum of the lattice quantum field theory associated with certain strongly coupled Euclidean Z^3 lattice gauge theories with Wilson action[2] was begun. Specifically the truncated plaquette–plaquette correlation function

$$G(x,\beta) \equiv G(y;z,\beta) = \lim_{\Lambda \uparrow Z^3} G_\Lambda(y;z,\beta),$$
$$x = y - z, \quad x,y,z \in \Lambda \subset Z^3$$

was considered. Here

$$G_\Lambda(y;z,\beta) = \langle \chi(g_{p_y}\chi(g_{p_z})\rangle_\Lambda - \langle \chi(g_{p_y})\rangle_\Lambda \langle \chi(g_{p_z})\rangle_\Lambda,$$

where χ is a real character associated with an r-dimensional irreducible representation of the compact gauge group, g_{p_x} is the oriented product of group elements along the boundary of the plaquette p_x located at x and perpendicular to the 1-direction. The expectations $\langle\cdot\rangle_\Lambda$ are finite lattice $\Lambda \subset Z^3$ averages in the Gibbs ensemble with Boltzmann factor $\exp(\beta_{PC}\,\Sigma_{p\subset\Lambda}\chi(g_p))$ (β is related to the coupling constant g by $\beta = 2/g^2$) and measure dg_Λ, the product of Haar measures $d\mu$ of the gauge group, one for each bond of Λ. The existence, β-analyticity and translation invariance of the $\Lambda \uparrow Z^3$ limit is established in Ref. 3 (or see Ref. 4) for small $|\beta|$. We denote points $x \in Z^3$ by $x = (x_1,\mathbf{x})$, $\mathbf{x} = (x_2,x_3)$ and $|x| = \Sigma_{i=1}^{d} |x_i| = |x_1| + |\mathbf{x}|$. We let $\sim$ denote the Fourier transform, i.e., $\tilde{G}(p,\beta) = \Sigma_{x\in Z^3} e^{ipx} G(x)$, where

$$px = \sum_{i=1}^{3} p_i x_i, \quad p = (p_1,\mathbf{p}),\ p_i \in (-\pi,\pi].$$

In Ref. 1 for $\beta > 0$ and small, it is shown that there is an isolated dispersion curve $\omega(\mathbf{p})$, real analytic in $\mathbf{p}$, $\mathbf{p} \in (-\pi,\pi]^2$, $\omega(\mathbf{p}) \geqslant \omega(\mathbf{0}) = m, m$ defined by

$$m \equiv \lim_{x_1\to\infty} \frac{-1}{x_1} \ln G(x = (x_1,\mathbf{0}),\beta), \quad \beta > 0,$$

where m is interpreted as the glueball rest mass. $\omega(\mathbf{p})$ is defined by

$$\omega(\mathbf{p}) = \lim_{x_1\to\infty} \frac{-1}{x_1} \ln\left(\sum_{\mathbf{x}} e^{i\mathbf{p}\cdot\mathbf{x}} G(x,\beta)\right).$$

Furthermore, the mass and dispersion curve satisfy

$$\lim_{\beta\downarrow 0} \frac{m(\beta)}{-4\ln\beta} = 1, \quad \lim_{\beta\downarrow 0} \frac{\omega(\mathbf{p})}{m} = 1,$$

uniformly in $\mathbf{p} \in (-\pi,\pi]^2$.

In Ref. 5 strong coupling expansions of $m(\beta)$ have been obtained for some gauge groups. Our results, stated as Theorems I and II, provide a quasianalytic, computable, convergent perturbation theory for the dispersion curve.

Theorem I: (a) There exists a function $r(\beta)$, depending on the representation of the gauge group, analytic at $\beta = 0$, such that, for all $\beta > 0$ and small, $m(\beta)$ admits the representation $m(\beta) = -4\ln\beta + r(\beta)$,

(b) There exists a $\beta' > 0$ such that for each $\beta \in (0,\beta')$ $m(\beta)$ is analytic.

Remark: β' may be larger than the radius of convergence of the $\beta = 0$ Taylor series of $r(\beta)$.

Theorem II:

$$b_n \equiv \frac{1}{n!}\frac{d^n r}{d\beta^n}(\beta = 0),$$

the nth $\beta = 0$ Taylor series coefficient of $r(\beta)$, can be computed by a finite algorithm. b_n depends on a finite number of the $\beta = 0$ Taylor series coefficients of the finite lattice correlation function at a finite number of points, increasing with n, of Z^3.

Remark 1: It will be seen in the course of the proof of Theorem II that, by using the polymer expansion of Ref. 4, $(d^k/d\beta^k)G(0;x,\beta=0) = d^k G_\Lambda(0;x,\beta=0)/d\beta^k$, where $|\Lambda|$, the number of points in Λ, depends on and increases with $|x|$ and k. We emphasize that there are no troublesome volume-dependent terms that have to be canceled in the thermodynamic limit as occurs, for example, in formally applying the Rayleigh–Schrödinger expansion to the time-continuum Hamiltonian version of the model.[6]

Remark 2: Similar results hold for the dispersion curve $\omega(\mathbf{p})$ and for abelian gauge groups with $\mathrm{Re}\,\chi$ replacing χ in the definition of the correlation function.

In Sec. II we give the proofs of Theorems I and II. In Sec. III we make some concluding remarks.

II. PROOFS OF THEOREMS I AND II

The proofs of Theorems I and II will be given after some preliminary lemmas. The results of Ref. 1 as well as ours follow from β analyticity and x-decay properties of $G(x,\beta)$ and $\Gamma(x,\beta)$, where $\Gamma(x,\beta) \equiv \Gamma(y;z,\beta)$, $x = y - z$ is the

 0022-2488/84/092741-03$02.50

convolution inverse of $G(x;y,\beta)$ interpreted as a matrix operator on $l_2(Z^3)$, i.e., $\Gamma(x;y) = G^{-1}(x;y)$, $\Sigma_z \Gamma(x;z)\, G(z;y) = \delta_{xy}$. Specifically $\Gamma(x,\beta)$ has faster x_1 falloff than $G(x,\beta)$. We state the results of Ref. 1 (see also Ref. 7) in the form needed here as Lemmas 1 and 2. Let $\|G\|$ ($\|\Gamma\|$) denote the $l_2(Z^3)$ operator norm of G (Γ). In what follows, all results are to be understood as holding for all sufficiently small $|\beta|$; $c, c', c_1, c_2 \ldots$ will denote strictly positive constants.

Lemma 1: (a) There exist c, c_1, c_2 such that $G(x,\beta)$ is analytic in β,

$$|G(x,\beta)| \leqslant c_1 |c\beta|^{4|x_1|+|\mathbf{x}|}, \quad \|G\| < c_2$$

and

$$G(x_1,\mathbf{x},\beta) = G(-x_1,\mathbf{x},\beta).$$

(b) $\tilde{G}(p,\beta)$ is jointly analytic in β and p_1, $|\mathrm{Im}\, p_1| < -4 \ln|c\beta|$, the x sum converges absolutely.

Lemma 2: (a) There exist c', c_3, c_4 such that $\Gamma(x,\beta)$ is analytic in β,

$$|\Gamma(x,\beta)| \leqslant c_3 |c'\beta|^{5|x_1|+|\mathbf{x}|}, \quad x \neq (\pm 1, 0);$$

for $x = (\pm 1, \mathbf{0})$ replace the 5 by 4, $\|\Gamma\| < c_4$ and $\Gamma(x_1,\mathbf{x},\beta) = \Gamma(-x_1,\mathbf{x},\beta)$.

(b) $\Gamma = G^{-1} = (I + P^{-1}(G-P))^{-1} P^{-1} = \Sigma_{n=0}^{\infty} (-1)^n [P^{-1}(G-P)]^n P^{-1}$ is convergent in norm where $P: l_2(Z^3) \to l_2(Z^3)$ has matrix elements $P(x;y) = G(x,y)\, \delta_{xy}$ and $|G(0,\beta)|$ is bounded away from zero.

(c) $\tilde{\Gamma}(p,\beta)$ is jointly analytic in β and p_1, $|\mathrm{Im}\, p_1| < -5 \ln|c'\beta|$, the x sum converging absolutely.

(d) $\tilde{\Gamma}(p,\beta)\, \tilde{G}(p,\beta) = 1$ in the analyticity region of $\tilde{G}(p,\beta)$.

(e) For $\beta > 0$ $p_1 = i\omega(\mathbf{p})$ satisfies $\tilde{\Gamma}(p_1 = i\omega(\mathbf{p}), \mathbf{p}, \beta) = 0$ and is the only zero of $\tilde{\Gamma}(p,\beta)$ in $0 < \mathrm{Im}\, p_1 < -5 \ln|c'\beta|$, $|\mathrm{Re}\, p_1| < \pi$, is simple and $(\partial\tilde{\Gamma}/\partial p_1)(p_1 = i\omega(\mathbf{p}), \mathbf{p}, \beta) = Z'(\mathbf{p}) > 0$.

The proofs of Theorems I and II are based on the implicit equation for $m(\beta)$ of Lemma 2(e) written as the $\beta = 0$ Taylor expansion of $\tilde{\Gamma}(p,\beta)$, where the terms up to and including order β^4 are explicited and are obtained from the $\beta = 0$ Taylor expansion of $\tilde{G}(p,\beta)$ (see also Ref. 8). We let

$$G_s(x,\beta) \equiv G(x,\beta) - \sum_{m=0}^{4} \frac{\beta^m}{m!} \frac{d^m G}{d\beta^m}(x, \beta = 0)$$

$$= \frac{\beta^5}{4!} \int_0^1 (1-t)^4 \frac{\partial^5 G}{\partial \xi^5}(x, \xi = \beta t)\, dt$$

and let $\tilde{G}_s(p,\beta)$ be the Fourier transform; $\Gamma_s(x,\beta)$ and $\tilde{\Gamma}_s(p,\beta)$ are defined similarly. We also let $\Gamma_s(n,\beta) \equiv \Sigma_{\mathbf{x}} \Gamma_s(x_1 = n, \mathbf{x}, \beta)$ for $n = 0, 1, 2, \ldots$ and set $\tilde{\Gamma}(p_1,\beta) \equiv \tilde{\Gamma}(p_1, \mathbf{p} = 0, \beta)$.

Lemma 3: There exist c, c_5, c_6 such that

(a) $|G_s(x,\beta)| \leqslant c_5 |\beta|^5$,

(b) For all p, $\mathbf{p}$ real, $|\mathrm{Im}\, p_1| < -(4 - \frac{1}{8}) \ln|c\beta|$,

$$\tilde{G}(p,\beta) = 1 + g_1\beta + g_2\beta^2 + g_3\beta^3 + g_4\beta^4 + (\beta^4/r^4)(e^{-ip_1} + e^{ip_1}) + \tilde{G}_s(p,\beta);$$

the x series of $\tilde{G}_s(p,\beta)$ converges absolutely and $|\tilde{G}_s(p,\beta)| \leqslant c_6 |\beta|^5$. g_i, $1 \leqslant i \leqslant 4$, are group representation-dependent constants given by, with $I_m \equiv \int \chi^m(g)\, d\mu(g)$,

$$g_1 = I_3, \quad g_2 = (I_4 - 3)/2, \quad g_3 = (I_5 - 10 I_3)/3!,$$

$$g_4 = (I_6 - 15 I_4 - 10 I_3^2 + 30)/4!.$$

Proof: (a) follows by a Cauchy estimate on $(\partial^5 G/\partial \xi^5)(x, \xi = \beta t)$ using Lemma 1(a).

(b) By Lemma 1(a) for x satisfying $4|x_1| + |\mathbf{x}| \geqslant 5$, $G_s(x,\beta) = G(x,\beta)$, $|G_s(x,\beta)| \leqslant c_1 |c\beta|^{4|x_1|+|\mathbf{x}|}$, and using (a) for the other x, $|\tilde{G}_s(p,\beta)| \leqslant c_6 |\beta|^5$ follows by summing these bounds, multiplied by $e^{|\mathrm{Im}\, p_1|\, |x_1|}$, over x. For x such that $4|x_1| + |\mathbf{x}| \leqslant 4$, by the cluster expansion of Ref. 3, it follows that $G(x,\beta) = \lim_{\Lambda \uparrow Z^3} G_\Lambda(0;x,\beta)$ uniformly in $|\beta|$ which implies that the $\beta = 0$ derivatives of $G(x,\beta)$ are equal to the Z^3 limits of the $\beta = 0$ derivatives of $G_\Lambda(0;x,\beta)$, $0, x \in \Lambda \subset Z^3$. The $\beta = 0$ Taylor series coefficients of $G_\Lambda(0;x,\beta)$ are obtained from the quotient series separately expanding the numerator and denominator (partition function) of the expectations and using the Peter–Weyl (PW) theorem.[9] Note that for finite Λ the denominator is analytic at $\beta = 0$. From Ref. 1 $G(x_1 = 1, \mathbf{x} = 0, \beta) = (1/r^4)\beta^4 + O(\beta^5)$; expanding the numerator of $G_\Lambda(x_1 = 0, \mathbf{x}, \beta)$, written in terms of duplicate variables, and applying the PW theorem we find that $G(x_1 = 0, \mathbf{x}, \beta) = O(\beta^5)$ for $|\mathbf{x}| \geqslant 2$. Thus we are left with $G(x = 0, \beta)$ and $G(x_1 = 0, \mathbf{x}, \beta)$, $|\mathbf{x}| = 1$. By direct calculation we find

$$G(x = 0, \beta) = 1 + g_1\beta + g_2\beta^2 + g_3\beta^3 + g_4\beta^4 + O(\beta^5),$$

$$G(x_1 = 0, \mathbf{x}, \beta) = O(\beta^5), \quad |\mathbf{x}| = 1.$$

Lemma 4: There exist c', c_7, c_8, c_9 such that

(a) $|\Gamma_s(x,\beta)| \leqslant c_7 |\beta|^5$,

(b) for all p, $\mathbf{p}$ real, $|\mathrm{Im}\, p_1| < (5 - \frac{1}{8}) \ln|c'\beta|$,

$$\tilde{\Gamma}(p,\beta) = 1 + \gamma_1\beta + \gamma_2\beta^2 + \gamma_3\beta^3 + \gamma_4\beta^4 - (\beta^4/r^4)(e^{-ip_1} + e^{ip_1}) + \tilde{\Gamma}_s(p,\beta);$$

the x series of $\tilde{\Gamma}_s(p,\beta)$ converges absolutely and $|\tilde{\Gamma}_s(p,\beta)| \leqslant c_8 |\beta|^5$; γ_i, $1 \leqslant i \leqslant 4$, are given by

$$\gamma_1 = -g_1, \quad \gamma_2 = -g_2 + g_1^2, \quad \gamma_3 = -g_3 + 2g_1g_2 - g_1^3,$$

$$\gamma_4 = -g_4 + g_1^4 - g_1^2 g_2 + g_2^2 - 2g_1^2 g_2 + 2g_1 g_3.$$

(c) $\Gamma_s(n,\beta)/\beta^{4n}$ is analytic, the $\mathbf{x}$ series converges absolutely, and

$$|\Gamma_s(n,\beta)| \leqslant c_9 |c\beta|^{5n}, \quad n \neq 0; \quad |\Gamma_s(n=0,\beta)| \leqslant c_9 |\beta|^5.$$

Proof: (a) follows by a Cauchy estimate on $(\partial^5 \Gamma/\partial \xi^5)(x, \xi = \beta t)$.

(b) The bound $|\tilde{\Gamma}_s(p,\beta)| \leqslant c_8 |\beta|^5$ follows as in the proof of Lemma 4(b) but uses Lemma 2(a). The coefficients of β^m, $0 \leqslant m \leqslant 4$, are found using Lemma 2(d) and 3(b).

(c) from Lemma 2(a) for x such that $5|x_1| + |\mathbf{x}| \geqslant 5$, $x \neq (\pm 1, \mathbf{0})$, $\Gamma_s(x,\beta) = \Gamma(x,\beta)$, $|\Gamma_s(x,\beta)| \leqslant c_3 |c'\beta|^{5|x_1|+|\mathbf{x}|}$ and using (a) for the other x the result follows on summing these bounds over $\mathbf{x}$ with $x_1 = n$.

We now give the proofs of the theorems.

Proof of Theorem I (b): $\tilde{\Gamma}(p_1,\beta)$ is jointly analytic in p_1 and β for $\beta \in (0, \beta')$, $p_1 = \mathrm{Im}(\beta)$, by Lemma 2(d). From Lemma 2(c) $(\partial\tilde{\Gamma}/\partial p_1)(p_1 = \mathrm{Im}(\beta), \beta) \neq 0$ so the result follows from the analytic implicit function theorem.[10,11]

Proof of Theorem I (a): We write $\tilde{\Gamma}(p_1,\beta) \equiv \tilde{\Gamma}(p_1, \mathbf{p} = 0, \beta)$ as

$$\tilde{\Gamma}(p_1,\beta) = 1 + \gamma_1\beta + \gamma_2\beta^2 + \gamma_3\beta^3 + \gamma_4\beta^4 - (\beta^4/r^4)(e^{-ip_1} + e^{ip_1}) + \Gamma_s(n=0,\beta) + \sum_{n=1}^{\infty} \Gamma_s(n,\beta)(e^{-ip_1 n} + e^{ip_1 n}).$$

Consideration of the asymptotic form of $m(\beta)$, i.e., $m(\beta) \sim -4\ln\beta$, motivates the introduction of the auxiliary complex variable w and function $H(w,\beta)$ such that $H(w=-(\beta^4/r^4)e^{-ip_1}+1,\beta)=\tilde{\Gamma}(p_1,\beta)$, where

$$H(w,\beta)=w+\gamma_1\beta+\gamma_2\beta^2+\gamma_3\beta^3+\gamma_4\beta^4 - \beta^8/r^8(1-w)+\Gamma_s(n=0,\beta) + \sum_{n=1}^{\infty}\Gamma_s(n,\beta)\left[\left(\frac{r^4(1-w)}{\beta^4}\right)^n + \left(\frac{\beta^4}{r^4(1-w)}\right)^n\right].$$

We show below that $H(w,\beta)$ is jointly in w,β for $|w|$, $|\beta|$ small, $H(0,0)=0$, $(\partial H/\partial w)(0,0)=1$. By the analytic implicit function theorem there exists a unique analytic function $w(\beta)$ *for* $|\beta|$ small such that $H(w(\beta),\beta)=0$, $w(0)=0$. As the only zero of $\tilde{\Gamma}(p_1,\beta)$ is at $p_1=\mathrm{im}(\beta)$ for $\beta>0$ by Lemma 2(e) we have $w(\beta)=-(\beta^4/r^4)e^{m(\beta)}+1$ or $m(\beta)=-4\ln\beta+4\ln r+\ln(1-w(\beta))$. From Lemma 4(c) each term of $H(w,\beta)$ is analytic for $|w|<1$ and by the ratio test the series is uniformly convergent for $|w|$ small. Thus $H(w,\beta)$ is analytic for $|w|$, $|\beta|$ small, and $H(0,0)=1$. A similar analysis shows that

$$\frac{\partial H}{\partial w}(w,\beta)=1-\frac{\beta^8}{r^8(1-w)^2}+\sum_{n=1}^{\infty}\Gamma_s(n,\beta) \times\left[\left(\frac{r^4}{\beta^4}\right)^n(-n)(1-w)^{n-1} + \left(\frac{\beta^4}{r^4}\right)^n\frac{n}{(1-w)^{n+1}}\right]$$

and $(\partial H/\partial w)(0,0)=1$.

Proof of Theorem II: From the above proof $m(\beta)=-4\ln\beta+r(\beta)$, $r(\beta)=4\ln r+\ln(1-w(\beta))$, $r(\beta)=\Sigma_{n=0}^{\infty}b_n\beta^n$, provides a convergent expansion for $m(\beta)$. b_n is determined from a finite number of the $(1/n!)(d^n w(0)/d\beta^n)$ which can be calculated in terms of w,β partial derivatives of H at $(w,\beta)=(0,0)$ (see Ref. 11) and involves a finite number of $\beta=0$ derivatives of $\Gamma_s(m,\beta)$. For examples the first two derivatives are, letting $D_\beta^n\equiv\partial^n/\partial\beta^n$, $D_w^m=\partial^m/\partial w^m$,

$$\frac{dw}{d\beta}=-D_\beta H(D_w H)^{-1},$$

$$\frac{d^2w}{d\beta^2}=-[D_\beta D_w H D_\beta w+D_\beta^2 H](D_w H)^{-1} + D_w H(D_w H)^{-2}[D_w^2 H D_\beta w+D_w D_\beta H].$$

Using the falloff in Lemma 4(c) only a finite number of m contribute and using the falloff of $\Gamma(m,\mathbf{x},\beta)$ only a finite number of $\mathbf{x}$ contribute. Using Lemma 1(a) only a finite number of terms in Lemma 2(b) contribute to $(d^k\Gamma/d\beta^k)(m,\mathbf{x},\beta=0)$ and again by Lemma 1(a) only a finite number of $\beta=0$ derivatives of $G(x,\beta)$ and a finite number of x contribute.

We now turn to the calculation of $g_k(x)$, where $G(0;x,\beta)=\Sigma_{k=0}^{\infty}g_k(x)\beta^k$. By the β analyticity of $G_\Lambda(0;x,\beta)$ and the uniform convergence in β of $G_\Lambda(0;x,\beta)$ to $G(0;x,\beta)$

$$g_k(x)=\lim_{\Lambda\uparrow Z^3}g_{\Lambda k}(x),$$

where $G_\Lambda(0;x,\beta)=\Sigma_k g_{\Lambda k}(x)\beta^k$. For sufficiently small $|\beta|$

$$g_{\Lambda k}(x)=\frac{1}{k!}\frac{d^k}{d\beta^k} \times\left\{L\equiv\left[\frac{\partial^2}{\partial\alpha_0\partial\alpha_x}\ln Z'_\Lambda(\alpha_0,\alpha_x)\right]_{\alpha_0=\alpha_x=0}\right\}\Bigg|_{\beta=0},$$

where $Z'_\Lambda(\alpha_0,\alpha_x)$ is the partition function with the Boltzmann factor $\exp\{\beta\Sigma_p\chi(g_p)+\alpha_0\chi(g_{p_0})+\alpha_x\chi(g_{p_x})\}$. By the polymer expansion of Ref. 4 $\ln Z'_\Lambda$ and L can be expressed as an infinite sum over connected sets of polymers containing 0 and x. However, as the activity of the polymer γ is less than $c'|\beta|^{c|\gamma|}$ for some c,c', where $|\gamma|$ denotes the size of γ, only a finite sum over a finite number of polymers occurs in $(d^k/d\beta^k)L|_{\beta=0}$.

III. CONCLUDING REMARKS

Similar methods, with similar results, apply to other lattice models, such as the high temperature[12] and pure phase low temperature[13] Ising model as well as to $P(\phi)$ models.[14] In the above models the mass is nondegenerate. The case of asymptotically ($\beta\downarrow 0$) degenerate masses as occurs in multicomponent spin, complex character gauge, and gauge–Higgs models requires a degenerate perturbation theory and is developed in Ref. 15. A convergent perturbation theory for the mass or masses of the time-continuum Hamiltonian versions of infinite space lattice spin, gauge,and gauge-matter models[6] has yet to be developed.

[1] R. Schor, Nucl. Phys. B **222**, 71–82 (1983).
[2] K. Wilson, Phys. Rev. D **10**, 2445 (1974).
[3] K. Osterwalder and E. Seiler, Ann. Phys., New York **110**, 440 (1978).
[4] E. Seiler, "Gauge Theories as a Problem of Constructive Quantum Field Theory and Statistical Mechanics," *Lecture Notes in Physics*, Vol. 159 (Springer, New York, 1982).
[5] G. Munster, Nucl. Phys. B **190** [FS3], 439–453 (1981).
[6] J. Kogut, Rev. Mod. Phys. **51**, 659 (1979).
[7] R. Schor, Commun. Math. Phys. **59**, 213 (1978).
[8] P. Paes-Leme, Ann. Phys., New York **115**, 367–387 (1978).
[9] M. A. Naimark, *Normed Rings* (Noordhoff, Gronigen, The Netherlands, 1970).
[10] E. Hille, *Analytic Function Theory* (Ginn, Boston, 1962), Vols. I and II.
[11] J. Dieudonné, *Foundations of Modern Analysis* (Academic, New York, 1969).
[12] M. O'Carroll, J. Stat. Phys. (to appear Feb. 1984).
[13] M. O'Carroll and W. Barbosa, J. Stat. Phys. (to appear Feb. 1984).
[14] J. Glimm and A. Jaffe, *Quantum Physics* (Springer, New York, 1981).
[15] M. O'Carrol, G. Braga, and R. Schor, "Glueball Mass Spectrum and Mass Splitting in Z^3 Strongly Coupled Lattice Gauge Theories," UFMG Dept. of Physics preprint, Sept. 1983.

A superspace approach to lattice supersymmetry

V. Alan Kostelecký
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545
Jeffrey M. Rabin
Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637

(Received 23 November 1983; accepted for publication 6 April 1984)

We construct a cubic lattice of discrete points in superspace, as well as a discrete subgroup of the supersymmetry group which maps this "superlattice" into itself. We discuss the connection between this structure and previous versions of lattice supersymmetry. Our approach clarifies the mathematical problems of formulating supersymmetric lattice field theories and suggests new methods for attacking them.

PACS numbers: 11.30.Pb, 11.15.Ha

I. INTRODUCTION

The recent results of Mandelstam[1] concerning the vanishing to all orders in perturbation theory of the β-function for the $N = 4$ supersymmetric Yang–Mills model make any nonperturbative analysis of great interest. Unfortunately, no satisfactory method of placing supersymmetric gauge theories on a lattice is known. The basic problem is associated with a general feature of any supersymmetry algebra, namely that the anticommutator of two supersymmetry transformations is a translation. The introduction of a lattice, which restricts the translations, therefore implies that the supersymmetry algebra cannot be preserved.

There have been several attempts to avoid this difficulty. One approach[2] involves using superspace, with the space-time component latticized; this method successfully places the Wess–Zumino model[3] on the lattice, but cannot readily be extended to gauge theories. Other authors[4] attempt to keep only the Hamiltonian part of the supersymmetry algebra on the lattice; unfortunately, the resulting models have difficulties with Lorentz invariance in the continuum limit. None of these methods are totally satisfactory.

The present work is based on the philosophy that one should not try to maintain the supersymmetry algebra in a lattice theory. Only internal symmetries are represented in lattice theories via the algebra of their infinitesimal generators; space-time symmetries are always represented at the group level rather than at the algebra level. In general, a continuum space-time symmetry group is represented on the lattice by a discrete subgroup which acts as a group of motions of the lattice points themselves. For example, the Poincaré group on the lattice is reduced to its discrete subgroup of unit translations and cubic rotations. Since supersymmetry is an extension of the Poincaré group, and is therefore a space-time symmetry, we believe it is correctly represented on the lattice as a discrete subgroup.

We shall construct a superspace lattice in which both the space-time and the anticommuting dimensions become discrete. The basic formalism relies on the mathematical analysis of continuum superspace given by Rogers.[5] It will be shown that a discrete subgroup of the supersymmetry group can be defined on the superspace lattice. The desirability of such a construction has been recognized by other authors,[4] but no explicit realization of it was given. We also show how to construct supersymmetry invariants by summing superfields over the superlattice sites. We then discuss the θ-expansion of our superfields. Within Rogers' formalism, and ours, the existence of a θ-expansion is *not* an automatic property of "the most general superfield"; rather, it is an extremely strong *constraint* imposed on a superfield. We consider this the most important lesson of the superlattice approach: constructing a supersymmetric lattice field theory amounts to defining an action of a discrete group on a constrained field, and constructing invariants. The constraints are conditions of analyticity in the Grassmann coordinates, and are so natural in continuum superspace that they are assumed without comment. In the lattice context, however, they become problematic.

The outline of this paper is as follows. In Sec. II we discuss the definition and properties of the superlattice. In Sec. III we discuss the relation between our approach and previous versions of lattice supersymmetry. In particular, we show how the results of Dondi and Nicolai[2] follow from ours once the analyticity conditions mentioned above are imposed on our superfields. However, our approach is more general and leaves open the possibility of weakening these constraints. This and other directions for future work are suggested in Sec. IV. We are particularly intrigued by the analogy between Rogers' notion of analyticity and ordinary complex analyticity, which deserves more exploration.

II. THE SUPERLATTICE

Our approach to lattice supersymmetry is based on the superspace construction of Rogers.[5] This section first reviews Rogers' formalism, and then discusses the superlattice.

Recall that a real Grassmann algebra B_L over R^L is defined as an associative algebra, under an operation $\wedge$, which satisfies the following two conditions.

(1) Each pair of vectors v, w in R^L satisfies $v \wedge w = -w \wedge v$.

(2) B_L contains all vectors in R^L and may be generated from them and from scalar multipliers using the operation $\wedge$.

A basis for B_L therefore consists of the identity $v_I = I$, of a

 0022-2488/84/092744-05$02.50

set of L vectors v_i, $i=1,\ldots,L$, and of all nonvanishing $\wedge$-products of these vectors, denoted $v_{ij},\ldots,v_{ij\cdots k},\ldots,v_{123\cdots L}$, $i<j<\cdots<k$. Here, since complex conjugation of a vector in B_L is defined to reverse the order of anticommuting factors, we define

$$v_{ij\cdots k}=v_i\wedge v_j\wedge\cdots\wedge v_k, \tag{1}$$

or

$$v_{ij\cdots k}=i\,v_i\wedge v_j\wedge\cdots\wedge v_k, \tag{2}$$

as necessary to maintain $v^*_{ij\cdots k}=v_{ij\cdots k}$. In this way the basis vectors of B_L are all real. There are 2^L such basis vectors; we will denote them collectively by v_Γ. The subset of vectors generated by I and by even $\wedge$-products of v_i spans what will be referred to as the even part ${}^0B_L$ of B_L, while the subset of vectors generated by v_i and by odd $\wedge$-products of v_i spans the odd part ${}^1B_L$ of B_L. In what follows, where no confusion can arise, we shall omit the $\wedge$-symbol.

Now, if A is a vector in B_L, then it may be expanded in terms of the basis v_Γ as follows:

$$A=\sum_\Gamma A_\Gamma v_\Gamma, \tag{3}$$

where $A_\Gamma\in\mathbb{R}$ are the components of A. Then a norm, denoted $\|\ \|$, can be defined on B_L by

$$\|A\|=\sum_\Gamma|A_\Gamma|. \tag{4}$$

Rogers chose this norm as it leads to a finer topology on superspace than those used by other authors.[6]

Two useful projection maps acting on B_L will now be introduced. The first, denoted ϵ, is defined by

$$\epsilon:B_L\to\mathbb{R},\quad \epsilon(A)=A_I. \tag{5}$$

The second, denoted s, is defined in terms of ϵ by

$$s:B_L\to B_L,\quad s(A)=A-\epsilon(A)I. \tag{6}$$

The projections $\epsilon(A)$ and $s(A)$ are the analogs for Grassmann numbers of the real and imaginary parts of complex numbers. As will be seen later, these maps play an important role in the definition and physical interpretation of superfields.

An (m,n)-superspace $B_L^{m,n}$, with m even and n odd coordinates, is then defined as the Cartesian product of m copies of ${}^0B_L$ and n copies of ${}^1B_L$. It is a $2^{L-1}(m+n)$-dimensional space. A point P in $B_L^{m,n}$ is specified by

$$P=(\hat{x}^1,\ldots,\hat{x}^m;\theta^1,\ldots,\theta^n), \tag{7}$$

where $\hat{x}^i\in{}^0B_L$ and $\theta^\alpha\in{}^1B_L$ for $i=1,\ldots,m$ and $\alpha=1,\ldots,n$. We use carets on the even elements $\hat{x}^i$ to distinguish them from their $\mathbb{R}^m$-space projections; thus, by definition,

$$\epsilon(\hat{x}^i)=x^i. \tag{8}$$

The norm of P is defined by

$$\|P\|=\sum_i\|\hat{x}^i\|+\sum_\alpha\|\theta^\alpha\|. \tag{9}$$

Here, we wish to stress the distinction between v_i and θ^α: the v_i are fixed elements in the basis of B_L, whereas the θ^α are variables taking values in ${}^1B_L$.

As an example, consider the simple case of the four-dimensional Grassmann algebra B_2 over $\mathbb{R}^2$. The basis for this algebra is the set of vectors $\{I,v_1,v_2,v_{12}=iv_1\wedge v_2\}$. An element A of B_2 may be written

$$A=aI+bv_1+cv_2+dv_{12}, \tag{10}$$

for $a,b,c,d\in\mathbb{R}$. The norm of A is then

$$\|A\|=|a|+|b|+|c|+|d|. \tag{11}$$

We can define a (1,1)-superspace $B_2^{1,1}$ based on B_2. A point P in this superspace is written

$$\begin{aligned}P&=(\hat{x},\theta)\\&=(aI+dv_{12},bv_1+cv_2),\end{aligned} \tag{12}$$

for $a,b,c,d\in\mathbb{R}$.

Rogers proceeds further with an investigation of functions on $B_L^{m,n}$. The starting point is the generalization of C^k functions on $\mathbb{R}^m$ to the so-called G^k functions on $B_L^{m,n}$. Eventually, the G^∞ functions on $B_L^{m,n}$ are shown to be those which have a θ-expansion with coefficients defined by a map z. The concept of a G^k function is not relevant to our discussion, however, so we will not review further its definition and properties. See Ref. 5 for more details. Nonetheless, we will use the θ-expansion later, and so we proceed now to outline it.

First consider the z-map of an arbitrary B_L-valued C^∞ function $f(x^i)$. This is defined by $z:C^\infty(\mathbb{R}^m)\to G^\infty(B_L^{m,n})$ with

$$\begin{aligned}&z[f](\hat{x}^i,\theta^\alpha)\\&\quad=\sum_{j_1=0\cdots j_m=0}^{L}\frac{1}{j_1!\cdots j_m!}\\&\qquad\times\left[\partial_1^{j_1}\cdots\partial_m^{j_m}f(x^i)\right]\wedge s(\hat{x}^1)^{j_1}\wedge\cdots\wedge s(\hat{x}^m)^{j_m}.\end{aligned} \tag{13}$$

Here ∂_j denotes the jth partial derivative, and $s(\hat{x}^i)^k$ denotes k $\wedge$-products of $s(\hat{x}^i)$. Note that $z[f]$ is independent of θ^α, i.e.,

$$z[f](\hat{x}^i,\theta^\alpha)=z[f](\hat{x}^i). \tag{14}$$

The z-map is introduced to make explicit the distinction between functions of x^i and functions of $\hat{x}^i$. It is best viewed as an analytic continuation map which extends functions of x^i to functions of $\hat{x}^i$ such that

$$z[f](x^iI)=f(x^i). \tag{15}$$

The analytic continuation is defined via Taylor expansion about $\hat{x}^i=x^iI$ in powers of $s(\hat{x}^i)=\hat{x}^i-x^iI$. Note that functions extended from space-time to superspace using the z-map enjoy all the analyticity properties assumed in standard superfield manipulations.

Now, the important feature of the z-map for our purposes is that it may be used to define θ-expansions. The relevant proposition [2.11(d) in Ref. 5] states that for G^∞ functions f there exist uniquely defined C^∞ functions f_μ such that

$$f(\hat{x}^i,\theta^\alpha)=\sum_\mu z[f_\mu](\hat{x}^i)\wedge\theta^\mu, \tag{16}$$

where μ is an index spanning all distinct nonzero $\wedge$-products of elements in the set $\{\theta^\alpha,I\}$. Rogers calls this the z-expansion of f; for $B_L^{m,n}=B_L^{4,4}$ it is equivalent to the usual superfield θ-expansion in (4,4)-superspace. Note that the dimension L, usually undefined, must satisfy $L\geqslant n$, or else the θ-expansion will have trivial terms. In this work we shall always assume L is sufficiently large to avoid such problems.

As an example, let us consider $B_2^{1,1}$. The z-map reduces to

$$z[f](\hat{x},\theta) = f(x) + \partial_x f(x)\cdot s(\hat{x}). \tag{17}$$

Note that the term $\frac{1}{2}\partial_x^2 f(x)\cdot s(\hat{x})\wedge s(\hat{x})$ is zero in this special case of $L=2$, because $s(\hat{x})$ is just a multiple of v_{12}. The θ-expansion becomes

$$f(\hat{x},\theta) = z[f_I](\hat{x}) + z[f_\theta](\hat{x})\wedge\theta. \tag{18}$$

For instance, the function $f(\hat{x})=\hat{x}^2$ is associated with uniquely defined $f_I(x)=x^2$ and $f_\theta(x)=0$, because $\hat{x}^2 = x^2 + 2x\cdot s(\hat{x}) = z[x^2](\hat{x})$. Note also that f_θ can be taken as a C^∞ function with values in the odd part of B_L. Then f_θ anticommutes with itself, as fermions should.

We now turn to the construction of the superspace lattice. It is defined as the subset of $B_L^{m,n}$ for which the components of each θ^α are integer multiples of a real number k and for which the components of each $\hat{x}^i$ are integer multiples of k^2. Thus the superlattice is spaced by k in the θ^α-directions, while it is spaced by k^2 in the $\hat{x}^i$ directions. For instance, in the previous example of $B_2^{1,1}$, Eq. (12), the components a and d are restricted to integer multiples of k^2 while b and c are integer multiples of k. In what follows, we shall set $k=1$ for brevity.

To motivate the superlattice definition, let us restrict ourselves to $B_L^{4,4}$, the usual (4,4)-superspace. Now, since B_L is defined over $\mathbb{R}^L$, $\theta^\alpha = \theta^{\alpha *}$. Therefore θ^α is a Majorana spinor. Accordingly, we will work in the Majorana representation of the Dirac γ^i-matrices in which $\gamma^{i*} = -\gamma^i$, so that the matrices $i\gamma^i$ have real integer entries. Then a supersymmetry transformation $Q(\eta)$, by a finite parameter η^α with the same discrete values as θ^α, is given by

$$Q(\eta)(\hat{x}^i,\theta^\alpha) = (\hat{x}^i - i\bar{\eta}\gamma^i\theta, \theta^\alpha + \eta^\alpha). \tag{19}$$

The transformed point is evidently still on the superlattice, provided there is no boundary; we therefore always work on an infinitely extended superlattice. The superlattice is then invariant under a discrete group of supersymmetry transformations.[7] This group is generated by $Q(\alpha)$ and by the translation operator $T(\hat{a})$ which acts as

$$T(\hat{a})(\hat{x}^i,\theta^\alpha) = (\hat{x}^i + \hat{a}^i, \theta^\alpha). \tag{20}$$

An arbitrary element of the group may be written in the form $Q(\alpha)T(\hat{a})$, by using the relations

$$Q(\alpha)Q(\beta) = Q(\alpha+\beta)T(-i\bar{\alpha}\gamma^i\beta), \tag{21}$$

$$Q(\alpha)T(\hat{a}) = T(\hat{a})Q(\alpha), \tag{22}$$

$$T(\hat{a})T(\hat{b}) = T(\hat{a}+\hat{b}). \tag{23}$$

The analog of the usual anticommutation relation is

$$Q(-\beta)Q(-\alpha)Q(\beta)Q(\alpha) = T(2i\bar{\alpha}\gamma^i\beta). \tag{24}$$

In this paper we consider only the proper supersymmetry group generated by the translations and supertranslations, and not its extension containing Lorentz transformations also. We note in passing that due to the half-angles involved in the spinor transformation law, the components of θ^α will not remain integral after a 90° rotation, and hence the superlattice does not admit the usual cubic rotation group. However, a discrete group of 180° rotations can be implemented.

For simplicity in constructing constrained superfields on the superlattice, it is useful to work with Weyl two-component spinors. A similarity transformation maps the Majorana representation into the Weyl one. In effect this complexifies the odd sector of the superlattice: instead of four copies of the odd part of the real Grassmann algebra B_L, there are two copies of the odd part of the complex Grassmann algebra $B_{L,C}$. Of course, there are still only four independent θ-variables. Note that since the similarity transformation involves a factor of $\sqrt{2}$, the components of the Weyl spinors on the superlattice are restricted to $(p+iq)/\sqrt{2}$ for integers p and q.

In this two-component formalism, a finite supersymmetry transformation by a parameter η is written as

$$Q(\eta_\alpha,\bar{\eta}^{\dot{\alpha}})(\hat{x}^i,\theta_\alpha,\bar{\theta}^{\dot{\alpha}}) = (\hat{x}^i + i\theta\sigma^i\bar{\eta} - i\eta\sigma^i\bar{\theta}, \theta_\alpha + \eta_\alpha, \bar{\theta}^{\dot{\alpha}} + \bar{\eta}^{\dot{\alpha}}). \tag{25}$$

The transformed point is still a lattice point because the Pauli matrices are Hermitian and have entries ± 1, $\pm i$.

It is easy to construct supersymmetry invariants from superfields on the superlattice. Let $\phi(\hat{x},\theta)$ be an arbitrary assignment of an element of B_L to each superlattice site. Then the sum over all superlattice sites

$$S = \sum_{\hat{x},\theta} \phi(\hat{x},\theta), \tag{26}$$

which exists if $\phi(\hat{x},\theta)\to 0$ fast enough as $\|(\hat{x},\theta)\|\to\infty$, is manifestly invariant under discrete supersymmetry transformations because such transformations simply permute the terms in the sum.

III. CONNECTIONS WITH PREVIOUS WORK

Let us now consider the problem of constructing a supersymmetric theory of a single scalar superfield—a superlattice version of the Wess–Zumino model. It would seem impossible to make a connection between such a superlattice theory and the Wess–Zumino model in continuum superspace unless constraints analogous to the existence of the continuum θ-expansion are imposed on the superlattice superfield. The most direct approach is to constrain the most general superlattice field $\phi(\hat{x},\theta)$ by requiring that $\phi(\hat{x},\theta)$ be a polynomial in the θ^α with coefficients of the form $z[f_\mu](\hat{x})$ such that $\phi(\hat{x},\theta)$ takes values in ${}^0B_L$, the even part of the Grassmann algebra. Here the z-map converts x-space lattice fields $f(x^i)$ to superlattice fields via the definition

$$z[f](\hat{x}^i) = \sum_{j_1,\ldots,j_4=0}^{L} \frac{1}{j_1!j_2!j_3!j_4!} \times [\nabla_1^{j_1}\nabla_2^{j_2}\nabla_3^{j_3}\nabla_4^{j_4} f(x^i)] \wedge s(\hat{x}^1)^{j_1}\wedge s(\hat{x}^2)^{j_2}\wedge s(\hat{x}^3)^{j_3}\wedge s(\hat{x}^4)^{j_4}, \tag{27}$$

where ∇_i may be any x-space lattice derivative. Of course, the class of fields obeying the constraints will be different with different choices of ∇_i.

The superlattice fields obeying these constraints enjoy analyticity properties in their arguments which are similar to those assumed for continuum superfields. We will adapt Rogers' terminology and call such fields superanalytic. For example, to derive the transformation properties of the component fields f_μ, it is necessary to know that $z[f](\hat{x}+\hat{a})$ with

$\epsilon(\hat{a}) = 0$ can be expanded about any point $\hat{x}$ in powers of the nilpotent element $\hat{a}$. This is guaranteed by the result

$$z[f](\hat{x}+\hat{a}) = \sum_{k_1,\ldots,k_4=0}^{L} \frac{1}{k_1!k_2!k_3!k_4!} \times z[\nabla_1^{k_1}\nabla_2^{k_2}\nabla_3^{k_3}\nabla_4^{k_4} f](\hat{x}) \wedge (\hat{a}^1)^{k_1} \wedge (\hat{a}^2)^{k_2} \wedge (\hat{a}^3)^{k_3} \wedge (\hat{a}^4)^{k_4}, \tag{28}$$

which is easily proved by using definition (27) and changing the summation variables. The derivation is analogous to the proof that a complex-analytic function, defined by its Taylor series about a point on the real axis, can be reexpanded about any point within the circle of convergence.

The superanalyticity constraints are very strong. The dynamical degrees of freedom are reduced from one independent element of B_L for each superlattice site to the component fields $f_\mu(x)$ on an ordinary space-time lattice. Furthermore, the dependence of $\phi(\hat{x},\theta)$ on θ and $s(\hat{x})$ is fixed to be polynomial. This allows the domain of $\phi(\hat{x},\theta)$ to be extended to continuous values of θ and $s(\hat{x})$ between the superlattice sites. It is then possible to define the action of supersymmetry transformations (19) with infinitesimal η on the field by the usual device of expanding the shifted field in a Taylor series and collecting terms. Once this is done, our formalism reduces to that of Dondi and Nicolai,[2] which was based on a superspace with discrete $\epsilon(\hat{x})$ but continuous $s(\hat{x})$ and θ.

Supersymmetry invariants of the type (26) do not exist for superanalytic superfields because polynomials in $s(\hat{x})$ and θ typically blow up as $\|(\hat{x},\theta)\| \to \infty$. For such fields one must follow the usual procedure of finding expressions which transform by a total divergence. It is possible to rewrite the invariant actions of Dondi and Nicolai in the form of sums over superlattice sites,[8] but it does not seem useful to do so because the dynamical degrees of freedom of the theory are clearly the component fields $f_\mu(x)$ which live on an ordinary space-time lattice. However, we feel that our formalism is conceptually useful because it separates the problem of constructing a lattice supersymmetry group from the problem of imposing suitable constraints on the fields. For example, Dondi and Nicolai encountered problems in working with products of superfields because of the failure of the Leibniz product rule for lattice derivatives:

$$\nabla_i(fg) \neq f\nabla_i g + (\nabla_i f)g. \tag{29}$$

This translates into $z[fg] \neq z[f]z[g]$, which implies that the product of two fields having θ-expansions will not obey the superanalyticity constraint. Our formalism allows the possibility of weakening the constraint so as to avoid this problem.

Next we would like to comment on the relation between our work and that of Kaku.[9] Both approaches share the motivation that superspace must be the natural geometrical setting for a discrete version of supersymmetry, as it is for its continuum version. However, Kaku works with a random superlattice while we construct a cubic one. This means that he hopes to obtain invariance under supersymmetry only after integrating over all positions of the lattice sites, whereas we have a discrete supergroup leaving our superlattice invariant. He also does not stress what we believe to be the crucial distinction between the class of all superfields and the subclass of superanalytic ones (the former being closed under multiplication, for example, but not the latter).

IV. DISCUSSION AND PROSPECTS

In this paper we have constructed a cubic superlattice in superspace which is left fixed by a discrete subgroup of the continuum supersymmetry group. A straightforward development of a field theory on this superlattice leads directly to the lattice version of the Wess–Zumino model discussed in Ref. 2. However, our formalism clarifies the mathematical structure of the theory and suggests new possibilities for constructing more satisfactory supersymmetric lattice theories. One obvious approach would be to seek a weaker version of the superanalyticity constraints which would be preserved under multiplication of superfields and would allow finite values for invariants of the form (26).

Recently it has become clear that there is topological content to the problem of constructing a lattice version of a given continuum field theory.[10] The lattice is viewed as a simplicial complex (random lattice) or a cell complex (cubic lattice) with support homeomorphic to the space-time manifold on which the continuum field theory is defined. Continuum field equations are transcribed to the lattice using precise correspondences between the algebras of differential forms on the continuum manifold and cochains on the lattice. This results in a canonical lattice transcription bearing a well-defined geometrical relationship to the continuum theory. Our formalism opens the possibility of extending this approach to supersymmetric theories, regarded as living naturally on a supermanifold rather than ordinary space-time. To this end we hope to exploit more fully the analogy between supermanifolds and complex manifolds already alluded to in this paper. This analogy consists of three specific points.

(1) Both the complex numbers and the Grassmann numbers contain subalgebras isomorphic to the real numbers: complex numbers with $\mathrm{Im}\, z = 0$, Grassmann numbers with $s(A) = 0$.

(2) Complex manifolds and supermanifolds can each be regarded as real manifolds of higher dimensionality: $2m$ for an m-dimensional complex manifold, $2^{L-1}(m+n)$ for an (m,n) supermanifold.

(3) Both complex manifolds and supermanifolds possess an analytic structure in addition to the real structure: complex analyticity for functions on a complex manifold, superanalyticity for functions on a supermanifold.

Of course, the analogy is spoiled by the crucial difference that the Grassmann algebra contains nilpotent elements while the complex numbers do not. Nevertheless, we believe a study of simplicial approximations to complex manifolds may shed light on the proper treatment of the superanalyticity constraints in lattice supersymmetry. Research on this possibility is in progress.

ACKNOWLEDGMENTS

One of us (J.M.R) is an Enrico Fermi Fellow. This research was supported in part by the United States Department of Energy under Contracts No. DE-AC02-76ER-

03075 and No. DE-AC02-82ER-40073, and the National Science Foundation under Contract No. PHY-83-01221.

[1]S. Mandelstam, Nucl. Phys. B **213**, 149 (1983).
[2]P. H. Dondi and H. Nicolai, Nuovo Cimento A **41**, 1 (1977).
[3]J. Wess and B. Zumino, Nucl. Phys. B **70**, 39 (1974).
[4]T. Banks and P. Windey, Nucl. Phys. B **198**, 226 (1982); V. Rittenberg and S. Yankielowicz, "Supersymmetry on the Lattice," CERN preprint TH.3263, March 1982; E. Elitzur, E. Rabinovici, and A. Schwimmer, Phys. Lett. B **119**, 165 (1982).
[5]A. Rogers, J. Math. Phys. **21**, 1352 (1980).
[6]B. Kostant, in *Differential Geometric Methods in Mathematical Physics*, edited by K. Bleuler and A. Reetz, Lecture Notes in Mathematics, Vol. 570 (Springer-Verlag, Berlin, 1977); M. Batchelor, Trans. Am. Math. Soc. **253**, 329 (1979); M. Batchelor, Trans. Am. Math. Soc. **258**, 257 (1980); F. A. Berezin and D. A. Leĭtes, Sov. Math. Dokl. **16**, 1218 (1975).
[7]Discrete supergroups of this type have been considered in a different context in A. Rogers, J. Math. Phys. **22**, 939 (1981). The discrete supersymmetry group has been independently constructed by D. M. Scott, "Lattices, Supersymmetry, and Kähler Fermions," University of Edinburgh preprint 83/255.
[8]V. A. Kostelecký and J. M. Rabin, "Supersymmetry on a Superlattice," Los Alamos preprint LA-UR-83-1373.
[9]M. Kaku, Phys. Rev. Lett. **50**, 1893 (1983).
[10]J. M. Rabin, Nucl. Phys. B **201**, 315 (1982); P. Becher and H. Joos, Z. Phys. C **15**, 343 (1982).

Symmetry properties of the configuration interaction space in relation to one- and two-particle operators: The splitting theorem

Tomislav P. Živković
Rudjer Bošković Institute, 41001 Zagreb, Yugoslavia

(Received 18 October 1983; accepted for publication 10 February 1984)

The configuration interaction (CI) space X_n built upon n electrons moving over $2n$ orthonormalized orbitals χ_i is considered. It is shown that the space X_n splits into two complementary subspaces X_n^+ and X_n^- having special properties: each state $\Psi^+ \in X_n^+$ and $\Psi^- \in X_n^-$ is "alternantlike" in the sense that it has a uniform charge density distribution over all orbitals χ_i and vanishing bond-orders between all orbitals of the same parity. In addition, matrix elements $\Gamma(ij;kl)$ of a two-particle density matrix vanish whenever four distinct orbitals are involved and there is an odd number of orbitals of the same parity. Further, $\Gamma(ij;lj) = \gamma(il)/4$ $(j \neq i,l)$, whenever (i) and (l) are of different parity. This last relation shows the connection between a two-particle (Γ) and a one-particle (γ) density matrix. "Elementary" alternant and antialternant operators are identified. These operators connect either only the states in the same subspace, or only the states in different subspaces, and each one- and two-particle symmetric operator can be represented by their linear combination. Alternant Hamiltonians, which can be represented as linear combinations of elementary alternant operators, have alternantlike eigenstates. It is also shown that each symmetric Hamiltonian possessing alternantlike eigenstates can be represented as such a linear combination. In particular, the PPP Hamiltonian describing an alternant hydrocarbon system is such a case. Complementary subspaces X_n^+ and X_n^- can be explicitly constructed using the so-called regular resonance structures (RRS's) which are normalized determinants containing mutually disjunct bond orbitals. Expressions for the derivation of matrix elements of one- and two-particle operators between different RRS's are also derived.

PACS numbers: 31.10. + z, 03.65. — w

I. INTRODUCTION

Quantum chemistry has two main goals: the development of fast and efficient procedures for the accurate predictions of molecular properties, and the development of concepts which reveal the inner structure and regularity of quantum systems. The Hartree–Fock self-consistent field molecular orbital (SCF-MO) model, configuration interaction (CI) model, perturbation configuration interaction of localized orbitals (PCILO) model,[1,2] coupled cluster expansion (CCE) model,[3,4] etc., fall mainly in the first category. Different symmetry considerations, Woodward–Hofmann rules, the notion of alternant and nonalternant compounds, etc., fall mainly into the second category. There is no clear-cut distinction between these two developments, and they are often interwoven. Thus, the use of group theory can highly enhance the efficiency of various computational methods. Similarly, the development of the Hückel and later PPP theory led to the formulation of the pairing theorem[5–7] in alternant hydrocarbon systems, and thus significantly contributed to the qualitative understanding of the remarkable properties of these compounds.

This paper discusses some general relations between the configuration interaction (CI) space and different operators acting in this space. The CI space X_n defined by n electrons moving over $2n$ orbitals is considered. For example, the CI space built upon the minimum basis set with one electron per atomic orbital is such a space.

In the second section the so-called regular resonance structures (RRS's) and matrix elements of one- and two-particle operators between these structures are considered. These structures span the CI space X_n, and they naturally lead to the splitting of this space into complementary subspaces X_n^+ and X_n^-. These subspaces are mutually orthogonal, and hence each state $\Psi \in X_n$ can be uniquely represented as a sum $\Psi = \Psi^+ + \Psi^-$, where $\Psi^+ \in X_n^+$ and $\Psi^- \in X_n^-$. In the third section the so-called splitting theorem is derived. This theorem explicitly gives the complete set of linearly independent symmetric operators having definite symmetry properties with respect to subspaces X_n^+ and X_n^-. Matrix elements of each of these operators vanish either between all the states contained in different subspaces, or they vanish between all the states contained in the same subspace. The former operators are called "alternant" and the latter "antialternant." In the fourth section different properties of states $\Psi \in X_n^+$ and $\Psi \in X_n^-$ are deduced from the splitting theorem. In particular, it is shown that each state $\Psi \in X_n^\pm$ has a uniform charge density distribution and vanishing bond-orders between all vertices of the same parity. These are properties of eigenstates of neutral alternant hydrocarbons, and hence we call states $\Psi \in X_n^\pm$ "alternantlike." In the fifth section the most general symmetric Hamiltonian possessing alternantlike eigenstates is explicitly given. The PPP Hamiltonian describing alternant hydrocarbon systems appears as a very special Hamiltonian. Simple conditions are derived so that an arbitrary symmetric Hamiltonian has alternantlike eigenstates. These conditions are expressed explicitly in terms of one- and two-particle integrals entering this Hamiltonian, and hence can be given a simple physical interpreta-

 0022-2488/84/092749-18$02.50

tion. In the sixth section nonalternant symmetric Hamiltonians are considered. Each such Hamiltonian can be represented as a sum of an alternant and an antialternant part. Accordingly, each nonalternant system can be considered to be some alternant system with antialternant perturbation. In conjunction with all other results obtained, this splitting leads to the special kind of perturbation expansion. In the seventh section the formal analogy between the present approach and the VB theory is considered. The advantage of the present approach is mainly due to the partial orthogonality of RRS's, since this orthogonality leads to the splitting of the space X_n into complementary subspaces. There is no such orthogonality in the case of the VB canonical structures.

II. CONFIGURATION INTERACTION SPACE X_n AND MATRIX ELEMENTS OF DIFFERENT OPERATORS BETWEEN REGULAR RESONANCE STRUCTURES

Let us consider the configuration interaction (CI) space X_n determined by n electrons moving over $2n$ orthonormalized orbitals $\chi_i \in B$ (see Appendix). One can partition the set B onto subsets B^o and B^x containing n "source" and n "sink" orbitals χ_i, respectively.[8,9] Next one defines regular resonance structures (RRS's): each RRS, $S \in \mathscr{R}(n)$, is a normalized determinant containing n mutually disjunct excited and/or nonexcited bond orbitals. Each bond orbital is a linear combination of one source orbital $\chi_i \in B^o$ and one sink orbital $\chi_i \in B^x$:

$$\phi_s = \phi_{ij} = (\chi_i + \chi_j)/\sqrt{2}, \quad \text{nonexcited BO}, \tag{1a}$$

$$\phi^*_s = \phi^*_{ij} = (\chi_i - \chi_j)/\sqrt{2}, \quad \text{excited BO},$$

where

$$\chi_i \in B^o \quad \text{and} \quad \chi_j \in B^x. \tag{1b}$$

The set of all RRS's, $S \in \mathscr{R}(n)$, spans the space X_n, though it is overdetermined.[9] Note that the definition of the set $\mathscr{R}(n)$ depends on the partition of the set B on subsets B^o and B^x. There is a one-to-one correspondence between these partitions (up to the exchange $B^o \leftrightarrow B^x$) and different sets $\mathscr{R}(n)$ of RRS's. All these sets are isomorphic, and they all span the CI space X_n (see Ref. 9). In what follows we will assume that the set $\mathscr{R}(n)$ is fixed with an appropriate choice of sets B^o and B^x.

In the molecular orbital resonance theory (MORT) approach the wave function Ψ is assumed to be the linear combination of regular resonance structures.[8–10] One is thus faced with the problem to calculate overlap integrals and matrix elements of different operators between RRS's. In Ref. 8 overlaps and matrix elements of one-particle operators between different resonance structures were derived. In the Appendix the rules for the calculation of matrix elements of two-particle symmetric operators are given. The derivation of these rules is quite time-consuming, and it will be omitted here. They are derived using a technique similar to the one used in the one-particle case, and the interested reader may refer to Ref. 8. For the sake of completeness, overlaps and matrix elements of one-particle operators are also included in the Appendix. Matrix elements of the following operators are considered:

$$\hat{A}_{ij} = \eta_i^+ \eta_j + \eta_j^+ \eta_i,$$
$$\hat{A}_{ij,kl} = \eta_i^+ \eta_j^+ \eta_k \eta_l + \eta_l^+ \eta_k^+ \eta_j \eta_i, \tag{2}$$

where the η_i^+ are creation operators associated with orbitals χ_i, while the η_i are annihilation operators associated with these orbitals. These operators satisfy the usual anticommutation relations

$$[\eta_i, \eta_j]_+ = [\eta_i^+, \eta_j^+]_+ = 0, \quad [\eta_i^+, \eta_j]_+ = \delta_{ij}. \tag{3}$$

Each symmetric one- and two-particle operator can be expressed as a linear combination of operators $\hat{A}_{ij}$ and $\hat{A}_{ij,kl}$. In particular, each symmetric Hamiltonian can be expressed in such a way.

Structures $S \in \mathscr{R}(n)$ are not mutually orthogonal (see Lemma A2), and the rules for the derivation of overlaps $S_{ab} = \langle S_a | S_b \rangle$ and matrix elements of different operators between MORT resonance structures resemble to some extent the rules for the derivation of the matrix elements between VB resonance structures. In both cases one forms the superposition between different structures, and these superpositions consist of disjunct even cycles. There are, however, some important differences: in the VB approach these superpositions, known as Rumer diagrams,[11] contain only oriented bonds. Each oriented bond corresponds to two paired electrons, one electron with the spin-α and another with the spin-β. In the MORT approach the superposition G_{ab} between regular structures S_a and S_b can contain oriented as well as nonoriented bonds. Each oriented bond corresponds to one electron occupying excited BO, while each nonoriented bond corresponds to one electron occupying nonexcited BO. There is hence a difference in the meaning of these superpositions, and there is more "structure" in the MORT superpositions.

In the MORT approach the notion of "active" and "passive" cycle is quite important in deriving overlaps and matrix elements between different RRS's. Each cycle $c_\mu \in G_{ab}$ is characterized by two numbers, n_μ and m_μ, where $(2n_\mu)$ is the number of bonds in the cycle c_μ, while m_μ is the number of oriented bonds in this cycle. Cycle $c_\mu \in G_{ab}$ is "active" if $(n_\mu + m_\mu)$ is odd and "passive" otherwise. This distinction of active and passive cycles is the essential characteristic of the MORT approach. For example, overlap S_{ab} between structures S_a and S_b vanishes if the superposition G_{ab} of these two structures contains at least one passive cycle. Due to this property the set $\mathscr{R}(n)$ of all RRS's can be partitioned into subsets $\mathscr{R}^+(n)$ and $\mathscr{R}^-(n)$ containing mutually orthogonal structures.[9] The space X_n^+ spanned by "positive" structures $S \in \mathscr{R}^+(n)$ is hence orthogonal to the space X_n^- spanned by "negative" structures $S \in \mathscr{R}^-(n)$. The splitting of the space X_n in complementary subspaces X_n^+ and X_n^- depends on the partition of the set B on subsets B^o and B^x, and there is again a one-to-one correspondence between this splitting (up to the exchange $X_n^+ \leftrightarrow X_n^-$) and different partitions (up to the exchange $B^o \leftrightarrow B^x$).[9] Expressions for matrix elements of one- and two-particle operators between different RRS's contain again the notion of active and passive cycles, similar to the expressions for overlaps.

This leads to particular regularities of these matrix elements with respect to complementary subspaces. As a result, many general rules concerning matrix elements between arbitrary states contained either in the same subspace (either X_n^+ or X_n^-), or in different subspaces can be derived. For example, the generalization of the well-known Hückel $(4m+2)$ rule,[8,12] the derivation of the pairing theorem in the case of the PPP Hamiltonian,[9] etc., can be obtained. In this paper we will deal with some other general properties which can be deduced from the particular form of one- and two-particle matrix elements.

Besides the notion of active and passive cycles the rules for the derivation of matrix elements between different RRS's contain some additional concepts. Thus one-particle operators can be "internal" or "external" with respect to the superposition G_{ab}. An operator $\hat{A}_{ij}$ is internal with respect to G_{ab} if vertices (i) and (j) are contained in the same cycle $c_\mu \in G_{ab}$, and it is external otherwise. Similarly, two-particle operators can be "connected" or "disconnected" with respect to the superposition G_{ab}. Roughly, a two-particle operator $\hat{A}_{ij,kl}$ is "connected" with respect to the superposition G_{ab} if the arrows associated with this operator together with some segments on G_{ab} form closed loop(s). Otherwise it is disconnected (see the Appendix). These notions are defined relative to the superposition G_{ab}, and one and the same one-particle operator can be internal with respect to one pair of RRS's, and external with respect to another. Similarly, one and the same two-particle operator can be connected with respect to one pair of RRS's, and disconnected with respect to another. The importance of these notions is revealed by Lemmas A4 and A7. Thus, if the operator $\hat{A}_{ij}$ is external with respect to the superposition G_{ab}, then matrix element $\langle S_a|\hat{A}_{ij}|S_b\rangle$ vanishes (Lemma A4). Similarly, if the operator $\hat{A}_{ij,kl}$ is disconnected with respect to G_{ab}, then matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ vanishes (Lemma A7). It is hence sufficient to derive matrix elements only for one-particle internal operators and for two-particle connected operators. Operators $\hat{A}_{ij}$ and $\hat{A}_{ij,kl}$ are further distinguished depending on the number of vertices $(i,j,k,$ and $l)$ which coincide. Operators $\hat{A}_{ii}$ are vertex operators, operators $\hat{A}_{ij}(i\neq j)$ are bond operators, operators $\hat{A}_{ij,ij} = -\hat{A}_{ij,ji}$ are vertex–vertex operators, operators $\hat{A}_{ij,jk} = -\hat{A}_{ij,kj}(i\neq k)$ are bond–vertex operators, while operators $\hat{A}_{ij,kl}$, where all indices $i,j,k,$ and l are mutually different, are bond–bond operators. These definitions are directly suggested by diagrammatic representations of different operators [see Figs. 2(a) and 3(c)] and they are also useful in the evaluation of matrix elements. All additional concepts needed for the derivation of matrix elements are defined in the Appendix.

III. SYMMETRY PROPERTIES OF MATRIX ELEMENTS $\langle S_a|\hat{A}_{ij}|S_b\rangle$ AND $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$: THE SPLITTING THEOREM

From various expressions for the derivation of matrix elements given in the Appendix, many symmetry properties follow. Thus, according to the Lemma A5, matrix element $A^s_{ab} = \langle S_a|\hat{A}_{kl}|S_b\rangle$ of internal bond operator $\hat{A}_{kl}$ is proportional to the overlap S_{ab} if vertices (k) and (l) are of opposite parity. However, the overlap S_{ab} between structures $S_a \in X_n^+$ and $S_b \in X_n^-$ vanishes. Moreover, since matrix element A^s_{ab} of an external operator vanishes (Lemma A4), it follows that $A^s_{ab} = 0$ whenever $S_a \in X_n^+$ and $S_b \in X_n^-$. Hence $\langle \Psi^+|\hat{A}_{kl}|\Psi^-\rangle = 0$ for arbitrary states $\Psi^+ \in X_n^+$ and $\Psi^- \in X_n^-$. Similarly, in the case when vertices (k) and (l) are of the same parity, matrix element $A^s_{ab} = \langle S_a|\hat{A}_{kl}|S_b\rangle$ vanishes if the cycle $c_\mu \in G_{ab}$ supporting the operator $\hat{A}_{kl}$ is active and/or at least one among other $(\rho-1)$ cycles contained in the superposition G_{ab} is passive (Lemma A6). However, the superposition G_{ab} of structures S_a and S_b of the same parity contains an even number of passive cycles. If now G_{ab} contains no passive cycle, then the cycle $c_\mu \in G_{ab}$ is also active, and hence $A^s_{ab} = 0$. If however G_{ab} contains two or more passive cycles, then at least one among remaining $(\rho-1)$ cycles should be passive, and hence again $A^s_{ab} = 0$. Thus $A^s_{ab} = 0$ whenever S_a and S_b are of the same parity. It follows that $\langle \Psi|\hat{A}_{kl}|\Psi'\rangle = 0$ for arbitrary states $\Psi,\Psi' \in X_n^+$ and arbitrary states $\Psi,\Psi' \in X_n^-$. Matrix elements of a bond operator $\hat{A}_{kl}$ connecting vertices of the same parity vanish between all states contained in the same subspace, either X_n^+ or X_n^-.

Applying similar arguments to all other matrix elements one can derive the following.

Theorem 1(a):

(1) Let the states Ψ_a and Ψ_b be contained in the same subspace, either both in the subspace X_n^+, or both in the subspace X_n^-. In this case matrix elements of all the following operators vanish between these states: (a) each bond operator $\hat{A}_{kl}$ connecting vertices (orbitals) of the same parity; and (b) each bond–bond operator $\hat{A}_{ij,kl}$ containing an odd number of source(sink) vertices.

(2) Let the states Ψ_a and Ψ_b be contained in different subspaces, one in the subspace X_n^+ and the other one in the subspace X_n^-. In this case matrix elements of all the following operators vanish between these states: (a) each bond operator $\hat{A}_{kl}$ connecting vertices (orbitals) of different parity; and (b) each bond–bond operator $\hat{A}_{ij,kl}$ containing an even number of source(sink) vertices.

According to the above theorem all bond operators $\hat{A}_{kl}$ and all bond–bond operators $\hat{A}_{ij,kl}$ fall into one of the two classes: they connect either only the states in the same subspace, or they connect only the states in different subspaces. Concerning vertex operators $\hat{A}_{kk}$, bond–vertex operators $\hat{A}_{ij,jl}$, and vertex–vertex operators $\hat{A}_{ij,ji}$ one can demonstrate using Lemmas A3, A11, and A12 that all these operators can have nonvanishing matrix elements between the states in the same subspace, as well as between the states in different subspaces. One can, however, define "reduced" vertex operators $\hat{R}_{kk}$, "reduced" bond–vertex operators $\hat{R}_{ij,jl}(i\neq j\neq l)$, and "reduced" vertex–vertex operators $\hat{R}_{ij,ji}(i\neq j)$:

$$\begin{aligned}\hat{R}_{kk} &= \hat{A}_{kk} - 1, \quad \hat{R}_{ij,jl} = 2\hat{A}_{ij,jl} - \hat{A}_{il},\\ \hat{R}_{ij,ji} &= -\hat{R}_{ij,ij} = 2\hat{A}_{ij,ji} - \hat{A}_{ii} - \hat{A}_{jj} + 1\\ &= 2\hat{A}_{ij,ji} - \hat{R}_{ii} - \hat{R}_{jj} - 1. \qquad (4)\end{aligned}$$

From Lemmas A2 and A3 it follows that matrix element $R^k_{ab} = \langle \Psi_a|\hat{R}_{kk}|\Psi_b\rangle$ vanishes whenever states Ψ_a and Ψ_b are contained in the same subspace. Concerning operators $\hat{R}_{ij,jl}$ and $\hat{R}_{ij,ji}$ one finds, using the anticommutation algebra

of creation and annihilation operators η_i^+ and η_i,

$$2\hat{A}_{ij,jl} = \hat{A}_{jj}\hat{A}_{il} \quad (i,l \neq j), \tag{5}$$

and hence

$$\begin{aligned} \hat{R}_{ij,jl} &= (\hat{A}_{jj} - 1)\hat{A}_{il} = \hat{R}_{jj}\hat{A}_{il} \quad (i \neq j \neq l), \\ \hat{R}_{ij,ji} &= (\hat{A}_{ii} - 1)(\hat{A}_{jj} - 1) = \hat{R}_{ii}\hat{R}_{jj} \quad (i \neq j). \end{aligned} \tag{6}$$

Assume now $\Psi^+ \in X_n^+$. If vertices (i) and (l) are of the opposite parity, then, according to Theorem 1a, $\langle \Psi^- | \hat{A}_{il} | \Psi^+ \rangle = 0$ for an arbitrary state $\Psi^- \in X_n^-$. Hence $\hat{A}_{il}\Psi^+ = \Psi' \in X_n^+$. Similarly one derives $(\hat{A}_{jj} - 1)\Psi' = \Psi'' \in X_n^-$. Hence $\langle \Psi_a | \hat{R}_{ij,jl} | \Psi_b \rangle = 0$ if vertices (i) and (l) are of different parity and states Ψ_a and Ψ_b are contained in the same subspace. Analogously all other cases can be treated. One thus obtains the following theorem.

Theorem 1b:

(1) Let the states Ψ_a and Ψ_b be contained in the same subspace, either both in the subspace X_n^+, or both in the subspace X_n^-. In this case (a) matrix elements $\langle \Psi_a | \hat{R}_{kk} | \Psi_b \rangle$ of each vertex operator $\hat{R}_{kk}$ vanish, and (b) matrix elements $\langle \Psi_a | \hat{R}_{ij,jl} | \Psi_b \rangle$ of each bond–vertex operator $\hat{R}_{ij,jl}$ such that (i) and (l) are vertices of opposite parity vanish.

(2) Let the states Ψ_a and Ψ_b be contained in different subspaces, one in the subspace X_n^+ and the other one in the subspace X_n^-. In this case (a) matrix elements $\langle \Psi_a | \hat{R}_{ij,ji} | \Psi_b \rangle$ of each vertex–vertex operator $\hat{R}_{ij,ji}$ vanish, and (b) matrix elements $\langle \Psi_a | \hat{R}_{ij,jl} | \Psi_b \rangle$ of each bond–vertex operator $\hat{R}_{ij,jl}$, such that (i) and (l) are vertices of the same parity, vanish.

Theorem 1b completes Theorem 1a. We jointly refer to these two theorems as the "splitting theorem." Anticipating results of the following two sections, we call operators having vanishing matrix elements between different subspaces "alternant," and we call operators having vanishing matrix elements within the same subspace "antialternant." According to the splitting theorem, bond operators $\hat{A}_{ij}$, where vertices (i) and (j) are of different parity, vertex–vertex operators $\hat{R}_{ij,ji}$, bond–vertex operators $\hat{R}_{ij,jl}$, where vertices (i) and (l) are of the same parity, and bond–bond operators $\hat{A}_{ij,kl}$ containing an even number of source (sink) vertices are alternant. By definition, a unit operator I is an alternant operator as well. Vertex operators $\hat{R}_{kk}$, bond operators $\hat{A}_{ij}$, where vertices (i) and (j) are of the same parity, bond–vertex operators $\hat{R}_{ij,jl}$, where vertices (i) and (l) are of opposite parity, and bond–bond operators $\hat{A}_{ij,kl}$ containing an odd number of source (sink) operators are antialternant. We jointly call all the above operators "reduced" operators.

According to the relations (4) vertex operators $\hat{A}_{kk}$, vertex–vertex operators $\hat{A}_{ij,ji}$, and bond–vertex operators $\hat{A}_{ij,jl}$ which have no definite symmetry properties with respect to subspaces X_n^+ and X_n^-, can be expressed as a linear combination of reduced operators $\hat{R}_{kk}$, $\hat{R}_{ij,ji}$, I, and $\hat{R}_{ij,jl}$ having such properties. Operators $\hat{A}_{kk}$, $\hat{A}_{ij,ji}$, and $\hat{A}_{ij,jl}$ together with a unit operator I, bond operators $\hat{A}_{ij}$, and bond–bond operators $\hat{A}_{ij,kl}$ are linearly independent, and an arbitrary symmetric one- and two-particle operator can be expressed as their linear combination. It follows that each one- and two-particle symmetric operator can be expressed in a unique way as a linear combination of reduced operators $\hat{R}_{kk},\ldots,\hat{A}_{ij,kl}$ having definite symmetry properties with respect to subspaces X_n^+ and X_n^-.

Note that the notion of alternant and antialternant operators $\hat{R}_{kk},\ldots,\hat{A}_{ij,kl}$ depends on the partition on source and sink vertices. However, whatever the partition, each of these operators has definite symmetry properties, merely the role of alternant and antialternant operators is permuted. In other words, the set of all reduced operators is invariant with respect to different possible splittings of the space X_n on complementary subspaces.[13] This is a rather amazing property which is not shared by arbitrary alternant and antialternant operators. For example, if reduced operators $\hat{A}_{ij}$ and $\hat{A}_{kl}$ are alternant with respect to some pair of complementary spaces X_n^+ and X_n^-, then the operator $\hat{O} = \hat{A}_{ij} + \hat{A}_{kl}$ is also alternant with respect to these spaces. In other words, matrix elements of the operator $\hat{O}$ vanish between the states $\Psi^+ \in X_n^+$ and $\Psi^- \in X_n^-$. However, another partition on source and sink vertices may define operator $\hat{A}_{ij}$ to be alternant and operator $\hat{A}_{kl}$ to be antialternant. Hence operator $\hat{O}$ has no definite symmetry properties with respect to new complementary spaces $X_n'^+$ and $X_n'^-$. In this respect reduced operators are unique. Moreover, whatever the partition, reduced operators $\hat{R}_{kk}$ are always antialternant, while reduced operators $\hat{R}_{ij,ij}$ are always alternant.

For the sake of completeness we give in the Appendix also matrix elements of reduced operators $\hat{R}_{kk}$, $\hat{R}_{ij,ji}$, and $\hat{R}_{ij,jl}$. These matrix elements can be derived directly using the technique in Ref. 8. It is however much easier to use relations (4) in conjunction with matrix elements for operators $\hat{A}_{kk}$, $\hat{A}_{ij}$, $\hat{A}_{ij,ji}$, and $\hat{A}_{ij,jl}$. It can be easily shown that the expressions for these matrix elements are in accord with the splitting theorem.

The splitting theorem was derived in Ref. 9 for operators involved in the PPP-type Hamiltonian. The present formulation is complete, as far as symmetric one- and two-particle operators are concerned.

IV. PROPERTIES OF STATES $\Psi^+ \in X_n^+$ AND $\Psi^- \in X_n^-$

The splitting theorem implies some special properties of states $\Psi^+ \in X_n^+$ and $\Psi^- \in X_n^-$. One-particle operators $\hat{A}_{ij}$ are up to the factor 2 identical to bond-order operators $\hat{p}_{ij}$, i.e.,

$$\hat{p}_{ij} = (\eta_i^+ \eta_j + \eta_j^+ \eta_i)/2 = \hat{A}_{ij}/2, \tag{7}$$

and the splitting theorem implies the following.

Corollary 1: Let $\hat{q}_i = \hat{p}_{ii} = \eta_i^+ \eta_i$ be a charge density operator. Then

$$\langle \Psi | \hat{q}_i | \Psi \rangle = \tfrac{1}{2}, \tag{8}$$

for each state $\Psi \in X_n^{\pm}$. In other words, each such state has a uniform charge density of $\frac{1}{2}$ over all vertices (orbitals) (i). In particular, consider the following model.

Model A: Let the set $B = \{\chi_i | i = 1,\ldots,2n\}$ contain $2n$ spin-orbitals χ_i such that

$$\chi_i = w_i\alpha, \quad \chi_{n+i} = \overline{\chi}_i = w_i\beta \quad (i = 1,\ldots,n), \tag{9a}$$

where the w_i are orthonormalized atomic orbitals, while α and β are spin-α and spin-β states, respectively. Annihilation operators $\eta_i (i = 1,\ldots,2n)$ can be now written in the form

$$a_i = \eta_i, \quad b_i = \eta_{n+i} \quad (i = 1,\ldots,n), \tag{9b}$$

where a_i are spin-α annihilation operators associated with the spin-α orbitals χ_i, while b_i are spin-β annihilation operators associated with the spin-β orbitals $\overline{\chi}_i$. In terms of these operators one can write charge-density and bond-order operators

$$\hat{q}_i^{\alpha} = a_i^{+} a_i, \quad \hat{q}_i^{\beta} = b_i^{+} b_i, \quad \hat{Q}_i = \hat{q}_i^{\alpha} + \hat{q}_i^{\beta},$$
$$\hat{p}_{ij}^{\alpha} = (a_i^{+} a_j + a_j^{+} a_i)/2, \quad \hat{P}_{ij} = \hat{p}_{ij}^{\alpha} + \hat{p}_{ij}^{\beta},$$
$$\hat{p}_{ij}^{\beta} = (b_i^{+} b_j + b_j^{+} b_i)/2 \quad (i \neq j)$$
$$\hat{p}_{ij}^{\alpha\beta} = (a_i^{+} b_j + b_j^{+} a_i)/2, \tag{10}$$

where the $\hat{q}_i^{\alpha}$ are spin-α charge density operators, the $\hat{q}_i^{\beta}$ are spin-β charge-density operators, etc. Operators $\hat{p}_{ij}^{\alpha\beta}$ are "cross" bond-order operators connecting spin-α and spin-β orbitals. Partition now the set B onto subsets B^o and B^x in such a way that spin-α and spin-β orbitals χ_i and $\overline{\chi}_i$ associated with the same atomic orbital w_i have opposite parity. Such a partition is always possible.[14] By convention, atomic orbital w_i is considered to have the same parity as a spin-α atomic orbital χ_i.

In conjunction with the above model Corollary 1 now implies

$$\langle \Psi | \hat{q}_i^{\alpha} | \Psi \rangle = \langle \Psi | \hat{q}_i^{\beta} | \Psi \rangle = \tfrac{1}{2},$$
$$\langle \Psi | \hat{Q}_i | \Psi \rangle = 1 \quad (i = 1,\ldots,n), \tag{8'}$$

for each state $\Psi \in X_n^{\pm}$. Spin-α and spin-β densities are uniformly $\frac{1}{2}$, and the total charge is uniformly 1 over all atomic orbitals w_i. Uniform charge-density distribution is, however, the property of eigenstates corresponding to neutral alternant hydrocarbons.[5-7] The states $\Psi \in X_n^{\pm}$ are thus likely candidates to be such eigenstates. However, according to Corollary 1, this property should be attributed to entire spaces X_n^{+} and X_n^{-}, rather than to particular eigenstates. Moreover, eigenstates of nonalternant systems are known to be polarized. Such eigenstates should hence contain nonvanishing components in both subspaces, X_n^{+} and X_n^{-}. Charge polarization is thus recognized to be due to the interference between complementary spaces X_n^{+} and X_n^{-}.

Another consequence of the splitting theorem is the following.

Corollary 2: Let $\hat{p}_{ij}$ be a bond-order operator and let vertices (orbitals) (i) and (j) be of the same parity. Then

$$\langle \Psi | \hat{p}_{ij} | \Psi \rangle = 0, \tag{11a}$$

for each state $\Psi \in X_n^{\pm}$. In other words, bond-orders between vertices (orbitals) of the same parity vanish for all states $\Psi \in X_n^{\pm}$.

In particular, in the case of neutral alternant hydrocarbons (AH) one can use Model A above, and atomic orbitals w_i can be partitioned on sink and source in accordance with the usual partition on starred and nonstarred atoms. Corollary 2 now implies

$$\langle \Psi | \hat{p}_{ij}^{\alpha} | \Psi \rangle = \langle \Psi | \hat{p}_{ij}^{\beta} | \Psi \rangle = \langle \Psi | \hat{P}_{ij} | \Psi \rangle = 0, \tag{11b}$$

for each state $\Psi \in X_n^{\pm}$, provided atoms (i) and (j) are of the same parity. In addition one obtains

$$\langle \Psi | \hat{p}_{ij}^{\alpha\beta} | \Psi \rangle = 0, \tag{11c}$$

for each state $\Psi \in X_n^{\pm}$, provided atoms (i) and (j) are of opposite parity. Relations (11b) express again the well-known property of eigenstates of neutral AH systems, the vanishing of bond-orders between atoms of the same parity.[5] This property should be also attributed to the entire spaces X_n^{+} and X_n^{-}, rather than to particular eigenstates.[9]

As suggested by the above two corollaries, we are justified to call each state $\Psi \in X_n^{\pm}$ "alternantlike".[9] We have shown here that nonalternant eigenstates are not contained in spaces X_n^{+} and X_n^{-}, but should rather have nonvanishing components in both spaces. Further, we have shown that alternant eigenstates *may be* contained in spaces X_n^{+} and X_n^{-}, since all the states in these spaces have the necessary properties. In the following section we will show that alternant eigenstates *are* contained in these spaces.

Uniform charge density distribution and vanishing bond-orders between atoms of the same parity are well-known properties of neutral AH systems. The splitting theorem however implies some additional properties of alternantlike states $\Psi \in X_n^{\pm}$.

Corollary 3: Let $\hat{R}_{ij,jl}$ be a bond–vertex operator such that vertices (i) and (l) are of the opposite parity. Then

$$\langle \Psi | \hat{R}_{ij,jl} | \Psi \rangle = 0, \tag{12}$$

for each state $\Psi \in X_n^{\pm}$.

Corollary 4: Let $\hat{A}_{ij,kl}$ be a bond–bond operator containing an odd number of source (sink) vertices. Then

$$\langle \Psi | \hat{A}_{ij,kl} | \Psi \rangle = 0, \tag{13}$$

for each state $\Psi \in X_n^{\pm}$.

Provided the state Ψ is real, the quantities $\gamma(i,j) = \langle \Psi | \hat{A}_{ij} | \Psi \rangle / 2$ and $\Gamma(ij;kl) = \langle \Psi | \hat{A}_{ij,lk} | \Psi \rangle / 4$ are one- and two-particle density matrices, respectively. Since eigenstates of symmetrical operators are real, Corollaries 1 and 2 express the properties of one-particle density matrices of alternantlike states $\Psi \in X_n^{\pm}$, while Corollaries 3 and 4 refer essentially to the properties of two-particle density matrices of these states. Thus Corollary 3 establishes the connection between some matrix elements of one-particle and some matrix elements of two-particle density matrices. Equation (12) is namely equivalent to

$$\Gamma(ij;lj) = \langle \Psi | \hat{A}_{ij,jl} | \Psi \rangle / 4$$
$$= \langle \Psi | \hat{A}_{il} | \Psi \rangle / 8 = \gamma(il)/4 \quad (j \neq i,l), \tag{12'}$$

where (i) and (l) are of different parity and the state Ψ is an arbitrary alternantlike state. Matrix elements $\Gamma(ij;lj)$ of a two-particle density matrix are thus uniquely determined by the matrix elements $\gamma(i,l)$ of the one-particle density matrix, whenever (i) and (l) are of different parity. Note the role played by one-particle density matrix $\gamma(i,l)$: if (i) and (l) are of the same parity, then $\gamma(i,l) = 0$ $(i \neq l)$ and $\gamma(i,i) = \frac{1}{2}$ (Corollaries 1 and 2). If however (i) and (l) are of the opposite parity, then $\gamma(i,l)$ determines certain matrix elements of the two-particle density matrix (Corollary 3). In addition, by corollary 4 matrix elements $\Gamma(ij;kl)$ $(i \neq j \neq k \neq l)$ of the two-particle density matrix are required to be zero whenever there is an odd number of source (sink) vertices. These are quite severe requirements on two-particle density matrices associated with alternantlike states. They are properties of entire spaces X_n^{+} and X_n^{-}, rather than the properties of particular eigenstates. Anticipating the result of the following section,

we see that these properties are also shared by eigenstates of alternant Hamiltonians.

V. ALTERNANT HAMILTONIANS

According to the splitting theorem all bond operators $\hat{A}_{ij}$ with vertices (i) and (j) of different parity, all vertex–vertex operators $\hat{R}_{ij,ij}$, all bond–vertex operators $\hat{R}_{ij,jl}$, where vertices (i) and (l) are of the same parity, and all bond–bond operators $\hat{A}_{ij,kl}$, with an even number of source (sink) vertices, have vanishing matrix elements between states contained in different subspaces. Each operator which can be expressed as an arbitrary function of these operators forms diagonal blocks in the base of the space X_n with respect to its subspaces X_n^+ and X_n^-. Eigenfunctions of such an operator are hence alternantlike,[15] i.e., they are contained in subspaces X_n^+ and X_n^- and they possess all the properties discussed in the previous section. In particular, each Hamiltonian which can be expressed in terms of the above operators possesses alternantlike eigenfunctions. In the Appendix we prove the following.

Theorem 2: Each symmetric Hamiltonian possessing the complete set of alternantlike eigenstates can be expressed in the form

$$H=\sum_{ij}^{2n-} a_{ij}\hat{A}_{ij}+\sum_{ij}^{2n} b_{ij}\hat{R}_{ij,ji} + \sum_{j}^{2n}\sum_{i,l}^{2n+} c_{j,il}\hat{R}_{ij,jl} + \sum_{i,j,k,l}^{2n+} d_{ij,kl}\hat{A}_{ij,kl}+\text{const}, \tag{14}$$

where $a_{ij}, b_{ij}, \ldots, d_{ij,kl}$ are arbitrary real parameters, and the following convention concerning different summations is used: double summations Σ_{ij}^+ and Σ_{ij}^- are performed over vertices (i) and (j) of the same and of the opposite parity, respectively. Quadripole summations Σ_{ijkl}^+ and Σ_{ijkl}^- are performed over distinct vertices (i), (j), (k), and (l). In the former summation an even number of these vertices is source (sink), while in the latter summation an odd number of these vertices is source (sink).

In other words, the most general form of the symmetric Hamiltonian possessing the complete set of alternantlike eigenstates is a linear combination of reduced alternant operators.

To be more specific, assume that we deal with a neutral AH system. Such a system can be treated within the model A above, where w_i are $2p_z$ atomic orbitals at different carbon atoms. The first term in the expression (14) is now recognized to be essentially the Hückel Hamiltonian. Coulson and Rushbrooke demonstrated that in the HMO approach neutral AH possess uniform charge density distribution and vanishing bond-orders between all atoms of the same parity.[5] Consider now the PPP approach. The PPP Hamiltonian of an alternant hydrocarbon system can be written in the form

$$H_p = K\sum_i \hat{Q}_i + \sum_{ij}\beta_{ij}\hat{P}_{ij} + \sum_{i<j}\gamma_{ij}\left[(\hat{Q}_i-1)(\hat{Q}_j-1)-1\right] + \sum_i \gamma_{ii}\left[\left(\hat{q}_i^\alpha-\frac{1}{2}\right)\left(\hat{q}_i^\beta-\frac{1}{2}\right)-\frac{1}{4}\right], \tag{15}$$

where K is the effective potential energy of a π-electron which is assumed to be the same on each carbon atom, the β_{ij} are resonance integrals, and the γ_{ij} are electron–electron repulsion integrals.[6] In the second term the summation is performed only over carbon atoms of the opposite parity. The first term of the Hamiltonian H_p contains the operator $\hat{N}=\Sigma_i\hat{Q}_i$, which is the operator of the total number of electrons. This operator is constant over the space X_n. The second term contains bond operators $\hat{A}_{ij}$ connecting vertices (i) and (j) of different parity, and using relations (6) one finds that the third and the fourth term contain only vertex–vertex operators $\hat{R}_{ij,ij}$. The PPP Hamiltonian H_p can hence be represented as a sum of the first two terms in the expression (14) plus a constant. McLachlan demonstrated that eigenstates of the Hamiltonian H_p corresponding to a neutral AH system possess uniform charge-density distribution and vanishing bond-orders between all atoms of the same parity.[6] Hamiltonian (14) is however much more general than either the Hückel or the PPP Hamiltonian, and according to Theorem 2 it is the most general symmetric Hamiltonian possessing the complete set of alternantlike eigenstates.

One can now ask the following question: Given a symmetric Hamiltonian H in its usual representation, what are the conditions on this Hamiltonian in order that it has alternantlike eigenstates? Such Hamiltonians are usually written in the form

$$H=\sum_{ij} h_{ij}\eta_i^+\eta_j+\sum_{i<j}\sum_{k<l} V_{ij,kl}\eta_i^+\eta_j^+\eta_l\eta_k, \tag{16a}$$

where h_{ij} and $V_{ij,kl}$ are integrals

$$h_{ij}=\langle i|h\,|j\rangle=\int\chi_i(1)h(1)\chi_j(1)d\tau_1,$$

$$V_{ij,kl}=\langle ij\|kl\rangle=\int\chi_i(1)\chi_j(2)V(1,2)\times[\chi_k(1)\chi_l(2)-\chi_l(1)\chi_k(2)]d\tau_1\,d\tau_2. \tag{16b}$$

Orbitals χ_i and one- and two-particle operators $h(1)$ and $V(1,2)$ are real. Hence the following symmetry relations hold:

$$h_{ij}=h_{ji},\quad V_{ij,kl}=-V_{ij,lk}=-V_{ji,kl}=V_{kl,ij}. \tag{16c}$$

One can rewrite Hamiltonian (16a) as a linear combination of alternant and antialternant reduced operators,

$$H=\frac{1}{2}\Bigg[\sum_i\Big(h_{ii}+\frac{1}{4}\sum_j V_{ij,ij}\Big)+\sum_i\Big(h_{ii}+\frac{1}{2}\sum_j V_{ij,ij}\Big)\hat{R}_{ii}+\sum_{i\neq j}\Big(h_{ij}+\frac{1}{2}\sum_l V_{il,jl}\Big)\hat{A}_{ij}+\frac{1}{2}\sum_{i<j}V_{ij,ij}\hat{R}_{ij,ji}+\frac{1}{2}\sum_i\sum_{j\neq l}V_{ij,il}\hat{R}_{ij,li}+\sum_{i<j}\sum_{k<l}V_{ij,lk}\hat{A}_{ij,kl}\Bigg], \tag{17}$$

where the last summation is performed over mutually differ-

ent vertices (i), (j), (k), and (l). This can be written in the form

$$H = H_{\mathrm{al}} + H_{\mathrm{nal}}, \tag{18a}$$

where H_{al} contains only alternant, while H_{nal} contains only antialternant reduced operators:

$$\begin{aligned} H_{\mathrm{al}} = \frac{1}{2}\Bigg[& \sum_i \Big(h_{ii} + \frac{1}{4}\sum_j V_{ij,ij}\Big) \\ & + \sum_{ij}{}^{-} \Big(h_{ij} + \frac{1}{2}\sum_l V_{il,jl}\Big)\hat{A}_{ij} \\ & + \frac{1}{2}\sum_{i<j} V_{ij,ij}\hat{R}_{ij,ji} \\ & + \frac{1}{2}\sum_l \sum_{i\neq j}{}^{+} V_{il,lj}\hat{R}_{il,jl} \\ & + \sum_{i<j,k<l}{}^{+} V_{ij,lk}\hat{A}_{ij,kl}\Bigg], \end{aligned} \tag{18b}$$

$$\begin{aligned} H_{\mathrm{nal}} = \frac{1}{2}\Bigg[& \sum_i \Big(h_{ii} + \frac{1}{2}\sum_j V_{ij,ij}\Big)\hat{R}_{ii} \\ & + \sum_{i\neq j}{}^{+} \Big(h_{ij} + \frac{1}{2}\sum_l V_{il,jl}\Big)\hat{A}_{ij} \\ & + \frac{1}{2}\sum_l \sum_{i\neq j}{}^{-} V_{il,lj}\hat{R}_{il,jl} \\ & + \sum_{i<j,k<l}{}^{-} V_{ij,lk}\hat{A}_{ij,kl}\Bigg]. \end{aligned} \tag{18c}$$

If the operator H is to have alternantlike eigenstates, then it should commute with projection operators P_n^+ and P_n^- projecting an arbitrary state $\Psi \in X_n$ on subspaces X_n^+ and X_n^-, respectively. However, according to the splitting theorem $[H_{\mathrm{al}}, P_n^{\pm}] = 0$. Moreover, operator H_{nal} vanishes over spaces X_n^+ and X_n^-, and hence $[H_{\mathrm{nal}}, P_n^{\pm}] = 0$ if and only if H_{nal} vanishes over the entire space X_n. Operator $\hat{N} - n = \frac{1}{2}\Sigma_i \hat{R}_{ii}$, which is a linear combination of antialternant reduced operators, vanishes over the space X_n. In the Appendix we show that this is the only operator with such a property. Hence, we get the following lemma.

Lemma 1: A necessary and sufficient condition for an arbitrary symmetric Hamiltonian (16a) to possess the complete set of alternantlike eigenstates is

$$H_{\mathrm{nal}} = \lambda(\hat{N} - n), \tag{19}$$

where λ is an arbitrary real constant.

Since antialternant operators entering H_{nal} are linearly independent, Lemma 1 is equivalent to the following set of conditions concerning integrals h_{ij} and $V_{ij,kl}$:

$$(1)\quad h_{ii} + \frac{1}{2}\sum_j^{2n} V_{ij,ij} = \mathrm{const}; \tag{20a}$$

$$(2)\quad h_{ij} + \frac{1}{2}\sum_l^{2n} V_{il,jl} = 0, \tag{20b}$$

(i) and (j) are of the same parity $(i \neq j)$;

$$(3)\quad V_{il,jl} = 0, \tag{20c}$$

(i) and (j) are of opposite parity; and

$$(4)\quad V_{ij,kl} = 0, \tag{20d}$$

odd number of source (sink) vertices $(i \neq j \neq k \neq l)$.

Conditions (20) are necessary and sufficient conditions for the symmetric Hamiltonian operator (16a) to have alternantlike eigenstates. In particular, within the model A above, and if $h(1)$ and $V(1,2)$ are spin independent operators, these conditions reduce to

$$(1)\quad \alpha_i + \frac{1}{2}\sum_j^n [2(ii|jj) - (ij|ij)] = \mathrm{const}, \tag{21a}$$

$$(2)\quad \beta_{ij} + \frac{1}{2}\sum_l^n [2(ij|ll) - (il|jl)] = 0, \tag{21b}$$

orbitals w_i and w_j are of the same parity and $i \neq j$;

$$(3)\quad (il|jl) = (ij|ll) = 0, \tag{21c}$$

orbitals w_i and w_j are of the opposite parity and $l \neq i,j$;

$$(4)\quad (ij|kl) = 0, \tag{21d}$$

odd number of starred (nonstarred) orbitals and $i \neq j \neq k \neq l$; where

$$\alpha_i = \int w_i(1) h(1) w_i(1) d1,$$

$$\beta_{ij} = \int w_i(1) h(1) w_j(1) d1, \quad i \neq j, \tag{21e}$$

$$(ij|kl) = \int w_i(1) w_j(1) V(1,2) w_k(2) w_l(2) d1\, d2 \tag{21f}$$

are one- and two-particle integrals over atomic orbitals w_i.

There is a straightforward interpretation of the above conditions. The first condition implies that the effective potential energy of an electron situated on the atomic orbital w_i is constant over all atomic orbitals. The second condition implies that the effective resonance interaction between orbitals of the same parity should be zero. Three center integrals $(il|jl)$ and $(ij|ll)$ can be set arbitrarily, and the Conditions 2 then simply determine resonance integrals β_{ij}. Or, vice versa, resonance integrals β_{ij} can be set arbitrarily, and their effect then can be offset by an appropriate choice of three center integrals. This is the generalization of the usual assumption that in an alternant system the resonance integrals β_{ij} vanish between atoms of the same parity. Conditions (3) imply that all three center integrals $(il|jl)$ and $(ij|ll)$, where w_i and w_j are of the opposite parity, should vanish. Likewise, conditions (4) imply that all four center integrals $(ij|kl)$ containing an odd number of starred (nonstarred) orbitals w_i vanish. Note that there is no condition concerning four center integrals $(ij|kl)$ containing an even number of starred (nonstarred) orbitals w_i. These integrals can hence be arbitrary.

In the PPP approach all three and four center integrals are neglected, resonance integrals β_{ij} are assumed to be zero between atoms of the same parity, and the effective potential energy of an electron is assumed to be constant on each carbon atom. This automatically satisfies Conditions 1–4. The above conditions are, however, much more general, and they apply to any system where the number of orbitals is twice the number of electrons. It should be noted that Koutecký independently obtained some general conditions to be satisfied by a symmetric Hamiltonian operator in order to possess pairing properties.[7] The conditions he obtained are, however, rather implicit, it is not quite clear how general they are, and what should be their interpretation "down to the earth" in terms of one- and two-particle integrals.

Theorem 2 and Lemma 1 answer two complementary questions: (a) What is the most general form of the symmet-

ric Hamiltonian possessing the complete set of alternantlike eigenstates, and (b) given the symmetric Hamiltonian in its usual representation, what are the conditions on this Hamiltonian in order to have the complete set of alternantlike eigenstates. In answering these questions we obtained some additional results which might prove to be quite useful. According to Theorem 2 and Lemma 1, in the case of alternant systems the eigenvalue problem can be formulated in complementary subspaces, rather than in the full CI space X_n. Of course, one is seldom going to consider the full CI space X_n, but one should rather formulate different approximation schemes involving, for example, only some set of energetically low-lying RRS's. However, whatever the approximation used, the above splitting further reduces the dimension of the space considered, and this should lead to considerable computational saving in the case of the actual calculations.[16]

VI. NONALTERNANT HAMILTONIANS

Let us now consider a general symmetric Hamiltonian, not necessarily alternant. From the relations (18) and Lemma 1 it follows that each symmetric Hamiltonian can be written in the form

$$H = H'_{\mathrm{al}} + H'_{\mathrm{nal}}, \tag{22a}$$

where

$$H'_{\mathrm{al}} = H_{\mathrm{al}} + \lambda_0(\hat{N} - n), \quad H'_{\mathrm{nal}} = H_{\mathrm{nal}} - \lambda_0(\hat{N}-n), \tag{22b}$$

and

$$\lambda_0 = \sum_i \left(h_{ii} + \frac{1}{2}\sum_j V_{ij,ij}\right)(2n)^{-1}. \tag{22c}$$

Operator H'_{al} is a proper alternant part of the Hamiltonian H, while operator H'_{nal} is a proper antialternant part of this Hamiltonian. In other words, $H'_{\mathrm{nal}} = 0$ if and only if Hamiltonian H is alternant. It can be shown that in most cases H'_{nal} is "small" with respect to H'_{al}, i.e., the corresponding quantum system is "weakly nonalternant."[16] The antialternant part H'_{nal} of the Hamiltonian H can hence be considered as a small perturbation to the alternant part H'_{al}. This provides a simple and efficient method for the calculation of many quantum mechanical properties. Eigenstates of the operator H'_{al} are namely alternantlike, and hence they possess all the properties discussed in Sec. IV. For example, these eigenstates have uniform charge-density distribution over all vertices (i). Hence, if one is interested in the charge-density distribution of some nonalternant system, this distribution is mainly determined by the perturbation involving H'_{nal}, while the corresponding eigenstate of the unperturbed Hamiltonian H'_{al} is not so important since it contributes uniformly to this distribution. As a result only a very crude knowledge of the unperturbed eigenstate is needed in order to obtain quite reliable charge-density distributions.[16] Similarly, all other properties like bond-orders, etc., can be treated. This approach can be used to evaluate directly the influence of the introduction of heteroatoms in AH systems, of different changes due to the introduction of a nonalternant bond in AH systems, etc.[16] The method is not restricted to π-electron systems. In general, each system which is described using a minimum basic set can be considered to be a perturbed alternant system. Relations (18) and (22) provide the necessary splitting and uniquely define which part of the Hamiltonian is to be considered "alternant" and which "antialternant." Due to the splitting of the corresponding CI space X_n on complementary subspaces X_n^+ and X_n^-, and in conjunction with some simple properties of the unperturbed eigenstates of the operator H'_{al}, a simple and efficient perturbation scheme can be developed. The details will be given elsewhere.[16]

VII. COMMENT ON THE RELATION BETWEEN THE VB AND THE MORT APPROACHES

It is interesting to compare the mathematical structures of the VB and MORT approaches. In the VB theory the overlap S_{ab}^{VB} between the two VB canonical structures equals[17]

$$S_{ab}^{\mathrm{VB}} = 2^{\rho - n}, \tag{23}$$

where ρ is the number of cycles in the Rumer diagram associated with these two structures, while n is the number of pairs of electrons in these structures. This expression is quite similar to the expression (A3) for the overlap between RRS's, with one important difference: according to (23) the overlap between VB structures is always nonzero, while the overlap S_{ab} between RRS's is zero whenever the superposition of these structures contains at least one passive cycle. As a consequence, the set $\mathscr{R}_{\mathrm{VB}}(n)$ of all VB canonical structures cannot be partitioned into two subsets containing mutually orthogonal structures. Hence the corresponding space cannot be partitioned in a natural way into two complementary subspaces. All the nice symmetry properties of different matrix elements are thus lost. There is no simple way to identify alternantlike states, and the properties of the eigenstates of alternant Hamiltonians cannot be derived within the VB approach, at least not without highly artificial and elaborate means. Similarly, the Hückel $(4m + 2)$ rule,[12] which can be derived within the MORT approach in its most general form, is untractable within the VB theory. Cyclobutadiene is thus predicted to be stable, contrary to this rule and contrary to experience. Sure enough, if one does include all the VB structures, this failure of the VB approach can be corrected. However, this is done only by paying the high price of extensive calculations and by losing all the intuitive insight into what is going on. In many other details the MORT approach can be shown to be superior to the VB theory. Formally, the difference between the two approaches is illustrated by relations (23) and (A3). However, conceptually, the difference is due to the different physical pictures behind the two approaches. In the VB approach the resonance structure is built upon the idea of the pairing of two electrons of opposite spins. As a consequence the resulting bonding is mainly due to the two-particle exchange integrals. The MORT approach follows in this respect more closely the MO picture. The bonding is mainly due to the one-particle resonance integrals. Physically, the latter picture should be more correct than the former one. To attribute the bonding to the two-particle integrals is highly doubtful. These integrals represent electron–electron interaction, and this interaction should lead to the destabilization, rather than to the stabili-

zation. In an *ab initio* VB approach this inconsistency is partly corrected due to the nonorthogonality of atomic orbitals. This is, however, still a highly artificial approach and it obscures the real cause of bonding. In short, the MORT approach treats correctly one- and two-particle energy contributions, while the VB theory treats these contributions in an unnatural order. This point has been discussed in more detail in Ref. 10.

VIII. CONCLUSION

The configuration interaction space X_n generated by n electrons moving over $2n$ orbitals χ_i is considered. Given the partition of the set $B = \{\chi_i | i = 1,...,2n\}$ on subsets B^o and B^x containing n orbitals each, there is a unique splitting of the space X_n on complementary subspaces X_n^+ and X_n^-. One- and two-particle symmetric operators $\hat{A}_{ij} = \eta_i^+ \eta_j + \eta_j^+ \eta_i$ and $\hat{A}_{ij,kl} = \eta_i^+ \eta_j^+ \eta_k \eta_l + \eta_l^+ \eta_k^+ \eta_j \eta_i$ are considered. It is shown that operators $\hat{R}_{ii} = \hat{A}_{ii} - 1$, operators $\hat{A}_{ij} (i \neq j)$, operators $\hat{R}_{ij,ij} = 2\hat{A}_{ij,ij} + \hat{A}_{ii} + \hat{A}_{jj} - 1 (i \neq j)$, operators $\hat{R}_{ij,jl} = 2\hat{A}_{ij,jl} - \hat{A}_{il} (i \neq j \neq l)$, and operators $\hat{A}_{ij,kl} (i \neq j \neq k \neq l)$ possess definite symmetry properties with respect to the complementary spaces X_n^+ and X_n^-. In particular, operators $\hat{A}_{ij}$ (i and j are of different parity), operators $\hat{R}_{ij,ij}$, operators $\hat{R}_{ij,jl} (i \neq j \neq l$, i and l are of the same parity) and operators $\hat{A}_{ij,kl} (i \neq j \neq k \neq l$, there is an even number of vertices of each parity) have vanishing matrix elements between complementary spaces X_n^+ and X_n^-. The same property has the unit operator I. Operators $\hat{R}_{ii}$, operators $\hat{A}_{ij} (i \neq j$, i and j are of the same parity), operators $\hat{R}_{ij,jl} (i \neq j \neq l$, i and l are of different parity), and operators $\hat{A}_{ij,kl} (i \neq j \neq k \neq l$, there is an odd number of vertices of each parity), have vanishing matrix elements between all states contained in the same complementary space (Theorem 1). The former operators are called "alternant," and the latter "antialternant." Jointly we call the above alternant and antialternant operators "reduced" operators. Each symmetric one- and two-particle operator can be expressed as a linear combination of these reduced operators. In particular, each symmetric Hamiltonian can be expressed as such a linear combination. There is a complementary role played by the reduced alternant and antialternant operators: reduced alternant operators define properties of states $\Psi \in X_n^+$ and $\Psi \in X_n^-$. More precisely, matrix elements of reduced antialternant operators $\hat{R}_{ii}$ and $\hat{A}_{ij}$ determine each state $\Psi \in X_n^{\pm}$ to have a uniform charge-density distribution and vanishing bond-orders between orbitals of opposite parity. These are well-known properties of the eigenstates corresponding to neutral alternant hydrocarbons, and hence we call states $\Psi \in X_n^{\pm}$ "alternantlike." To be "alternantlike" is thus recognized as a property of the entire spaces X_n^+ and X_n^-, rather than the property of particular eigenstates. In addition, each alternantlike state $\Psi \in X_n^{\pm}$ satisfies $\langle \Psi | \hat{A}_{ij,kl} | \Psi \rangle = 0$ and $\langle \Psi | \hat{R}_{ij,jl} | \Psi \rangle = 2\langle \Psi | \hat{A}_{ij,jl} | \Psi \rangle - \langle \Psi | \hat{A}_{il} | \Psi \rangle = 0$ whenever operators $\hat{A}_{ij,kl}$ and $\hat{R}_{ij,jl}$ are antialternant. The former relation refers to matrix elements of a two-particle density matrix. According to the latter relation, matrix elements $\langle \Psi | \hat{A}_{ij,jl} | \Psi \rangle (j \neq i,l)$ of the two-particle density matrix are uniquely determined by the matrix elements $\langle \Psi | \hat{A}_{il} | \Psi \rangle$ of the one-particle density matrix whenever i and l are of different parity. These are also the properties of entire spaces X_n^+ and X_n^-, i.e., of all the states in these spaces. Alternant operators play another role: each operator which can be expressed as some function of these operators has its eigenstates in complementary subspaces X_n^+ and X_n^-. In the case of a symmetric Hamiltonian the inverse is also true: each symmetric Hamiltonian possessing the complete set of alternantlike eigenstates can be expressed as a linear combination of the reduced alternant operators (Theorem 2). This provides an explicit construction of all such Hamiltonians and also presents a simple test to verify whether a given symmetric Hamiltonian is alternant or not. Eigenstates of alternant Hamiltonians possess all the properties mentioned above. Eigenstates of nonalternant Hamiltonians have nonvanishing components in both spaces, X_n^+ and X_n^-. However, each symmetric Hamiltonian can be in a simple way represented as a sum of an alternant and antialternant part. This splitting, together with all other results obtained, is a base for the development of an efficient perturbation scheme.[16] Accordingly, each quantum chemical system can be considered to be some alternant system with an antialternant perturbation. This perturbation is usually small.[16]

In order to obtain these results, complementary spaces X_n^+ and X_n^- have first to be constructed. These spaces are spanned by "positive" regular resonance structures (RRS's), $S \in \mathscr{R}^+(n)$, and by "negative" RRS's, $S \in \mathscr{R}^-(n)$, respectively. Each RRS is defined to be a normalized determinant containing n mutually disjunct bond orbitals. Each bond orbital is a linear combination of one "source" orbital $\chi_i \in B^o$ and one "sink" orbital $\chi_i \in B^x$. The set $\mathscr{R}(n)$ of all these structures spans the CI space X_n, and the simple superposition criteria is used in order to determine whether the two structures are of the same parity or not. As a next step matrix elements of operators $\hat{A}_{ij}$ and $\hat{A}_{ij,kl}$ between these structures are derived. Analyzing symmetry properties of these matrix elements, all the above conclusions can be obtained. It should be noted that once these conclusions are reached, they are independent of the particular structures used. Structures $S \in \mathscr{R}(n)$ present only a very convenient base with which these properties can be expressed.

There are two restrictions on the generality of the above conclusions. The first restriction is related to the nature of the orbitals $\chi_i \in B$ which are assumed to be orthonormalized. Another restriction is related to the nature of the CI space X_n: only those spaces are considered which can be generated by n electrons moving over exactly $2n$ orbitals. The first restriction is not a serious one. One can always orthonormalize orbitals χ_i using, for example, the symmetric orthogonalization procedure.[18] This merely leads to a redefinition of reduced alternant and reduced antialternant operators. The second restriction seems to be more serious. However, this restriction can be also overcome, and the generalization of the above conclusions to the arbitrary CI space X_n^N generated by n particles moving over N orbitals (n and N are arbitrary) can be obtained.[16]

ACKNOWLEDGMENTS

This work was done mainly during the author's stay at the Quantum Theory Project, University of Florida, Gaines-

ville, Florida. I wish to thank Dr. H. Monkhorst and Dr. M. Zerner for making this stay possible and for many helpful discussions during the preparation of the paper.

Support is acknowledged from the National Scientific Foundation, Grants No. F 6 F 00 6F and No. CHE 8207720.

APPENDIX

1. Configuration interaction space X_n and regular resonance structures

Let $B = \{\chi_i | i = 1,...,2n\}$ be an orthonormalized set of $2n$ orbitals χ_i. The set of all n-particle determinants containing n out of these $2n$ orbitals spans the CI space X_n. In the second quantization formalism these determinants can be written in the form

$$\Psi_{\{i\}} = |i_1, i_2,...,i_{2n}\rangle, \tag{A1}$$

where $i_k = 1$ if the orbital χ_{ik} is occupied, and $i_k = 0$ otherwise. In addition, and since determinants (A1) represent n-particle states,

$$\sum_k^{2n} i_k = n. \tag{A2}$$

a. Regular resonance structures (RRS's)

Partition the set B containing $2n$ orbitals χ_i into subsets B^o and B^x containing n orbitals each. Call each orbital χ_i "source" if $\chi_i \in B^o$, and "sink" if $\chi_i \in B^x$. Form excited and nonexcited bond orbitals (BO),

$$\begin{aligned} &\phi_s = \phi_{ij} = (1/\sqrt{2})(\chi_i + \chi_j), \quad \text{nonexcited BO}, \\ &\phi^*_s = \phi^*_{ij} = (1/\sqrt{2})(\chi_i - \chi_j), \quad \text{excited BO}, \end{aligned} \tag{A3}$$

satisfying the condition

$$\chi_i \in B^o \quad \text{and} \quad \chi_j \in B^x. \tag{A4}$$

Each normalized determinant containing n mutually disjunct BO's (A3) satisfying the condition (A4) is a regular resonance structure (RRS). The set $\mathscr{R}(n)$ of all n-particle RRS's spans the CI space X_n (see Ref. 9). We denote structures $S \in \mathscr{R}(n)$ as

$$S = |s_1, s_2,...,s_n\rangle, \tag{A5}$$

where each s_k $(k = 1,...,n)$ represents either an excited or a nonexcited BO.

b. Superposition of RRS's; active and passive cycles

Orbitals χ_i are graphically represented as vertices. Excited and nonexcited BO's are represented as oriented and nonoriented bonds, respectively. In the case of excited BO, the end vertex of the corresponding oriented bond is associated with the sink orbital. Superposition of RRS's S_a and S_b is a graph G_{ab} which is obtained by superimposing graphical representations of these structures, and it consists of disjunct even cycles $c_\mu \in G_{ab}$ (see Ref. 8). Each cycle $c_\mu \in G_{ab}$ is characterized by two numbers, n_μ and m_μ, where $(2n_\mu)$ is the number of bonds in the cycle c_μ, while m_μ is the number of oriented bonds in this cycle. Cycle c_μ is "passive" if $(n_\mu + m_\mu)$ is even, and "active" if $(n_\mu + m_\mu)$ is odd (Fig. 1).

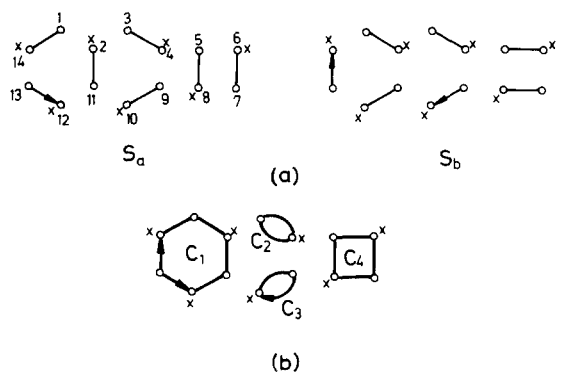

FIG. 1. Graphical representation of regular resonance structures. Vertices (1,3,5,7,9,11,13) represent source orbitals $\chi_i \in B^o$, while vertices (2,4,6,8,10,12,14) represent sink orbitals $\chi_i \in B^x$. (a) An example of two RRS's. Graphically, each structure contains oriented and/or nonoriented bonds connecting one source and one sink vertex. (b) The superposition G_{ab} of structures S_a and S_b consists of disjunct even cycles. Cycles c_1 and c_2 are active, while cycles c_3 and c_4 are passive. Cycles c_2 and c_3 containing only two vertices are called γ-cycles.[8]

c. Complementary spaces X_n^+ and X_n^-

The set $\mathscr{R}(n)$ of all RRS's can be consistently partitioned into subsets $\mathscr{R}^+(n)$ and $\mathscr{R}^-(n)$ using the superposition criteria[8,9]: if the superposition G_{ab} of RRS's S_a and S_b contains an even number of passive cycles, these two structures are contained in the same subset, either in $\mathscr{R}^+(n)$ or in $\mathscr{R}^-(n)$. Otherwise, i.e., if the superposition G_{ab} contains an odd number of passive cycles, structures S_a and S_b are contained in different subsets, one in the subset $\mathscr{R}^+(n)$ and another in the subset $\mathscr{R}^-(n)$.[8,9] The notion of "positive" and "negative" subset is relative.[8,9] From Lemma A2 it follows that the space X_n^+ spanned by all positive structures $S \in \mathscr{R}^+(n)$ is orthogonal to the space X_n^- spanned by all negative structures $S \in \mathscr{R}^-(n)$. Spaces X_n^+ and X_n^- have the same dimension, they are uniquely (up to the exchange $X_n^+ \leftrightarrow X_n^-$) defined by the subset B^o (or B^x) of the set B, and each state $\Psi \in X_n$ can be uniquely written as a sum $\Psi = \Psi^+ + \Psi^-$, where $\Psi^+ \in X_n^+$ and $\Psi^- \in X_n^-$.

d. Normal phase between RRS's

Each permutation of BO's $|s_i) \in S$ and $|s_j) \in S$ in a RRS $S = |...,s_i,...,s_j,...\rangle$ changes the sign of this structure. All overlaps and matrix elements between different RRS's are hence ambiguous up to the sign $(+1)$ or (-1). In order to lift this ambiguity, one has to fix in a consistent way relative phases between different RRS's. According to the definition, each structure $S \in \mathscr{R}(n)$ can be obtained from some fixed nonexcited structure $S_0 \in \mathscr{R}(n)$ in two steps: first one has to make an appropriate permutation of sink vertices, retaining source vertices fixed. Next one has to replace, wherever necessary, nonexcited BO's with excited BO's. The set $\mathscr{R}(n)$ of all RRS's can hence be written in such a way that all ith $(i = 1,...,n)$ BO's in structures $S \in \mathscr{R}(n)$ have the same source orbital in common. In other words, given any two structures $S_a = |s_1,...,s_i,...\rangle \in \mathscr{R}(n)$ and $S_b = |\,p_1,...,p_i,...\rangle \in \mathscr{R}(n)$, BO's $|s_i)$ and $|\,p_i)$ $(i = 1,...,n)$ contain the same source orbital. We will refer to this representation of RRS's as a "normal phase convention." Relative phase between two structures obeying

the normal phase convention is by definition "normal." Normal phase convention fixes relative phases between all RRS's to be normal. All the subsequent expressions for overlaps and matrix elements of different operators between RRS's S_a and S_b are derived under the assumption that the relative phase between these structures is normal.

Besides RRS's, one can define the so-called normal resonance structures (NRS's).[8,9] These structures are defined analogously to the RRS's, but without the condition (A4). The set $\mathcal{N}(n)$ of all NRS's is hence larger than the set $\mathcal{R}(n)$ of all RRS's. Moreover, this set does not depend on the partition of the set B on sink and source subsets. The set $\mathcal{R}(n) \subset \mathcal{N}(n)$ is however fixed by the particular partition of the set B; given another partition of this set, one obtains another set $\mathcal{R}'(n) \subset \mathcal{N}(n)$ of RRS's. Consider now the set $\mathcal{L}(n) = \mathcal{R}(n) \cap \mathcal{R}'(n)$ containing all the structures which are contained in the set $\mathcal{R}(n)$ as well as in the set $\mathcal{R}'(n)$. For the set $\mathcal{L}(n)$ normal phase can be defined either in accord with the convention applied to the set $\mathcal{R}(n)$, or in accord with the convention applied to the set $\mathcal{R}'(n)$. Moreover, given any two structures $S_a \in \mathcal{L}(n)$ and $S_b \in \mathcal{L}(n)$, the transition from the set $\mathcal{R}(n)$ to the set $\mathcal{R}'(n)$ corresponds to the exchange of the role of sink and source vertices in some cycles $c_\mu \in G_{ab}$. One can now ask the following question: what is the change in normal phase between structures S_a and S_b, if the set $\mathcal{R}(n)$ is substituted with the set $\mathcal{R}'(n)$? This question is answered by the following lemma.

Lemma A1: Exchange of sink and source vertices in a cycle $c_\mu \in G_{ab}$ changes the normal phase between structures S_a and S_b by a factor $(-1)^{n_\mu + m_\mu + 1}$, where $(2n_\mu)$ is the number of bonds in the cycle c_μ, while m_μ is the number of oriented bonds in this cycle.[8]

Note finally that in Refs. 8 and 9 somewhat different phase convention is given. However, in the case of RRS's the two conventions are in accord, and they lead to the same phase fixation.

2. Rules for overlaps and matrix elements

We give here rules for the derivation of overlaps and matrix elements of one- and two-particle symmetrical operators. Structures S_a and S_b are n-particle regular resonance structures. Their superposition G_{ab} contains ρ cycles $c_1,\ldots,c_\rho$. Cycle $c_\mu \in G_{ab}$ contains $(2n_\mu)$ bonds out of which m_μ are oriented.

All cases which are not explicitly given can be obtained either by using Lemma A1, or by using some of the symmetry relations. For example, if matrix element $\langle S_a | \hat{A}_{ij,kl} | S_b \rangle$ is not explicitly considered, look for the matrix element $\langle S_a | \hat{A}_{ij,lk} | S_b \rangle$ and use $\hat{A}_{ij,kl} = -\hat{A}_{ij,lk}$, etc.

Lemma A2: Overlap $S_{ab} = \langle S_a | S_b \rangle$ equals

$$S_{ab} = 2^{\rho - n} \begin{cases} 1, & \text{all } \rho \text{ cycles } c_\mu \in G_{ab} \text{ are active,} \\ 0, & \text{otherwise.} \end{cases} \tag{A6}$$

a. Matrix elements of one-particle operators $\hat{A}_{kl} = \eta_k^+ \eta_l + \eta_l^+ \eta_k$

Operator $\hat{A}_{kl}$ is diagrammatically represented as a wiggly line connecting vertices (k) and (l). If $k = l$, operator $\hat{A}_k = \hat{A}_{kk}$ is a vertex operator. If $k \neq l$, operator $\hat{A}_{kl}$ is a bond operator. With respect to the superposition G_{ab}, bond operator $\hat{A}_{kl}$ can be either internal or external. It is internal if vertices (k) and (l) are contained in the same cycle, i.e., if $\hat{A}_{kl}$ is supported by a single cycle $c_\mu \in G_{ab}$. It is external otherwise (Fig. 2).

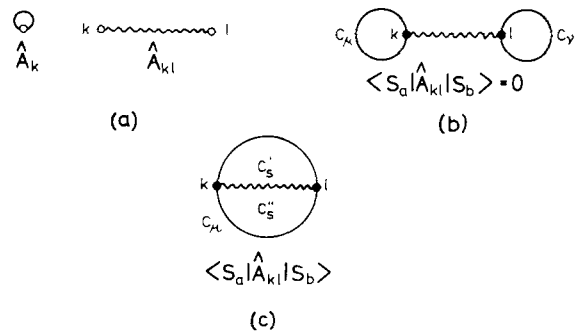

FIG. 2. Diagrammatic representation of operators $\hat{A}_{kl}$ and matrix elements $\langle S_a | \hat{A}_{kl} | S_b \rangle$. (a) Vertex and bond operators. (b) The matrix element of an external bond operator vanishes (Lemma A4). Only cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$ supporting the operator $\hat{A}_{kl}$ are drawn. (c) Matrix elements of an internal bond operator. Only the cycle $c_\mu \in G_{ab}$ supporting the operator $\hat{A}_{kl}$ is drawn. The "bridge" $(s) = (k,l)$ forms two inner cycles over $c_\mu \in G_{ab}$. The cycle c_s' contains the upper segment $(k\text{-}l)$ and a bond (bridge) $(s) = (k,l)$. The cycle c_s'' contains the lower segment $(k\text{-}l)$ and a bond $(s) = (k,l)$. If vertices (k) and (l) are of opposite parity, $(2n_\mu')$ is the number of bonds in the cycle c_s' (Lemma A5). If vertices (k) and (l) are of the same parity, $(2n_\mu' + 1)$ is the number of bonds in the cycle c_s' (Lemma A6). Relations (A8) and (A9) are invariant with respect to the exchange of cycles c_s' and c_s''.

Lemma A3: Let $\hat{A}_k = 2\eta_k^+ \eta_k$ be a vertex operator and let vertex (k) be contained in the cycle $c_\mu \in G_{ab}$. Matrix element $A_{ab}^k = \langle S_a | \hat{A}_k | S_b \rangle$ equals zero if at least one among $(\rho - 1)$ cycles $c_\kappa (\kappa \neq \mu)$ is passive. Otherwise, i.e., if all these cycles are active, matrix element A_{ab}^k equals

$$\langle S_a | \hat{A}_k | S_b \rangle = 2^{\rho - n} \begin{cases} 1, & (k) \text{ is source,} \\ (-1)^{n_\mu + m_\mu + 1}, & (k) \text{ is sink.} \end{cases} \tag{A7}$$

Note that the case when (k) is sink can be derived from the case when (k) is source using Lemma A1.

Lemma A4: Matrix element $A_{ab}^s = \langle S_a | \hat{A}_s | S_b \rangle$ of an external operator $\hat{A}_s = \hat{A}_{kl}$ vanishes [Fig. 2(b)].

If $\hat{A}_s = \hat{A}_{kl}$ is a bond operator internal to some cycle $c_\mu \in G_{ab}$, vertices (k) and (l) form two segments over this cycle. These segments, together with the "bridge" $(s) = (k,l)$ define two cycles, cycle c_s' and c_s''. Quantities n_s' and m_s' in Lemmas A5 and A6 refer to a cycle c_s'. By convention, bridge $(s) = (k,l)$ is considered to be an unoriented bond contained in c_s'. The notion of cycles c_s' and c_s'' is relative, and relations (A5) and (A6) are invariant with respect to a permutation of these two cycles [Fig. 2(c)].

Lemma A5: Let $\hat{A}_s = \hat{A}_{kl}$ be an internal bond operator and let vertices (k) and (l) be of opposite parity. Matrix element A_{ab}^s equals

$$A_{ab}^s = (-1)^{n_s' + m_s' + 1} S_{ab}, \tag{A8}$$

where $(2n_s')$ is the number of bonds in a cycle c_s', while m_s' is the number of oriented bonds in this cycle.

Lemma A6: Let $\hat{A}_s = \hat{A}_{kl}$ be a bond operator internal to a cycle $c_\mu \in G_{ab}$ and let vertices (k) and (l) be of the same

parity. Then (a) if the cycle c_μ is active and/or at least one among other $(\rho-1)$ cycles is passive, matrix element A^{s}_{ab} vanishes, and (b) otherwise, i.e., if the cycle c_μ is passive while all other cycles are active, matrix element A^{s}_{ab} equals

$$A^{s}_{ab} = 2^{\rho-n}\begin{cases}(-1)^{n'_s+m'_s}, & (k)\text{ and }(l)\text{ are source,}\\ (-1)^{n'_s+m'_s+1}, & (k)\text{ and }(l)\text{ are sink,}\end{cases} \tag{A9}$$

where $(2n'_s+1)$ is the number of bonds in the cycle c'_s formed by a bridge $(s)=(k,l)$ over the cycle $c_\mu \in G_{ab}$, while m'_s is the number of oriented bonds in this cycle.

b. Matrix elements of two-particle operators

$\hat{A}_{ij,kl} = \eta_i^+ \eta_j^+ \eta_k \eta_l + \eta_l^+ \eta_k^+ \eta_j \eta_i$

Operator $\hat{A}_{ij,kl}$ is diagramatically represented by two arrows connected by a wiggly line. One arrow connects vertices (i) and (k), and the other connects vertices (j) and (l). Vertices (i) and (j) are associated with the beginning of these arrows, and they are called creation vertices. Vertices (k) and (l) are associated with their ends, and they are called annihilation vertices [Fig. 3(a)]. The definition of the operator $\hat{A}_{ij,kl}$ implies

$$\hat{A}_{ij,kl} = -\hat{A}_{ij,lk} = -\hat{A}_{ji,kl} = \hat{A}_{kl,ij}, \tag{A10}$$

i.e., each diagram changes a sign if either creation or annihilation vertices are permuted. It does not change a sign if the direction of both arrows is simultaneously changed [Fig. 3(b)]. We distinguish three types of operators $\hat{A}_{ij,kl}$. If all four vertices (i), (j), (k), and (l) are different from each other, it is a bond–bond operator. If two out of these four vertices coincide, it is a bond–vertex operator. If in addition another two vertices coincide, it is a vertex–vertex operator [Fig. 3(c)].

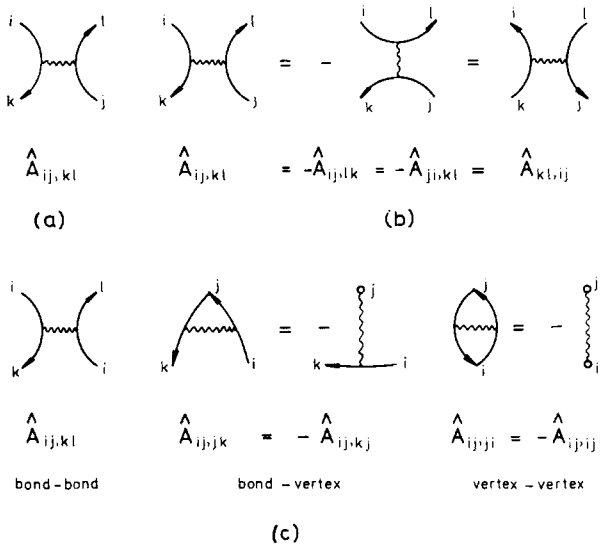

FIG. 3. Diagrammatic representations of two-particle operators $\hat{A}_{ij,kl}$. (a) Diagrammatic representation of operator $\hat{A}_{ij,kl}$. One arrow connects vertices (i) and (k), while the other arrow connects vertices (j) and (l). Creation vertices (i) and (j) are at the beginning of these arrows. (b) Each diagram changes a sign if either creation or annihilation vertices are permuted. It does not change a sign if the direction of both arrows is inverted. (c) Different types of operators $\hat{A}_{ij,kl}$ and their diagrammatic representations.

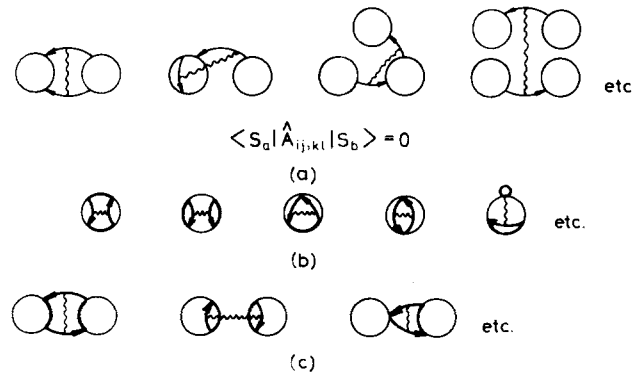

FIG. 4. Diagrammatic representations of matrix elements $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$. (a) Disconnected operators. The two arrows and different segments do not form closed loops. At least one of the cycles supporting these operators contains a different number of creation and annihilation vertices. (b) Internal operators. These operators are supported by a single cycle $c_\mu \in G_{ab}$. Closed loop(s) formed by two arrows and different segments is (are) drawn with a heavy line. (c) External operators. These operators are supported by two cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$. Each of these cycles contains one creation and one annihilation vertex. Closed loop(s) formed by two arrows and different segments is (are) drawn with a heavy line. Internal and external operators are connected operators.

Note that the two creation vertices cannot coincide since (A10) implies $\hat{A}_{ii,kl}=0$. Similarly, the two annihilation vertices cannot coincide either.

With respect to a superposition G_{ab}, operator $\hat{A}_{ij,kl}$ can be either connected or disconnected. If going in the direction indicated by the two arrows and then along the segments defined by vertices (i), (j), (k), and (l) in the superposition G_{ab} one can form either one or two closed loops, operator $\hat{A}_{ij,kl}$ is connected. Otherwise it is disconnected (Fig. 4). Connected operators can be supported by one or at most two cycles $c_\kappa \in G_{ab}$. Each of these cycles contains the same number of creation and annihilation vertices. Connected operators supported by a single cycle are internal, while connected operators supported by two cycles are external.

Lemma A7: If the operator $\hat{A}_{ij,kl}$ is disconnected with respect to the superposition G_{ab}, then the matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ vanishes [Fig. 4(a)].

According to the above lemma, in order to find matrix elements of two-particle operators it is sufficient to find matrix elements of connected operators. These operators are either internal or external.

i. Matrix elements of internal two-particle operators.

Lemma A8: Let $\hat{A}_{ij,kl}$ be internal to the cycle $c_\mu \in G_{ab}$. If creation vertices (i) and (j) [and likewise annihilation vertices (k) and (l)] are adjacent on this cycle, then the matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ vanishes [Fig. 5(a)].

Consider now an operator $\hat{A}_{ij,kl}$ supported by a cycle c_μ, and let creation vertices (i) and (j) [and likewise annihilation vertices (k) and (l)] be not adjacent on this cycle. Vertices (i), (j), (k), and (l) form four segments on the cycle c_μ. In order to express matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ we define a cycle c'_μ which is formed by two out of these four segments. By convention, cycle c'_μ contains segments $(i\text{-}l)$ and $(k\text{-}j)$, i.e., these connecting the beginning of one and the end of another arrow [Fig. 5(b)]. Cycle c'_μ is characterized by two numbers, n'_μ

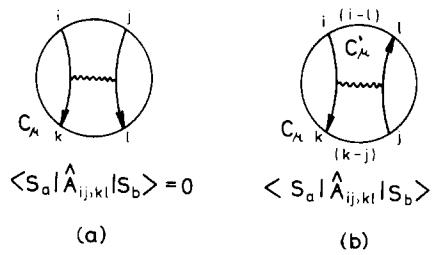

FIG. 5. Two topologically different types of matrix elements of internal operators $\hat{A}_{ij,kl}$. (a) Matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ of an internal operator supported by a cycle $c_\mu \in G_{ab}$ vanishes if creation (annihilation) vertices are adjacent to each other on this cycle (Lemma A8). (b) Otherwise, i.e., if creation (annihilation) vertices are not adjacent to each other on this cycle, the matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ is generally different from zero. In this case one defines the cycle c'_μ in order to express these matrix elements.

and m'_μ. If the cycle c'_μ contains an even number of bonds, then $(2n'_\mu)$ equals this number. If the cycle c'_μ contains an odd number of bonds, then $(2n'_\mu + 1)$ equals this number. In both cases m'_μ is the number of oriented bonds in c'_μ. Note that cycle c'_μ is considered to be formed only by different segments of the cycle c_μ, i.e., it does not contain any part of the diagrammatic representation of the operator $\hat{A}_{ij,kl}$. In the case of bond–vertex operators $\hat{A}_{ij,jl}$ this cycle reduces to a single segment $(i\text{-}l)$, while in the case of vertex–vertex operators $\hat{A}_{ij,ji}$ it contains no segment at all, i.e., $n'_\mu = m'_\mu = 0$ in this case.

Throughout Lemmas (A9)–(A12) the above convention and notation is used.

Lemma A9: Let $\hat{A}_{ij,kl}$ be a bond–bond operator internal to a cycle $c_\mu \in G_{ab}$. Let further operator $\hat{A}_{ij,kl}$ contain an even number of source (sink) vertices. In this case matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ equals

$$\langle S_a|\hat{A}_{ij,kl}|S_b\rangle = S_{ab}\begin{cases}(-1)^{n'_\mu + m'_\mu}, & \text{vertices } (i),\ (j)\ (k), \text{ and} \\ & (l) \text{ are of the same parity,} \\ (-1)^{n'_\mu + m'_\mu + 1}, & \text{two vertices are sink,} \\ & \text{and another two are source,}\end{cases} \tag{A11}$$

where $(2n'_\mu)$ is the number of bonds in the cycle c'_μ [Figs. 6(a) and 6(b)].

Lemma A10: Let $\hat{A}_{ij,kl}$ be a bond–bond operator internal to a cycle $c_\mu \in G_{ab}$. Further let operator $\hat{A}_{ij,kl}$ contain only one sink vertex. Then (a) if the cycle c_μ is active and/or at least one among other $(\rho - 1)$ cycles is passive, the matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ vanishes; or (b) otherwise, i.e., if the cycle c_μ is passive while all other cycles are active, the matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ is equal to

$$\langle S_a|\hat{A}_{ij,kl}|S_b\rangle = (-1)^{n'_\mu + m'_\mu} 2^{\rho - n}, \tag{A12}$$

where $(2n'_\mu + 1)$ is the number of bonds in the cycle c'_μ [Fig. 6(c)].

Lemma A11: Let $\hat{A}_{ij,jk}$ be a bond–vertex operator internal to a cycle $c_\mu \in G_{ab}$. Then (a) if at least one among $(\rho - 1)$ cycles $c_\kappa (\kappa \neq \mu)$ is passive, the matrix element $\langle S_a|\hat{A}_{ij,jk}|S_b\rangle$ vanishes; or (b) otherwise, i.e., if all $(\rho - 1)$ cycles $c_\kappa (\kappa \neq \mu)$ are active, the following cases can be considered.

(b1) Vertices (i), (j), and (k) are source. In this case

$$\langle S_a|\hat{A}_{ij,jk}|S_b\rangle = (-1)^{n'_\mu + m'_\mu} 2^{\rho - n - 1}, \tag{A13a}$$

where $(2n'_\mu)$ is the number of bonds in the cycle c'_μ, i.e., in the segment $(i\text{–}k)$ [Fig. 7(a)].

(b2) Vertex (j) is source, while vertices (i) and (k) are sink. In this case

$$\langle S_a|\hat{A}_{ij,jk}|S_b\rangle = (-1)^{n'_\mu + m'_\mu + 1} 2^{\rho - n - 1}, \tag{A13b}$$

where $(2n'_\mu)$ is the number of bonds in the cycle c'_μ [Fig. 7(b)].

(b3) Vertex (j) is source, while vertices (i) and (k) are of mutually opposite parity, one sink and the other one source. In this case

$$\langle S_a|\hat{A}_{ij,jk}|S_b\rangle = (-1)^{n'_\mu + m'_\mu} 2^{\rho - n - 1}, \tag{A13c}$$

where $(2n'_\mu + 1)$ is the number of bonds in the cycle c'_μ [Fig. 7(c)].

Note that in the case of bond–vertex operator $\hat{A}_{ij,jk}$ supported by a cycle $c_\mu \in G_{ab}$, vertices (i), (j), and (k) form only three segments on this cycle. By definition, cycle c'_μ coincides with the segment $(i\text{–}k)$.

Lemma A12: Let $\hat{A}_{ij,ji}$ be a vertex–vertex operator internal to a cycle $c_\mu \in G_{ab}$. Then (a) matrix element $\langle S_a|\hat{A}_{ij,ji}|S_b\rangle$ vanishes if at least one among $(\rho - 1)$ cycles $c_\kappa (\kappa \neq \mu)$ is passive and/or vertices (i) and (j) are of opposite parity [Fig. 7(d)]; or (b) otherwise, i.e., if all $(\rho - 1)$ cycles $c_\kappa (\kappa \neq \mu)$ are active and also vertices (i) and (j) are of the same parity, matrix element $\langle S_a|\hat{A}_{ij,ji}|S_b\rangle$ equals [Fig. 7(e)]

$$\langle S_a|\hat{A}_{ij,ji}|S_b\rangle = 2^{\rho - n}\begin{cases}1, & (i) \text{ and } (j) \text{ are source,} \\ (-1)^{n_\mu + m_\mu + 1}, & (i) \text{ and } (j) \text{ are sink.}\end{cases} \tag{A14}$$

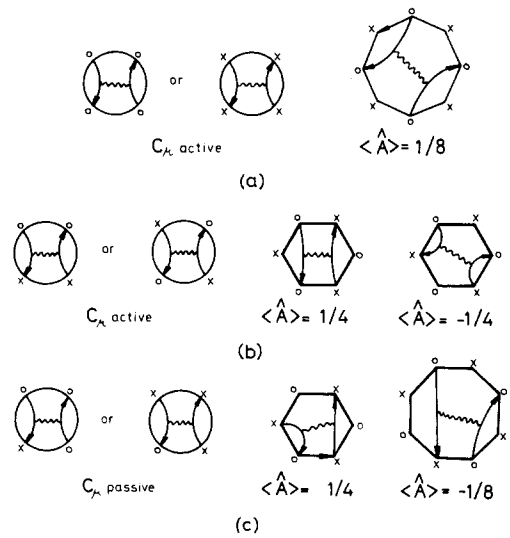

FIG. 6. Diagrammatic representation of matrix elements of internal bond–bond operators. Only the cycle c_μ supporting the operator $\hat{A}_{ij,kl}$ is drawn. If the operator $\hat{A}_{ij,kl}$ contains an even number of source (sink) vertices, then the matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ can be nonzero only provided the cycle c_μ is active [diagrams (a) and (b)]. If the operator $\hat{A}_{ij,kl}$ contains an odd number of source (sink) vertices, matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ can be nonzero only provided the cycle c_μ is passive [diagram (c)]. In the above numerical examples it is assumed that the superposition G_{ab} contains only the cycle c_μ.

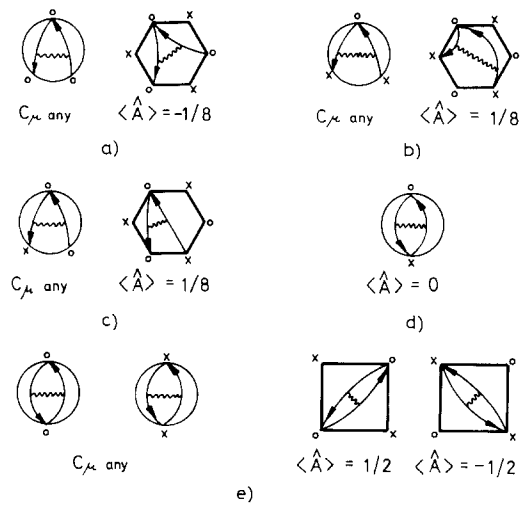

FIG. 7. Diagrammatic representation of matrix elements of internal bond–vertex and vertex–vertex operators. Only the cycle c_μ supporting these operators is drawn. In the above numerical examples it is assumed that the cycle c_μ is the only cycle contained in the superposition G_{ab}. Diagrams (a)–(c) represent matrix elements of internal bond–vertex operators, while diagrams (d) and (e) represent matrix elements of internal vertex–vertex operators. Matrix elements of the internal vertex–vertex operator $\hat{A}_{ij,ji}$ vanish if vertices (i) and (j) are of the opposite parity [diagram (d)].

ii. Matrix elements of external two-particle operators. Let $\hat{A}_{ij,kl}$ be an external operator with respect to the superposition G_{ab}, and let cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$ support this operator. Further let each of these two cycles contain one creation and one annihilation vertex. In addition, we assume that the two arrows corresponding to the operator $\hat{A}_{ij,kl}$ connect these two cycles, i.e., vertices (i) and (k) and likewise vertices (j) and (l) are contained in different cycles (Fig. 8). Vertices (i), (j), (k), and (l) form four segments, two on a cycle c_μ and another two on the cycle c_ν. We define the cycle c' to contain one segment in the cycle c_μ and another in the cycle c_ν. The following convention is used in order to define this cycle, i.e., to fix the corresponding two segments.

Consider creation vertex (i) and let it be contained in the cycle $c_\kappa (\kappa = \mu,\nu)$. In the structure S_a there is a BO containing orbital χ_i. If vertex (i) is source, then by definition cycle c' contains this BO. If however vertex (i) is sink, then by definition cycle c' does not contain this BO. Provided $i \neq l$ this convention uniquely defines that segment in the cycle c_κ which is contained in the cycle c'. If however $i = l$, then the cycle c' is assumed to contain void segment in the cycle c_κ. Creation vertex (j) in conjuncture with the structure S_a likewise defines another segment (Fig. 8).

The cycle c' plays in the case of external two-particle operators the same role as the cycle c'_μ does in the case of internal two-particle operators. It is also characterized by two numbers, n' and m'. The former number (n') is related to the number of bonds in this cycle, while the latter (m') is the number of oriented bonds in the cycle c'.

Throughout Lemmas (A13)–(A17) the above convention and notation is used.

Lemma A13: Let $\hat{A}_{ij,kl}$ be a bond–bond operator supported by cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$. Let further operator $\hat{A}_{ij,kl}$ contain an even number of source (sink) vertices. In this case (a) matrix element $\langle S_a | \hat{A}_{ij,kl} | S_b \rangle$ vanishes if cycles c_μ and c_ν are of different parity (one passive and another active) and/or at least one among remaining $(\rho - 2)$ cycles is passive; or (b) otherwise, i.e., if cycles c_μ and c_ν are of the same parity (either both active or both passive) and all other cycles are active, matrix element $\langle S_a | \hat{A}_{ij,kl} | S_b \rangle$ is equal to

$$\langle S_a | \hat{A}_{ij,kl} | S_b \rangle = 2^{\rho - n - 1} \begin{cases} (-1)^{n'+m'}, & \text{vertices } (i),\ldots,(l) \\ & \text{are all of the same parity,} \\ (-1)^{n'+m'+1}, & \text{each cycle contains} \\ & \text{one source and one sink vertex,} \end{cases} \tag{A15}$$

where $(2n')$ is the number of bonds in the cycle c' [Figs. 9(a) and 9(b)].

Lemma A14: Let $\hat{A}_{ij,kl}$ be a bond–bond operator supported by cycles c_μ and c_ν. Further let operator $\hat{A}_{ij,kl}$ contain only one sink vertex. In this case (a) matrix element $\langle S_a | \hat{A}_{ij,kl} | S_b \rangle$ vanishes if cycles c_μ and c_ν are of the same parity and/or at least one among remaining $(\rho - 2)$ cycles is passive; or (b) otherwise, i.e., if cycles c_μ and c_ν are of opposite parity and all other cycles are active, matrix element $\langle S_a | \hat{A}_{ij,kl} | S_b \rangle$ is equal to

$$\langle S_a | \hat{A}_{ij,kl} | S_b \rangle = (-1)^{n'+m'} 2^{\rho - n - 1}, \tag{A16}$$

where $(2n' + 1)$ is the number of bonds in the cycle c' [Fig. 9(c)].

Consider now external bond–vertex operators $\hat{A}_{ij,jl}$. Note that the cycle c' reduces to a single segment $(i\text{-}l)$. The following two lemmas apply to this case.

Lemma A15: Let $\hat{A}_{ij,jl}$ be a bond–vertex operator supported by cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$. Further let vertex (j) be contained in the cycle c_μ and let all three vertices (i), (j), and (l) be source. In this case (a) matrix element $\langle S_a | \hat{A}_{ij,jl} | S_b \rangle$ vanishes if the cycle c_ν is active and/or one among $(\rho - 2)$ cycles $c_\kappa (\kappa \neq \mu,\nu)$ is passive; or (b) otherwise, i.e., if the cycle c_ν is passive and all $(\rho - 2)$ cycles $c_\kappa (\kappa \neq \mu,\nu)$ are active, matrix element $\langle S_a | \hat{A}_{ij,jl} | S_b \rangle$ is equal to

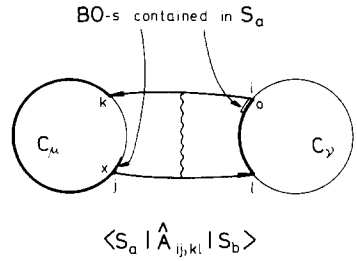

FIG. 8. Matrix elements of an external operator $\hat{A}_{ij,kl}$ supported by cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$. Only cycles c_μ and c_ν are drawn. It is assumed that the two arrows representing the operator $\hat{A}_{ij,kl}$ connect these two cycles. In order to define the cycle c' the following convention is used: Consider two BO's adjacent to creation vertices (i) and (j) and contained in the structures S_a. If the particular creation vertex is source, than c' should contain the corresponding BO. Otherwise, i.e., if the particular creation vertex is sink, than c' should not contain the corresponding BO. In the above example (i) is source and hence c' contains the corresponding BO, while (j) is sink and hence c' does not contain the corresponding BO. The two segments forming the cycle c' are drawn with heavy lines.

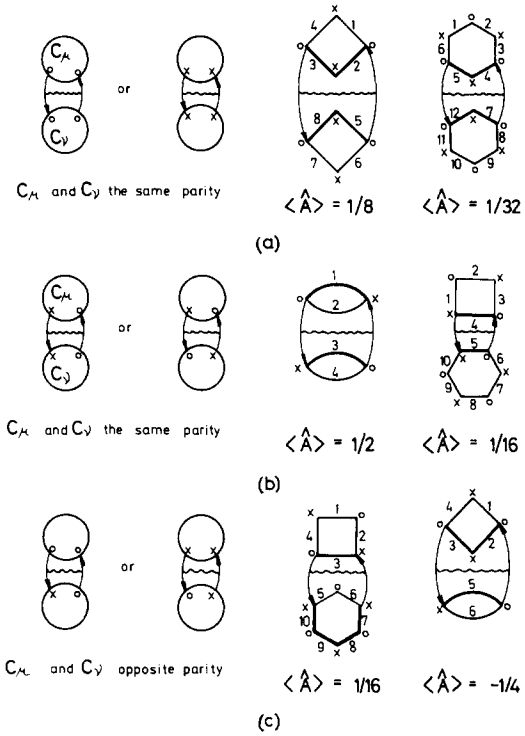

FIG. 9. Diagrams representing matrix elements of external bond–bond operators. Only the cycles c_μ and c_ν supporting the operator $\hat{A}_{ij,kl}$ are drawn. In the numerical examples it is assumed that the superposition G_{ab} contains only these two cycles. Odd numbers denote BO's contained in the structure S_a, while even numbers denote BO's contained in the structure S_b. Bonds contained in the cycle c' are depicted with heavy lines. If the operator $\hat{A}_{ij,kl}$ contains an even number of source (sink) vertices, then the matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ can be nonzero only provided cycles c_μ and c_ν are of the same parity [diagrams (a) and (b)]. If the operator $\hat{A}_{ij,kl}$ contains an odd number of source (sink) vertices, then the matrix element $\langle S_a|\hat{A}_{ij,kl}|S_b\rangle$ can be nonzero only provided cycles c_μ and c_ν are of the opposite parity [diagram (c)].

$$\langle S_a|\hat{A}_{ij,jl}|S_b\rangle = (-1)^{n'+m'}2^{\rho-n-1}, \tag{A17}$$

where $(2n')$ is the number of bonds in the cycle c' [Fig. 10(a)].

Lemma A16: Let $\hat{A}_{ij,jl}$ be a bond–vertex operator supported by cycles c_μ and c_ν and let vertex (j) be contained in the cycle c_μ. Assume further that vertex (j) is source while vertices (i) and (l) are of mutually opposite parity. In this case (a) matrix element $\langle S_a|\hat{A}_{ij,jl}|S_b\rangle$ vanishes if at least one among $(\rho-1)$ cycles $c_\kappa(\kappa\neq\mu)$ is passive; or (b) otherwise, i.e., if all $(\rho-1)$ cycles $c_\kappa(\kappa\neq\mu)$ are active, matrix element $\langle S_a|\hat{A}_{ij,jl}|S_b\rangle$ is equal to

$$\langle S_a|\hat{A}_{ij,jl}|S_b\rangle = (-1)^{n'+m'}2^{\rho-n-1}, \tag{A18}$$

where $(2n'+1)$ is the number of bonds in the cycle c' [Fig. 10(b)].

Finally in the case of external vertex–vertex operators we have the following lemma.

Lemma A17: Let $\hat{A}_{ij,ji}$ be a vertex–vertex operator supported by cycles c_μ and c_ν, and let vertices (i) and (j) be source [Fig. 10(c)]. Then (a) matrix element $\langle S_a|\hat{A}_{ij,ji}|S_b\rangle$ vanishes if at least one among $(\rho-2)$ cycles $c_\kappa(\kappa\neq\mu,\nu)$ is passive; or (b) otherwise, i.e., if all $(\rho-2)$ cycles $c_\kappa(\kappa\neq\mu,\nu)$ are active, matrix element $\langle S_a|\hat{A}_{ij,ji}|S_b\rangle$ equals

$$\langle S_a|\hat{A}_{ij,ji}|S_b\rangle = 2^{\rho-n-1}. \tag{A19}$$

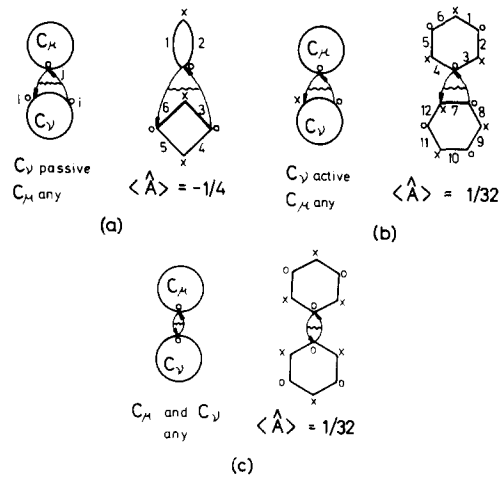

FIG. 10. Diagrammatic representation of matrix elements $\langle S_a|\hat{A}|S_b\rangle$ of external bond–vertex and vertex–vertex operators. Only cycles c_μ and c_ν supporting these operators are drawn. In the numerical examples it is assumed that the superposition G_{ab} contains only these two cycles. Odd numbers denote BO's contained in the structure S_a, while even numbers denote BO's contained in the structure S_b. Diagrams (a) and (b) represent matrix elements of external bond–vertex operators, while diagram (c) represents matrix elements of external vertex–vertex operators.

iii. Matrix elements of reduced operators $\hat{R}_{kk}$, $\hat{R}_{ij,ji}$ and $\hat{R}_{ij,jl}$.

Lemma A18: Let $\hat{R}_{kk} = \hat{A}_{kk} - 1$ be a reduced vertex operator, and let vertex (k) be contained in the cycle $c_\mu \in G_{ab}$. Then (a) if the cycle c_μ is active and/or at least one among other $(\rho-1)$ cycles is passive, matrix element $\langle S_a|\hat{R}_{kk}|S_b\rangle$ vanishes; or (b) otherwise, i.e., if the cycle c_μ is passive while all other cycles are active, a matrix element $\langle S_a|\hat{R}_{kk}|S_b\rangle$ is equal to [Fig. 11(a)]

$$\langle S_a|\hat{R}_{kk}|S_b\rangle = 2^{\rho-n}\begin{cases} 1, & (k) \text{ is source,} \\ -1, & (k) \text{ is sink.} \end{cases} \tag{A20}$$

Lemma A19: Let $\hat{R}_{ij,ji} = 2\hat{A}_{ij,ji} - \hat{A}_{ii} - \hat{A}_{jj} + 1$ be a reduced vertex–vertex operator supported by a cycle $c_\mu \in G_{ab}$ [Figs. 11(b) and 11(c)]. Then

$$\langle S_a|\hat{R}_{ij,ji}|S_b\rangle = S_{ab}\begin{cases} 1, & \text{vertices } (i) \text{ and } (j) \text{ are of the same parity,} \\ -1, & \text{vertices } (i) \text{ and } (j) \text{ are of opposite parity.} \end{cases} \tag{A21}$$

Lemma A20: Let $\hat{R}_{ij,ji}$ be a reduced vertex–vertex operator supported by cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$ [Fig. 11(d)]. Then (a) if either the cycle c_μ or the cycle c_ν is active and/or at least one among other $(\rho-2)$ cycles is passive, matrix element $\langle S_a|\hat{R}_{ij,ji}|S_b\rangle$ vanishes; or (b) otherwise, i.e., if cycles c_μ and c_ν are both passive, while all other cycles are active, matrix element $\langle S_a|\hat{R}_{ij,ji}|S_b\rangle$ equals

$$\langle S_a|\hat{R}_{ij,ji}|S_b\rangle = 2^{\rho-n}\begin{cases} 1, & (i) \text{ and } (j) \text{ are of the same parity,} \\ -1, & (i) \text{ and } (j) \text{ are of opposite parity.} \end{cases} \tag{A22}$$

Lemma A21: Let $\hat{R}_{ij,jl} = 2\hat{A}_{ij,jl} - \hat{A}_{il}$ be a reduced bond–vertex operator supported by a cycle $c_\mu \in G_{ab}$. Then (a)

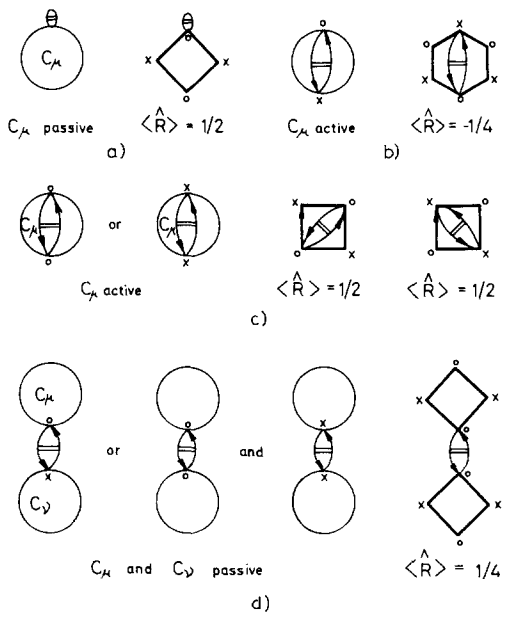

FIG. 11. Diagrammatic representation of reduced vertex and vertex–vertex operators. Only the cycle(s) supporting these operators is (are) drawn. In the above numerical examples it is assumed that the superposition G_{ab} contains only the cycle(s) supporting the reduced operator. (a) Matrix elements of a reduced vertex operator. These matrix elements can be nonzero only if the cycle c_μ is passive. (b) and (c) Matrix elements of internal reduced vertex–vertex operator $\hat{R}_{ijji}$. These matrix elements are different from zero if and only if all the cycles contained in the superposition G_{ab} are active. (d) Matrix elements of external reduced vertex–vertex operators $\hat{R}_{ijji}$. These matrix elements are different from zero if and only if the cycles c_μ and c_ν supporting the operator $\hat{R}_{ijji}$ are passive, while all other cycles are active.

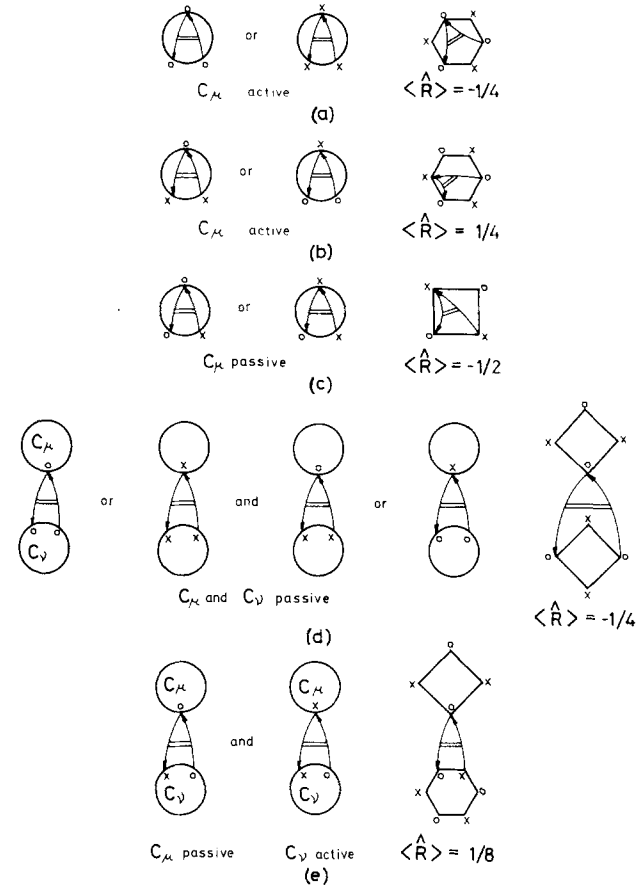

FIG. 12. Diagrammatic representation of reduced bond–vertex operators. Only the cycle(s) supporting these operators is (are) drawn. In the numerical examples it is assumed that the superposition G_{ab} contains only the cycle(s) supporting the reduced operator. Diagrams (a)–(c) represent matrix elements of reduced internal operators, while diagrams (d) and (e) represent matrix elements of reduced external operators. For details see text.

if vertices (i) and (l) are of the same parity matrix element $\langle S_a|\hat{R}_{ijjl}|S_b\rangle$ equals

$$\langle S_a|\hat{R}_{ijjl}|S_b\rangle = S_{ab}\begin{cases}(-1)^{n'_\mu+m'_\mu}, & (j) \text{ is the same parity as } (i) \text{ and } (l),\\ (-1)^{n'_\mu+m'_\mu+1}, & (j) \text{ is opposite parity to } (i) \text{ and } (l),\end{cases} \tag{A23a}$$

where $(2n'_\mu)$ is the number of bonds in the cycle c'_μ [Figs. 12(a) and 12(b)]; or (b) if vertices (i) and (l) are of the opposite parity then matrix element $\langle S_a|\hat{R}_{ijjl}|S_b\rangle$ vanishes, unless the cycle c_μ is passive and all other cycles are active. If, however, cycle c_μ is passive and all other cycles are active, then

$$\langle S_a|\hat{R}_{ijjl}|S_b\rangle = 2^{\rho-n}\begin{cases}(-1)^{n'_\mu+m'_\mu}, & (j) \text{ is source},\\ (-1)^{n'_\mu+m'_\mu+1}, & (j) \text{ is sink},\end{cases} \tag{A23b}$$

where $(2n'_\mu+1)$ is the number of bonds in the cycle c'_μ [Fig. 12(c)].

Lemma A22: Let $\hat{R}_{ijjl}$ be a reduced bond–vertex operator supported by cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$. Further let vertex (j) be contained in the cycle c_μ and let vertices (i) and (l) be of the same parity. Then (a) if either c_μ or c_ν is active and/or at least one among other $(\rho-2)$ cycles is passive, matrix element $\langle S_a|\hat{R}_{ijjl}|S_b\rangle$ vanishes; or (b) otherwise, i.e., if c_μ and c_ν are both passive, while all other cycles are active, matrix element $\langle S_a|\hat{R}_{ijjl}|S_b\rangle$ equals

$$\langle S_a|\hat{R}_{ijjl}|S_b\rangle = 2^{\rho-n}\begin{cases}(-1)^{n'+m'}, & (i),(j), \text{ and } (l) \text{ are of the same parity},\\ (-1)^{n'+m'+1}, & (j) \text{ is one parity, } (i) \text{ and } (l) \text{ are another parity},\end{cases} \tag{A24a}$$

where $(2n')$ is the number of bonds in the cycle c' [Fig. 12(d)].

Lemma A23: Let $\hat{R}_{ijjl}$ be a reduced bond–vertex operator supported by cycles $c_\mu \in G_{ab}$ and $c_\nu \in G_{ab}$. Further let vertex (j) be contained in the cycle c_μ, and let vertices (i) and (l) be of opposite parity. Then (a) if the cycle c_μ is active and/or at least one among other $(\rho-1)$ cycles is passive, matrix element $\langle S_a|\hat{R}_{ijjl}|S_b\rangle$ vanishes; or (b) otherwise, i.e., if the cycle c_μ is passive while all other $(\rho-1)$ cycles $c_\kappa (\kappa \neq \mu)$ are active, matrix element $\langle S_a|\hat{R}_{ijjl}|S_b\rangle$ equals

$$\langle S_a|\hat{R}_{ijjl}|S_b\rangle = 2^{\rho-n}\begin{cases}(-1)^{n'+m'}, & (j) \text{ is source},\\ (-1)^{n'+m'+1}, & (j) \text{ is sink},\end{cases} \tag{A24b}$$

where $(2n'+1)$ is the number of bonds in the cycle c' [Fig. 12(e)].

Proof of Lemma 1 and Theorem 2

In order to prove Lemma 1 it is sufficient to show that $H_{\rm nal}=0$ over X_n implies

$$H_{\rm nal}=\lambda(\hat{N}-n), \tag{A25}$$

where λ is an arbitrary real constant. Note that $H_{\rm nal}=0$ over X_n if and only if

$$\Psi\in X_n \Rightarrow H_{\rm nal}\Psi=0. \tag{A26}$$

The space X_n is spanned by vectors (A1) satisfying the condition (A2). In general vectors (A1) satisfy[19]

$$\eta_k^+|i_1,\ldots,i_k,\ldots,i_{2n}\rangle = (-1)^{\Sigma_k}(1-i_k)|i_1,\ldots,i_k+1,\ldots,i_{2n}\rangle,$$

$$\eta_k|i_1,\ldots,i_k,\ldots,i_{2n}\rangle = (-1)^{\Sigma_k}i_k|i_1,\ldots,i_k-1,\ldots,i_{2n}\rangle, \tag{A27a}$$

where

$$\Sigma_k=i_1+i_2+\ldots+i_{k-1}. \tag{A27b}$$

Operator $H_{\rm nal}$ is a linear combination of reduced antialternant operators, i.e., it is of the form

$$H_{\rm nal}=\sum_i\lambda_{ii}\hat{R}_{ii}+\sum_{ij}\lambda_{ij}\hat{A}_{ij}+\sum_{ijl}\lambda_{ij,jl}\hat{R}_{ij,jl}+\sum_{ijkl}\lambda_{ij,kl}A_{ij,kl}, \tag{A28}$$

where $\lambda_{ij},\ldots,\lambda_{ij,kl}$ are real constants, and different summations are performed only over mutually distinct antialternant operators. Consider now the action of the antialternant operator $\hat{A}_{ij,kl}$ on vectors (A1) satisfying (A2). Assume first that $i,j>n$, and $k,l\leqslant n$, and consider the vector

$$\Psi_0=|1,1,\ldots,1_n,0,\ldots,0\rangle\in X_n. \tag{A29}$$

Operator $\hat{A}_{ij,kl}$ creates two particles in empty orbitals χ_i and χ_j (since $i,j>n$), and annihilates two particles in occupied orbitals χ_k and χ_l (since $k,l\leqslant n$). In other words, it transforms the state Ψ_0 into the state $\Psi_{k\to i,l\to j}$,

$$\Psi_{k\to i,l\to j}=|1,1,\ldots,1,0_k,1,\ldots1,0_l,1,\ldots,1_n,0,\ldots,0,1_i,0,\ldots,0,1_j,0,\ldots\rangle.$$

None of the other antialternant operators contained in $H_{\rm nal}$ acting on the state Ψ_0 creates the state $\Psi_{k\to i,l\to j}$, and hence $H_{\rm nal}\Psi_0=0$ implies $\lambda_{ij,kl}=0$ whenever $i,j>n$ and $k,l\leqslant n$. Changing the state Ψ_0 one finds that the condition (A26) implies that $H_{\rm nal}$ contains no bond–bond operator $\hat{A}_{ij,kl}$, i.e., the last term in the relation (A28) is identically zero. Consider now the action of the antialternant operator $\hat{R}_{ij,jl}$ on vectors (A1). Assume first that $l,j\leqslant n$ and $i>n$. Relations (4) and (A27) imply

$$\hat{R}_{ij,jl}\Psi_0=(-1)^{l+n}\Psi_{l\to i}, \tag{A30}$$

where

$$\Psi_{l\to i}=|1,\ldots,1,0_l,1,\ldots,1_n,0,\ldots,0,1_i,0,\ldots\rangle\in X_n.$$

Relation (A30) holds for each $j\leqslant n(j\neq i,l)$. Moreover, and since indices i and l in an antialternant operator $\hat{R}_{ij,jl}$ are of the opposite parity, none of the remaining antialternant operators acting on the state Ψ_0 creates the state $\Psi_{l\to i}$. Hence $H_{\rm nal}\Psi_0=0$ implies

$$\sum_{j\neq i}^{n}\lambda_{ij,jl}=0. \tag{A31a}$$

Changing the state Ψ_0 one finds

$$\sum_{j\neq i,l}{}'\lambda_{ij,jl}=0, \tag{A31b}$$

where the summation is performed over arbitrary $(n-1)$ indices j satisfying $j\neq i,l$. It is easy to show that conditions (A31b) imply $\lambda_{ij,jl}=0(i\neq j\neq l)$, and hence $H_{\rm nal}$ contains no bond–vertex operator $\hat{R}_{ij,jl}$. Similarly one finds that it does not contain any bond operator $\hat{A}_{ij}$, i.e., only the first term in the relation (A28) remains. Finally one finds

$$\hat{R}_{ii}\Psi_0=\begin{cases}\Psi_0, & i\leqslant n,\\ -\Psi_0, & i>n,\end{cases} \tag{A32}$$

and hence

$$\sum_{i=1}^{n}\lambda_{ii}-\sum_{i=n+1}^{2n}\lambda_{ii}=0. \tag{A33a}$$

If $\Psi_1=\Psi_{1\to n+1}=|0,1,\ldots,1,1_{n+1},0,\ldots,0\rangle\in X_n$ is substituted for Ψ_0 the analogous condition reads

$$\sum_{i=2}^{n+1}\lambda_{ii}-\sum_{i=n+2}^{2n}\lambda_{ii}-\lambda_{11}=0. \tag{A33b}$$

Conditions (A33) imply $\lambda_{11}=\lambda_{n+1,n+1}$. This relation can be easily generalized to

$$\lambda_{ii}=\lambda/2=\text{const}, \tag{A33c}$$

and hence

$$H_{\rm nal}=\lambda\frac{1}{2}\sum_{i}^{2n}\hat{R}_{ii}=\lambda(\hat{N}-n), \tag{A34}$$

which proves Lemma 1. From this lemma it follows that each symmetric Hamiltonian possessing the complete set of alternantlike eigenstates is of the form

$$H=H_{\rm al}+\lambda(\hat{N}-n)+\text{const}. \tag{A35}$$

Since $\hat{N}$ is constant over X_n, and since in addition $H_{\rm nal}$ is an arbitrary linear combination of reduced alternant operators, this proves Theorem 2.

[1] S. Diner, J. P. Malrieu, P. Claverie, and F. Jordan, Chem. Phys. Lett. **2**, 319 (1968); S. Diner, J. P. Malrieu, and P. Claverie, Theor. Chim. Acta **13**, 1 (1969); J. P. Malrieu, P. Claverie, and S. Diner, Theor. Chim. Acta **13**, 18 (1969).

[2] J. P. Malrieu, in *Modern Theoretical Chemistry, Semiempirical Methods of Electronic Structure Calculation* (Plenum, New York, 1977), Vol. 7.

[3] F. Coester, Nucl. Phys. **7**, 421 (1958); F. Coester and H. Kümmel, *ibid.* **17**, 477 (1960); H. Kümmel, in *Lectures on the Many-Body Problem*, edited by E. R. Caianiello (Academic, New York, 1962), p. 265; J. Čižek, J. Chem. Phys. **45**, 4256 (1966); Adv. Chem. Phys. **14**, 35 (1969).

[4] F. E. Harris, H. J. Monkhorst, and D. L. Freeman (unpublished).

[5] C. A. Coulson and G. S. Rushbrooke, Proc. Cambridge Philos. Soc. **36**, 193 (1940).

[6]A. D. McLachlan, Mol. Phys. **4**, 49 (1961); A. D. McLachlan, Mol. Phys. **2**, 271 (1959).
[7]J. Koutecký, J. Chem. Phys. **44**, 3702 (1966).
[8]T. P. Živković, Croat. Chem. Acta **56**, 29 (1983); **56**, 525 (1983).
[9]T. P. Živković, Croat. Chem. Acta (in press).
[10]T. P. Živković, Theor. Chim. Acta (Berlin) **61**, 363 (1982); **62**, 335 (1983); Int. J. Quant. Chem. **23**, 679 (1983).
[11]G. Rumer, Göttingen Nachr., 377 (1932).
[12]T. P. Živković, Chem. Phys. Lett. (in press).
[13]It is of course not invariant with respect to the unitary transformation of the set B.
[14]In the case when there is an even number of atomic orbitals w_i, the partition can be always chosen such that orbitals χ_i and $\overline{\chi}_i$ are of the same parity, which seems most natural. However, in the case when n is odd this is not possible since sets B^o and B^x are required to contain the same number of orbitals. In order to be able to consider odd cases as well, the partition is defined such that χ_i and $\overline{\chi}_i$ are of opposite parity.
[15]These eigenstates are alternantlike only provided there is no accidental degeneracy involving subspaces X_n^+ and X_n^-. If namely $\Psi^+ \in X_n^+$ and $\Psi^- \in X_n^-$ are two degenerate eigenstates, then $\Psi = \Psi^+ + \Psi^-$ is an eigenstate as well, and it is not alternantlike. One can however always find the complete set of alternantlike eigenstates.
[16]T. P. Živković (unpublished results).
[17]R. McWeeny and I. Cooper, J. Chem. Phys. **45**, 226 (1966); **45**, 3484 (1966).
[18]P. O. Löwdin, Arkiv Mat. Astr. Fis. A **35**, 9 (1947); J. Chem. Phys. **18**, 365 (1950).
[19]Richard D. Mattuck, *A Guide to Feynman Diagrams in the Many-Body Problem* (McGraw–Hill, New York, 1967).

Inverse problem of image processing

Harold H. Szu
Naval Research Laboratory, Washington, D.C. 20375

(Received 5 January 1984, accepted for publication 13 April 1984)

A Gel'fand–Levitan inverse problem of image processing is used to determine an optimum trade-off potential $U(\mathbf{f})$ or $V(\mathbf{f})$ from the knowledge of the restored object $Q(\mathbf{f})$ or the output transfer function $\Psi(\mathbf{f})$ and the design knowledge of the optical transfer function (OTF) $S(\mathbf{f})$. Analytic examples and exact solutions are given for the coherent and incoherent image restorations. The minimum mean-square image spread (MMIS) filter is found to yield an identical minimum mean square estimation-error (MMSE) filter of Wiener in the special case, similar to the visual MTF or contrast sensitivity function, peaked at the middle band of spatial frequency channels. The trade-off potential for the linear and noisy motion-blurred image restoration is obtained. By means of the Gel'fand–Levitan inverse transform technique for the reference potential, the design of the energy constrained MMIS filter follows. Lastly, both the direct and inverse problems of incoherent imaging are simultaneously solved with a specific example of the dc incision, namely the central dark field method of Abbe and Zernike. It illustrates the main concept that a useful and efficient new approach of image processing requires a systematical design of both the imaging OTF and image restoration filters under noise.

PACS numbers: 42.30.Va, 02.30. + g

I. INTRODUCTION

The direct problem is, given a trade-off potential, to derive the output transfer function Ψ from which an energy-constrained minimum mean-square image spread (MMIS) filter follows,

$$M(\mathbf{f}) = S_0(\mathbf{f})^{-1}\Psi(\mathbf{f}) . \tag{1}$$

In other words, the noise-to-signal trade-off potential is known,

$$V_0(\mathbf{f}) = \langle |N|^2 \rangle / |S_0|^2 , \tag{2}$$

in terms of the ratio of the power spectral densities of noise and of the optical transfer function (OTF) $S_0(\mathbf{f})$. The solution of the direct problem is obtained by solving the Schrödinger equation

$$-\nabla^2\Psi + V_0(\mathbf{f})\Psi = E\Psi , \tag{3}$$

which balances the noise-to-signal energy ratio with the image spread in terms of the radius of gyration. The solution is called an apodization function Ψ used to soften the exact inverse filter Eq. (1). Also, it is called the output transfer function, in the 2-D spatial frequency domain

$$\mathbf{f} = (\nu, \mu) = (f_x, f_y) \tag{4}$$

because it transfers the object O to the estimated object Q,

$$Q(\mathbf{f}) = M(\mathbf{f})I(\mathbf{f}) = \Psi(\mathbf{f})O(\mathbf{f}) + M(\mathbf{f})N(\mathbf{f}) , \tag{5}$$

if one assumes the linear and shift invariant image equation

$$I(\mathbf{f}) = S_0(\mathbf{f})O(\mathbf{f}) + N(\mathbf{f}) . \tag{6}$$

An object-matched MMIS filter is slightly different from Eq. (3). We have derived

$$-\nabla^2 Q + U(\mathbf{f})Q = EQ , \tag{3'}$$

where the object-dependent U is the deviation of Wiener's Ψ_w from the unity

$$U(\mathbf{f}) = 1 - \Psi_w . \tag{2'}$$

The inverse problem is to derive the trade-off potential V_0 knowing the output transfer Ψ. The necessary and sufficient condition of the unique inverse solution (from E, Ψ to solve V) has been generally known for the Sturm–Liouville equation of which the Schrödinger equation is a special case.[1] It is known in the quantum mechanical potential scattering problem[2] as the Gel'fand–Levitan inverse (scattering) transform technique,[3] which will be applied to the design of imaging and restoration filters for the first time.[4–6]

Firstly, we compare Wiener filtering with dynamic filtering. Can the energy-constrained dynamic filter be identical to the least squares filter of Wiener? What is the imaging optical transfer function $S(\mathbf{f})$ that fulfills both the LS and the constrained LS? To answer both, an equivalence differential equation is derived for arbitrary object spectrum $P = \langle |O|^2 \rangle$. Two specific questions will be answered. The first is related to the comparison between Wiener filtering and dynamic.

Second, we study the noisy image restoration. What is the trade-off potential that includes the ideal restoration $\Psi(\mathbf{f}) = 1$ as its eigenfunctions? Since the exact restoration $Q = 0$ implies $\Psi = 1$ and the ideal inverse filtering

$$M = S_0^{-1}\Psi = S_0^{-1} , \tag{7}$$

then it is desirable to design the OTF for the exact restoration. But the presence of noise can be overly amplified by the zero of the optical transfer function S_0. As a result, the exact inverse filtering S_0^{-1} must be usually apodized by the output transfer function $\Psi(\mathbf{f})$. Therefore, the second question amounts to find a new OTF S which admits $\Psi = 1$ and a new potential V by means of the Gel'fand–Levitan inverse transform. The physical meaning of these exact solutions will be discussed, and an experiment modifying the imaging condition will be indicated.

II. EXACT SOLUTION OF INVERSE PROBLEM

It is interesting to find a consistent trade-off potential between the two approaches, namely the LS method of Wie-

ner and the energy-constrained LS method. The possibility of identical Wiener and dynamic filter is a specific case of the inverse problem that can be answered with a straightforward algebra as follows. The Wiener filter has also been casted as the apodization of inverse filtering, similar to Eq. (1),

$$M_w(\mathbf{f}) = S(\mathbf{f})^{-1}\Psi_w(\mathbf{f}), \tag{8}$$

where the Wiener apodization function is rewritten for arbitrary object $P = \langle |O|^2 \rangle$ in terms of the tradeoff potential $V(\mathbf{f})$:

$$\Psi_w(\mathbf{f}) = (1 + V(\mathbf{f})/P(\mathbf{f}))^{-1}. \tag{9}$$

On the other hand, the formal inverse solution of the Schrödinger equation (3) may be written as

$$V(\mathbf{f}) = \nabla^2\Psi/\Psi + E. \tag{10}$$

If

$$\Psi = \Psi_w,$$

then solving Eq. (9) for V,

$$V = (\Psi(\mathbf{f})^{-1} - 1)P(\mathbf{f}). \tag{11}$$

Equating Eq. (10) with Eq. (11), we obtain the following differential equation:

$$\nabla^2\Psi + (P + E)\Psi = P. \tag{12}$$

Simple solutions for a point object $P = 1$ are given as follows:

$$\Psi = [\nabla^2 + (1 + E)]^{-1}. \tag{13}$$

Since two-dimensional generalizations are possible due to the integrations of a second-order differential equation of constant coefficients, the one-dimensional real and symmetric result is explicitly written as follows:

$$\Psi(\nu) = 1/(1 + E) - \cos(\sqrt{E + 1}\nu). \tag{14}$$

The trade-off potential is found from Eq. (11) to be

$$V(\nu) = \frac{E/(1 + E) + \cos(\sqrt{E + 1}\,\nu)}{1/(1 + E) - \cos(\sqrt{E + 1}\,\nu)}, \tag{15}$$

and the optical transfer function is, when chosen to be real,

$$S(\nu) = \left[\frac{1/(1 + E) - \cos(\sqrt{1 + E}\,\nu)}{E/(1 + E) + \cos(\sqrt{1 + E}\,\nu)}\right]^{1/2}. \tag{16}$$

Then, by definition (1), the identical filter for both performance criteria follows, from Eqs. (14) and (16):

$$M(\nu) = [1/(1 + E) - \cos(\sqrt{1 + E}\,\nu)]^{1/2} \times [E/(1 + E) + \cos(\sqrt{1 + E}\,\nu)]^{1/2}. \tag{17}$$

In particular, these results at the ground state $E = 0$ are listed as follows:

$$\Psi(\nu) = 1 - \cos(\nu), \tag{18}$$

$$V(\nu) = \cos\nu/(1 - \cos\nu), \tag{19}$$

$$S(\nu) = \sqrt{(1 - \cos\nu)/\cos\nu}, \tag{20}$$

$$M(\nu) = \sqrt{\cos\nu(1 - \cos\nu)}, \tag{21}$$

where

$$|\nu| < \pi/2.$$

It is straightforward to verify that

$$-\frac{d^2}{d\nu^2}\Psi + V\Psi = 0. \tag{22}$$

We make three comments as follows.

(i) For the unbounded domain, the complex apodization function that satisfies the equivalent equation (12) is

$$\Psi(\nu) = 1/(1 + E) + e^{i\sqrt{1 + E}\,\nu}. \tag{23}$$

(ii) For arbitrary object spectral density the equivalent differential equation (12) can be expressed in terms of the net broadening, $\Psi - 1$,

$$\nabla^2(\Psi - 1) + (E + P)(\Psi - 1) = E, \tag{24}$$

which is again a Schrödinger equation of a constant inhomogeneity E.

(iii) It is interesting to observe that the matched filtering is similar to the visual contrast-sensitivity function. Note that the optical transfer function is a high pass filter which is unusual in optical imaging, but is possible using computed generated holograms. The equivalence differential equation (12) makes the analytical comparison between two types of filters possible. In Sec. III, the physical meaning of E will be given by a useful application. The other solution of the inverse problem will be given in Sec. IV using the Gel'fand–Levitan inverse transform technique.

III. MOTION-BLURRED NOISE IMAGE RESTORATION

The concept of reference filter[5,6] is closely related to the reference potential that goes into the nutshell of the inverse problem of quantum mechanics. The exact relationship is made possible by generalizing the MMSE filter of Wiener to the energy-constrained MMIS filter. Then the apodization function of the inverse filter satisfies a Schrödinger-like equation in the spatial frequency ν domain, rather in the spatial x domain. To appreciate the importance of this fact, we consider a special case of noiseless imaging. Since the trade-off potential is zero in the case of no noise, then the output transfer function Ψ of the Schrödinger equation is simply integrated to give the important phase factor

$$\Psi(\nu) = \exp(ik\nu), \tag{25}$$

where

$$k = (E)^{1/2}. \tag{26}$$

The restoring output image becomes, by definition (5),

$$Q = MI = MSO \equiv \Psi O, \tag{27}$$

$$Q(\nu) = \exp(ik\nu)O(\nu). \tag{28}$$

The important "gauge" phase freedom means the invariant intensity of spectral density

$$|Q(\nu)|^2 = |O(\nu)|^2, \tag{29}$$

and implies that the physical meaning of E is, in the spatial domain, a shifted but perfect restoration of the object

$$q(x) = o(x + (k/2\pi)) = o(x + (\sqrt{E}/2\pi)). \tag{30}$$

The further significance of such a solution can be best seen in that it unifies earlier results of Harris[7] and Swindell[8] for the case of linear motion blur and consequently a new "de-ghost" filter is obtained as follows.

The point spread function is, by definition of linear motion blur,

$$s(x) = \text{rect}(x/2L)/2L\,, \tag{31}$$

which gives, in the Fourier domain, the optical transfer function, with $K \equiv 2\pi L$,

$$S(\nu) = \sin(K\nu)/K\nu\,. \tag{32}$$

Accordingly, the restoring processor filter, $M \equiv \Psi/S$, has the real and imaging parts

$$M(\nu) = K\nu\cos k\nu/\sin K\nu + iK\nu\sin k\nu/\sin K\nu\,. \tag{33}$$

The real part gives Swindell's result while the imaginary part gives Harris' result. In the spatial domain, one can invert the solution Eq. (33) and therefore derive the general solution

$$m(x) = L\frac{d}{d\nu}\left[\sum_{n=0}^{\infty}\delta(x-(2n+1)L+(k/2\pi))\right] + L\frac{d}{d\nu}\left[\sum_{n=0}^{\infty}\delta(x+(2n+1)L+(k/2\pi))\right], \tag{34}$$

where use is made of the straightforward long division $1/(\exp(ik\nu) - \exp(-ik\nu))$ producing the expansion

$$\frac{1}{\sin K\nu} = 2\sum_{n=0}^{\infty}\sin((2n+1)K\nu)\,. \tag{35}$$

Thus, the Swindell's result is derived from the ground state ($k = \sqrt{E} = 0$) output transfer function $\Psi = 1$, while the Harris result follows the central portion ($|x| < 2L$) of the excited state ($k = K$) output transfer function, i.e., the shifted version of the ideal inverse filter.

If $k = K$ in the spatial frequency domain, then the first term of $M(\nu)$ is the "de-ghost" filter which cancels the ghost image produced by the second term:

$$iK\nu \leftrightarrow L\frac{d}{d\nu}\delta(x)\,,$$

namely the spatial differential operator of Harris.

For white noise the trade-off potential may be approximated by a void impurity to a monoatomic crystal lattice (Kronig–Penny model[9]), where $a = 1/2L$, as

$$V(\nu) = \left(\frac{\nu}{\sin\nu}\right)^2\langle|N|^2\rangle \approx \sum_{n=-\infty}^{\infty} a\delta(\nu - na) - a\delta(\nu)\,. \tag{36}$$

The void "impurity" is due to $S(0)\neq 0$ and consequently $V(0)\neq\delta(\nu)$. Using the Green's function, for no defect,

$$G(\nu,\zeta) = (i/2k)\exp(ik|\nu-\zeta|)\,, \tag{37}$$

and the Bloch theorem

$$\psi(\nu + na) = \exp(i\mu na)\psi(\nu)\,, \tag{38}$$

the output transfer function is known:

$$\psi(\nu) = (a/2k)\psi(0)\exp(i\mu(\nu-\nu')) \times\frac{\exp(i\mu a)\sin k\nu' - \sin k(\nu'-a)}{\cos\mu a - \cos ka}\,, \tag{39}$$

where $\nu' = \nu$ − a·integral part of $[\nu/a]$. Then, the standard perturbative approach follows from the work of Saxon and Hunter[10] and I will not repeat it here.

IV. DIRECT AND INVERSE PROBLEMS OF COHERENT IMAGE RESTORATION

The optical transfer function of coherent (laser) imaging is often represented by an ideal low pass filter

$$S_c(\nu) = \text{rect}(\nu/\pi)\,, \tag{40}$$

which is 1 if $|\nu| < \pi/2$ and zero outside the range where $V_c = \infty$. Four cases are considered as follows.

(i) In the first case of negligible phase (speckle) noise, we solve the equivalent Schrödinger equation, $H_0 = -d^2/d\nu^2 + V_c$, for a particle in a box, where $Vc\simeq 0$ in the zeroth order of approximation, and obtain

$$-\frac{d^2}{d\nu^2}\left[\left(\frac{2}{\pi}\right)^{1/2}\cos(n\nu)\right] = n^2\left[\left(\frac{2}{\pi}\right)^{1/2}\cos(n\nu)\right]. \tag{41}$$

Thus the even output transfer function is

$$\Psi_n(\nu) = \left(\frac{2}{\pi}\right)^{1/2}\cos(n\nu)\,, \tag{42}$$

which vanishes at the wall of the infinite potential well at $\nu = \pm\pi/2$ such that

$$\int_{-\infty}^{\infty}\psi_n(\nu)^2\,d\nu = C_n = 1\,, \tag{43a}$$

and

$$E_n = n^2\,. \tag{43b}$$

Similarly the odd eigenfunction follows:

$$\Psi_n(\nu) = (2/\pi)^{1/2}\sin(n\nu)\,. \tag{44}$$

Thus, the associated restoration filters become

$$M_n(\nu) = \Psi_n(\nu)S_c^{-1}\,, \tag{45}$$

which were first derived by Papoulis[11] for noiseless coherent imagery.

(ii) Due to the constrained moment formulation,[12] we can go beyond Papoulis by including the arbitrary speckle noise power spectral density as the trade-off perturbation potential:

$$V_c = \langle N^2\rangle/|S_c|^2\,. \tag{46}$$

Then in the second case of an arbitrary speckle noise power spectrum the convention perturbation method can be used to calculate the eigenfunction Ψ_c of $H_c = H_0 + V_c$ and the eigenvalue E_c. In particular, if $\langle N^2\rangle\simeq\epsilon_0$ a piecewise constant < 1, then a fast covergent series solution follows:

$$\Psi_{cn} = \Psi_n + \epsilon_0\sum_{m\neq n}\frac{\Psi_m}{(n^2-m^2)}\,, \tag{47}$$

$$E_{cn} = n^2 + \epsilon_0\,. \tag{48}$$

In contrast to Ψ_1, the result Ψ_{c1} becomes bipolar, which provides the necessary cancellation of noise.

(iii) In order to solve the inverse problem of imaging we recapitulate the Gel'fand–Levitan (GL) integral transform for the 1-D spatial frequency domain

$$\Phi(\nu) = \Psi(\nu) + \int_{-\infty}^{\nu}K(\nu,\mu)\Psi(\mu)\,d\mu\,. \tag{49}$$

The kernel of the GL transform itself satisfies the Fredholm integral equation of the second kind, at a fixed $\nu\geqslant\mu > 0$,

$$K(\nu,\mu) = -\Omega(\nu,\mu) - \int_{-\infty}^{\nu} K(\nu,\nu')\Omega(\nu,\mu)\,d\nu' , \quad (50)$$

where Ω in general is the experimental data. Solving K, the trade-off potential follows:

$$V(\nu) = 2\frac{dK(\nu,\nu)}{d\nu} . \quad (51)$$

For simplicity, discrete real states and normalization constants b_n or c_n, respectively, are assumed for the inverse imaging problem

$$\Omega_2(\nu,\nu') = \sum_n \frac{\Phi_n \Phi_n'}{b_n} - \sum_n \frac{\Psi_n \Psi_n'}{c_n} . \quad (52)$$

Solving K_2, the difference of trade-off potentials, V_2 follows:

$$K_2(\nu,\mu) = -\Omega_2(\nu,\mu) - \int_{-\infty}^{\nu} K_2(\nu,\nu')\Omega_2(\nu,\mu)\,d\nu' , \quad (53)$$

$$V_2(\nu) = 2\frac{dK_2(\nu,\nu')}{d\nu} = V - V_1 , \quad (54)$$

where V_1 is associated with Ψ. Since we are not going to willfully alter the other eigenfunctions Ψ_n, unless they are necessarily adjusted due to the required new potential $V = V_1 + V_2$, then we assume tacitly for the time being an identical set of formal eigenfunctions denoted by the lower case ϕ_n,

$$\phi_n = \Psi_n; \quad b_n = c_n; \quad n = 1,\ldots , \quad (55)$$

except that we have demanded the extra ground state, for example, Eq. (40),

$$\phi_0(\nu) = S_c(\nu) . \quad (56)$$

Consequently we have arrived at a single factorized term

$$\Omega_2(\nu,\nu') = \sum_{n=0} \frac{\phi_n \phi_n'}{b_n} - \sum_{n=1} \frac{\Psi_n \Psi_n'}{c_n} = \frac{\phi_0 \phi_0'}{b_0} . \quad (57)$$

Solving the degenerated Fredholm integral equation, we obtain

$$K_2(\nu,\mu) = \frac{-\phi_0(\nu)\phi_0(\mu)}{b_0 + \int_{-\infty}^{\nu} \phi_0(\nu')^2\,d\nu'} . \quad (58)$$

The actual eigenfunctions denoted by the upper case Φ_n are

$$\Phi_n(\nu) = \Psi_n(\nu) + \int_{-\infty}^{\nu} K_2(\nu,\mu)\Psi_n(\mu)\,d\mu , \quad (59)$$

due to the add-on difference potential

$$V_2(\nu) = 2\frac{dK_2(\nu,\nu)}{d\nu} . \quad (60)$$

Given

$$\Omega_2(\nu,\mu) = \phi_0(\nu)\phi_0(\mu)/b_0 , \quad (61)$$

the solution $K_2(\nu,\mu)$, when evaluated at the diagonal value $\mu = \nu$, has a simple form

$$K_2(\nu,\nu) = -\frac{d}{d\nu}\ln\left(b_0 + \int_{-\infty}^{\nu} \phi_0(\nu')^2\,d\nu'\right) , \quad (62)$$

and then the solution of the difference potential becomes

$$V_2(\nu) = -2\frac{d^2}{d\nu^2}\ln\left(b_0 + \int_{-\infty}^{\nu} \phi_0(\nu')^2\,d\nu'\right) . \quad (63)$$

If we add to the first case of coherent imaging, viz. a particle in a box, the perfect restoration ground state, we obtain, from substituting Eq. (56) into the above formula (62),

$$K_2(\nu,\mu) = \frac{-\operatorname{rect}(\nu/\pi)\operatorname{rect}(\mu/\pi)}{b_0 + \pi/2 + \nu} = -\Phi_0(\nu)\phi_0(\mu) , \quad (64)$$

where, as obviously identified from Eqs. (53), (59), and (61),

$$\Phi_0(\nu) = \operatorname{rect}(\nu/\pi)/(b_0 + (\pi/2) + \nu) . \quad (65)$$

Consequently

$$V_2(\nu) = -2(b_0 + \pi/2 + \nu)^{-2} , \quad (66)$$

where the normalization constant is

$$b_0 = [(\pi^2 + 4\pi)^{1/2} - \pi]/2 . \quad (67)$$

(iv) If we delete the ground state Ψ_1 then we obtain, instead of Eqs. (61)–(63),

$$\Omega_2(\upsilon,\mu) = -\Psi_1(\nu)\Psi_1(\mu)/C_0 , \quad (68)$$

$$K_2(\nu,\nu) = -\frac{d}{d\nu}\ln\left(C_0 - \int_{-\infty}^{\nu} \Psi_1(\nu')^2\,d\nu\right) , \quad (69)$$

$$V_2(\nu) = -\frac{d^2}{d\nu^2}\ln\left(C_0 - \int_{-\infty}^{\nu} \Psi_1(\nu')^2\,d\nu'\right) . \quad (70)$$

For the particle-in-a-box example of coherent imaging, i.e., to delete the Papoulis ground state

$$\Psi_1 = (2/\pi)^{1/2}\cos\nu; \quad c_1 = 1 , \quad (71)$$

we obtain the following potential slightly different from the result of Abraham and Moses (1980),[13]

$$V_2(\nu) = \frac{8\sin 2\nu + 32(\cos\nu)^4}{(2\nu - \pi + \sin 2\nu)^2} . \quad (72)$$

V. DIRECT AND INVERSE PROBLEMS OF INCOHERENT IMAGING

The incoherent imaging point spread function $s_0(x)$ is the intensity of the diffraction pattern $s_i(x) = (\sin \pi x/\pi x)^2$, behind the ideal shuttle aperture, e.g., 1-D slit,

$$\operatorname{rect}(\nu) = 1, \quad \text{if } |\nu| \leqslant \tfrac{1}{2} . \quad (73)$$

Then the optical transfer function is the roof-top function due to the Fourier transform (FT) convolution theorem, depicted in Fig. 1:

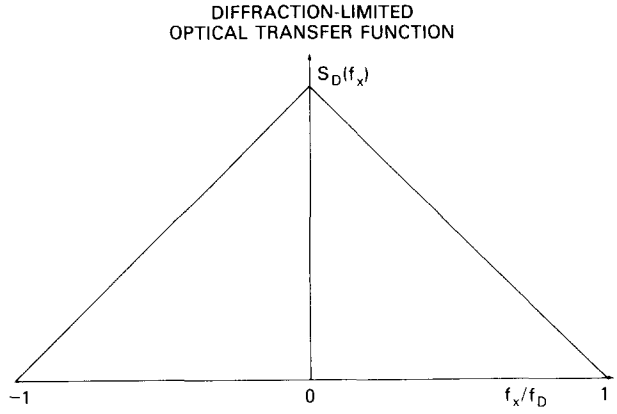

FIG. 1. Diffraction-limited optical transfer function (OTF) of a 1-D clear slit. The formalism should be clearly adaptable to a radially symmetric case.

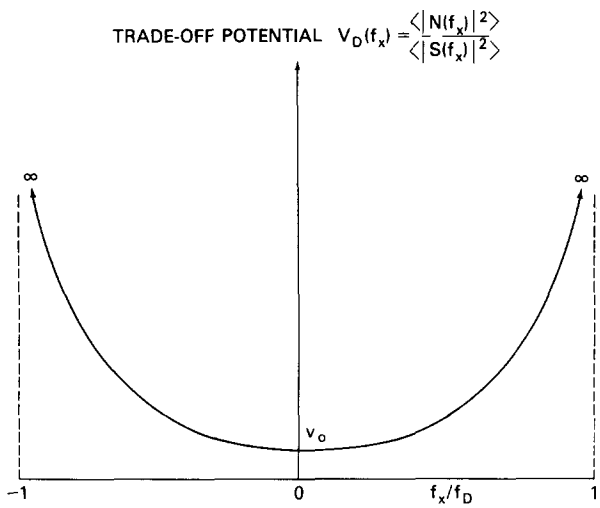

FIG. 2. Trade-off potential between the low-frequency noisy spikes and the high-frequency cutoff singularity.

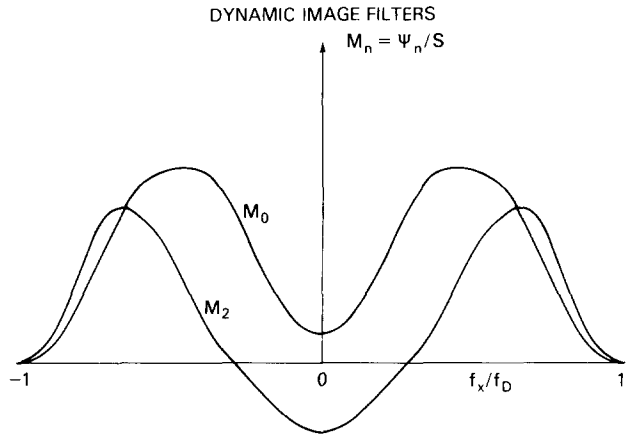

FIG. 4. Energy constrained MMIS Filters which go beyond classical Wiener MMSE matched filtering so as to be dynamically balanced at two conflicting goals: (1) suppressing noise and the oversensitivity at the high-frequency spectrum, and (2) maintaining the resolution by reducing the distortion and the desensitivity near the low-frequency spectrum for between-class object identification.

$$S_i(\nu) = FT\{s_i(x)\} = \mathrm{rect}(\nu)*\mathrm{rect}(\nu) \equiv \Lambda(\nu) = 1 - |\nu| . \tag{74}$$

Thus, the trade-off potential is a symmetric hyperbolic well shown in Fig. 2:

$$V_i(\nu) = \frac{\langle N^2\rangle}{|S_i|^2} = \frac{1}{\mathrm{SNR}(1-|\nu|)^2} . \tag{75}$$

Schematically the eigenfunctions are shown in Fig. 3 together with filters shown in Fig. 4.

Since it is desirable to incise the dc spectrum from the OTF $S_i(\nu)$, then the new OTF $S(\nu)$, without the subscript i, gives the new trade-off potential

$$V(\nu) = \frac{\langle N^2\rangle}{|S|^2} \simeq \frac{1}{\mathrm{SNR}}\left[\frac{1}{(1-|\nu|)^2} + \delta(\nu)\right] . \tag{76}$$

It is straightforward to solve, for weak SNR $= \frac{1}{2}$ the ground state $E = 0$,

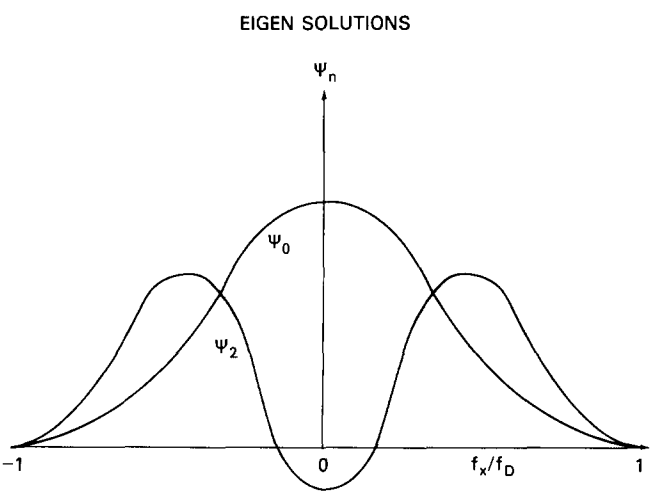

FIG. 3. Eigenfunctions of the positive potential well shown in Fig. 2.

$$\left[-\frac{d^2}{d\nu^2} + \frac{2}{(1-|\nu|^2} + 2\delta(\nu)\right]\left[\frac{1}{1-|\nu|}\right] = E\left[\frac{1}{1-|\nu|}\right] = 0 . \tag{77}$$

Thus, we obtain exactly $E = 0$ the ground state apodization function

$$\Phi_0(\nu) = \frac{1}{1-|\nu|} = \frac{1}{S_i(\nu)} , \tag{78}$$

which turns out to be the ideal inverse filter of the incoherent imaging. Consequently, the dynamic filter for the dc-incision incoherent imaging follows:

$$M_0 = \Phi_0 S^{-1} = (S_i S)^{-1} . \tag{79}$$

Since we use the principle value (P) identity to represent

$$\delta(\nu) = \mathrm{Re}\left[\frac{1}{-i\pi}\left(\mathrm{P}\frac{1}{\nu} - \frac{1}{\nu - ia}\right)\right] = \frac{a}{\pi(\nu^2 + a^2)} , \tag{80}$$

we obtain

$$\frac{1}{|S(\nu)|^2} = \frac{1}{(1-|\nu|)^2} + \delta(\nu) = \frac{(1 + a/\pi)\nu^2 - (2a/\pi)|\nu| + a^2 + a/\pi}{(1-|\nu|)^2(\nu^2 + a^2)} . \tag{81}$$

Solving the singular algebraic equation, we find

$$S(\nu) = (1-|\nu|)[(\nu + ia)/(\nu - c)] , \tag{82}$$

where

$$c = \frac{a}{\pi + a} + i\frac{\sqrt{\pi a}}{\pi + a}(1 + a\pi + a^2)^{1/2} . \tag{83}$$

The OTF at zero frequency is equivalent to a phase plate because in the limit $a = 0$:

$$S(0) = -i\lim_{a\to 0}(a/c) = -i\pi , \tag{84}$$

and due to the incision of dc, the real part vanishes,

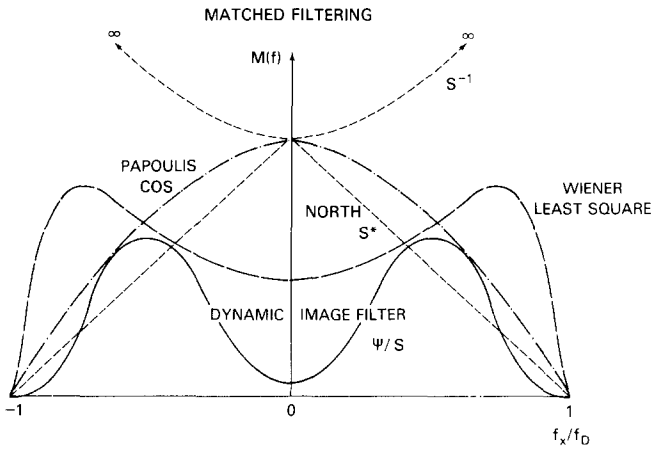

FIG. 5. Comparison among various matched filters. Papoulis' was taken from Ref. 11 to compare a smooth object in noise-free case. Wiener least-squares filtering was taken from Ref. 14. The energy-constrained MMIS filter is the quotient of the restoration output transfer function $\Psi(f)$ and the optical transfer function $S(f)$.

$$\mathrm{Re}[S(0)] = 0 . \tag{85}$$

Therefore, we conclude that a new incoherent imaging, with the suppression of dc is similar to the dark field electron microscope, and the image can be restored by means of the almost ideal inverse filter $|S_i|^{-2}$:

$$M_0 = (1 - |\nu|)^{-2}[(\nu - c)/(\nu + ia)] , \tag{86}$$

where the difference is the complex phase factor (83).

VI. CONCLUSION

In conclusion, we plot various filters[14] in Fig. 5. The simple rule of thumb of any man-created sciences is the following: derive an appropriate Sturm–Liouville differential equation and then apply the Gel'fand–Levitan inverse transform to both the direct and the inverse problems. In the image processing, the output transfer function $\Psi(\mathbf{f})$ is governed to the lowest-order moments by a Schrödinger-like equation, rather of the image filter $M(\mathbf{f})$ itself. Then it is straightforward to apply the Gel'fand–Levitan inverse transform to synthesize both designs of the imaging OTF $S(\mathbf{f})$ and the noisy image restoration filters $M(\mathbf{f})$, as illustrated in Sec. V. If someone argues that one should not squeeze the physics into the mathematics in order to bridge the gap between the direct and the inverse problems, the author agrees in principle but disagrees in the present case, because both the image and the signal are man-created sciences and therefore man can create them with their processing in mind. If we properly formulate the problem with a solution in mind, then the present systematic approach is bound to be useful, and so to speak pass the basic selection rule of utility filter for the man-made disciplines, as opposed to the natural sciences. In order to be not only useful but also effiicient, the present approach provides a complete system design of the pre-imaging system and the post-imaging restoration. This viewpoint is reminiscent of both the signal spectrum preshaping before sending through a channel and the receiving signal restoration and decoding. The present analyses beyond matched filtering open up a new prospect in image and signal processing.

[1]H. H. Szu, C. E. Carroll, C. C. Yang, and S. Ahn, J. Math. Phys. **17**, 1236 (1976).
[2]K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Spring, New York, 1977), Chap. 3.
[3]L. D. Faddeyev, J. Math. Phys. **4**, 72 (1963).
[4]F. Dyson, J. Opt. Soc. Am. **65**, 551 (1975).
[5]M. Kanal, H. E. Moses, and S. K. Mitter, Transp. Theory Stat. Phys. (USA) **8**, 163 (1979).
[6]T. Kailath, IEEE Trans. Inf. Theory **IT-20**, 146 (1974). H. E. Moses, J. Math. Phys. **24**, 2550 (1983).
[7]J. L. Harris, J. Opt. Soc. Am. **56**, 569 (1966).
[8]W. Swindell, Appl. Opt. **9**, 2459 (1970).
[9]R. de L. Kronig and W. G. Penny, Proc. R. Soc. London Ser. A **130**, 499 (1931).
[10]D. S. Saxon and R. A. Hunter, Philips Res. Rept. **4**, 81 (1949). R. G. Newton, J. Math. Phys. **24**, 2152 (1983).
[11]A. Papoulis, J. Opt. Soc. Am. **62**, 1423 (1972).
[12]See H. H. Szu, J. Math. Phys. **25**, 2773 (1984) for a systematic approach and comparisons.
[13]P. B. Abraham and H. E. Moses, Phys. Rev. A **22**, 1333 (1980); H. E. Moses and R. T. Prosser, J. Math. Phys. **24**, 2146 (1983).
[14]C. W. Helstron, J. Opt. Soc. Am. **57**, 297 (1967).

Direct problem of image processing

Harold H. Szu
Naval Research Laboratory, Washington, D.C. 20375

(Received 5 January 1984; accepted for publication 13 April 1984)

Matched filters of image processing are revisited from the viewpoint of Wiener's regulations of the ill-posed problem and the superresolution. For both the object-independent image processing and the object-matched pattern recognition, the minimum mean-square image spread (MMIS) filters are derived from a general constrained moment expansion method, which includes Wiener's method based on the performance criteria: minimum mean square estimation errors (MMSE) as a special case.

PACS numbers: 42.30.Va, 02.30. + g

I. A TREND OF IMAGE PROCESSING

Image processing of a single frame redistributes the information content for easier interpretation. In order to further enhance the desired information by means of its temporal behavior, our previous efforts increase the number of degrees of freedom from a two-dimensional (2-D) space (x, y) to a 3-D space (x, y, t). The local instances of good seeing technique for multiple frames (video/movie) have been developed[1,2] to overcome the blurring due to time integration of images formed through a turbulent medium.[3] The technique is a piecewise shift-and-add and rejection which does not exploit specific object knowledge, but rather uses the statistical mean value of a set of multiple frames as the reference for the piecewise shift-and-add. In order to compensate for absorption and other effects,[4] we have developed a cross-spectral Wiener filter[5] which is obtained by the minimum mean-square estimation (MMSE) technique in the energy wavelength space λ. This further increases in the dimensionality and leads to processing in 4-D space (x, y, t, λ). The present effort deals with single frame image processing, and the direct and inverse problems are systematically treated for the first time.

II. INTRODUCTION TO MATCHED FILTERS

We would like to implement a spatial filter $m(\mathbf{r})$ which can take into account a set of constraints based on *a priori* knowledge. For example, a useful constraint is a constant output noise from a processor when the input noise power spectrum is filtered by $m(\mathbf{r})$. Since a positive real function can be identified as an energy if it is derived from the squared modulus of another function, then the spatial integrated noise power spectrum will be referred to as a total noise energy. A technique which applies the constraint of constant output noise energy to the method of minimum radius of gyration will be presented. It will be shown to be a special case of a general constrained moment method when used to measure the image spread. Thereby, general applications to image restoration and pattern recognition become possible.

The performance criterion of the technique will be identified as a minimum mean-square *image spread* (MMIS) as opposed to the conventional minimum mean square *estimation errors* (MMSE).[6–10] Since the present criterion does not depend on the unknown estimation errors, it is convenient for image restoration. The resulting filter is a bandpass filter which is analytic in the spatial frequency domain. It will be referred to as an energy-constrained filter because the total energy is traded between the output noise energy and the energy associated with the radius of gyration. Such an energy-constrained filter that belongs to the class of constrained least squares filters[11,12] cannot be entirely new,[13,14] but a systematical treatment with applications to image restoration and pattern recognition problems[15,16] has not been explored. Hence, a general method of constrained moment expansion will be presented to construct a regulated inverse filter for image restoration filter to reject a known power spectrum of noise and background clutter. Both old object-independent and new object-matched MMIS filters are derived from the constrained moment method.

Since a spatial filter is usually compared to the Wiener matched filter used commonly for image restoration and traditional pattern recognition, a Wiener model,[17–19] as used by Helstrom, is adopted here for the necessary comparison. A two-dimensional (2-D) stochastic object $o(\mathbf{r})$ is imaged with a lens point spread function (psf) denoted by $s_0(\mathbf{r})$ and an independent, additive, Gaussian, white noise of zero mean denoted by $n(\mathbf{r})$. Under the assumption of linear, incoherent, and space-invariant imaging, the noisy image is given by

$$i(\mathbf{r}) = s_0(\mathbf{r}) * o(\mathbf{r}) + n(\mathbf{r}) , \tag{1a}$$

where the star stands for the convolution integral

$$s_0(\mathbf{r}) * o(\mathbf{r}) = \int s_0(\mathbf{r} - \mathbf{r}') o(\mathbf{r}') d\mathbf{r}' . \tag{1b}$$

By the assumption of statistical independence and zero mean

$$\begin{aligned} \langle n(\mathbf{r}) o(\mathbf{r}') \rangle &= 0 , \\ \langle n(\mathbf{r}) \rangle &= 0 , \end{aligned} \tag{2}$$

where angular brackets specify an ensemble average. A slight generalization of the Wiener model to include a stochastic psf as done by Slepian should entail no difficulty in the present problem. However, for simplicity of presentation, the discussion will be limited to a deterministic psf. But white noise $n(\mathbf{r})$ is generalized to the form

$$\tilde{n}(\mathbf{r}) = c(\mathbf{r}) + n(\mathbf{r}) , \tag{3}$$

where $c(\mathbf{r})$ denotes clutters that are treated as a known statistical pattern noise. For this $\tilde{n}(\mathbf{r})$, we would like to construct a deterministic spatial filter $m(\mathbf{r})$ which when convolved with a

stochastic image $i(\mathbf{r})$ will produce a stochastic object estimator $q(\mathbf{r})$ in the form

$$q(\mathbf{r}) = m(\mathbf{r}) \star i(\mathbf{r}) . \tag{4}$$

However, such a simple deconvolution of Eq. (1) can be ill-conditioned. As a matter of fact, the exact inverse filter $m_i(\mathbf{r})$ indicated by the subscript i and used for deconvolution (4) is ill conditioned for two reasons: it does not exist everywhere and it produces noise amplification. Since both are important aspects for subsequent development, we shall briefly review the divergent nature of ill-posed problems[20] and then describe what may be called a Wiener's method of regulation. We shall compare both aspects to the energy-constrained minimization of the radius of gyration in Sec. III. The general theory of the constrained moment method will be presented in Sec. IV, and then object-independent and object-matched dynamic filters follow in Secs. V and VI, respectively. In order to reject background clutters, we will fabricate an overall system psf by means of the Gel'fand–Levitan transform between two trade-off potentials in the sequel paper[21] as follows:

$$s(\mathbf{r}) = s_0(\mathbf{r}) \star s_p(\mathbf{r}) , \tag{5a}$$

where $s_p(\mathbf{r})$ is a preimaging filter for example the incision of *dc* spectrum to get rid of the frame aperture background

$$s_p(\mathbf{r}) = s_0(\mathbf{r})^{-1} \star s(\mathbf{r}) = s(\mathbf{r}) - \int s(\mathbf{r}) d\mathbf{r} . \tag{5b}$$

New results will be recapitulated, and the future work will be indicated, in Sec. VII.

III. NOISE CONSTRAINED PERFORMANCE CRITERIA FOR INVERSE FILTERS: WIENER'S REGULATION METHOD

This section begins with the definition of an exact inverse filter $m_i(\mathbf{r})$ to indicate reasons why it is ill conditioned. Then from the regulation viewpoint of the ill-conditioned deconvolution, Wiener's performance criterion: minimum mean-square estimation errors, $(q - o)^2$, is presented. As a result, an alternative and practical performance criterion, namely a minimum mean-square image spread, follows. Such an output energy-constrained criterion, related to a human visual performance, will be modeled by a composite filter $M(\mathbf{f})$, namely the (high-pass) inverse filter $S_0(\mathbf{f})^{-1}$ being apodized by a (Gaussian) regulation function $\Psi(\mathbf{f})$:

$$M(\mathbf{f}) = S_0(\mathbf{f})^{-1} \Psi(\mathbf{f}) . \tag{6}$$

Here and henceforth upper case letters denote Fourier transforms in the 2-D spatial frequency domain $\mathbf{f} = (f_x, f_y) = (\nu, \mu)$ alternatively. The Fourier transform of the point spread function $s_0(\mathbf{r})$ is commonly called the optical transfer function (OTF) $S_0(\mathbf{f})$. Since the imaging OTF $S_0(\mathbf{f})$ has the effect of low-pass filter, then the inverse of OTF $S_0(\mathbf{f})$ has the effect of a high-pass filter such that their product is the unity, all-pass filter.

An exact inverse filter $m_i(\mathbf{r})$ will be defined in terms of the Dirac delta function $\delta(\mathbf{r})$. Substituting Eq. (1) into Eq. (4), one has

$$q(\mathbf{r}) = m(\mathbf{r}) \star s_0(\mathbf{r}) \star o(\mathbf{r}) + m(\mathbf{r}) \star n(\mathbf{r}) . \tag{7}$$

An exact inverse filter may be defined as one with the property of no image spread,

$$m_i(\mathbf{r}) \star s_0(\mathbf{r}) = \delta(\mathbf{r}) . \tag{8}$$

Postponing for the moment consideration of the existence of m_i, one can formally substitute definition (8) into (7) and obtain

$$q(\mathbf{r}) = o(\mathbf{r}) + m_i(\mathbf{r}) \star n(\mathbf{r}) . \tag{9}$$

Taking the ensemble average of Eq. (9), one would then obtain a perfect restoration of no image spread

$$\langle q(\mathbf{r}) \rangle = \langle o(\mathbf{r}) \rangle . \tag{10}$$

However, due to the convolution theorem, the Fourier transform of definition (8) is

$$M_i(\mathbf{f}) S_0(\mathbf{f}) = 1 , \tag{11}$$

so that

$$M_i(\mathbf{f}) = 1/S_0(\mathbf{f}) . \tag{12}$$

Consequently, the function $M_i(\mathbf{f})$ does not exist everywhere for it becomes divergent at the zeros and also outside the band limits of the OTF $S_0(\mathbf{f})$. A simple 1-D example of noiseless motion blur shows an *ad hoc* regulation method as follows.

An example of the divergence is illustrated by the noisy motion blur. Since the blur psf $s_0 = \text{rect}(x)$ has the OTF $S_0 = \sin(\pi\nu)/\nu\pi$ associated with an infinite set of zeros where the sine function vanishes, the inverse filter (12) becomes divergent if dividing 1, instead of the regulation function Ψ, with zeros. But it has been fabricated by Tsujichu's technique of clipping peaks, namely a valid approximation for a band-limited restoration. Thus, the first problem of ill conditioning that concerns the divergence of m_i has been piecewise regulated in an *ad hoc* fashion.

Nevertheless, more serious is the noise amplification problem due to the high pass character of the inverse filter. From the Fourier transform of Eq. (9) and the result of Eq. (12) follow

$$Q(\mathbf{f}) = O(\mathbf{f}) + N(\mathbf{f})/S_0(\mathbf{f}) . \tag{13}$$

Thus, the ever decreasing signal-to-noise ratio at higher and higher spatial frequencies, prevents the exact inverse filter from giving a faithful restoration.

A simple way to insure the overall existence of $m_i(\mathbf{r})$ as well as regulate the divergence of output noise $m_i(\mathbf{r}) \star n(\mathbf{r})$ is to allow the replacement of the ideal image-spread δ-function used for (8), $m \star s_0$, with a broader regulation function $\psi(\mathbf{r})$,

$$m(\mathbf{r}) \star s_0(\boldsymbol{\delta}) = \psi(\mathbf{r}) , \tag{14a}$$

or, after Fourier deconvolution, the ideal image spread function is replaced by the regulation function

$$M(\mathbf{f}) S_0(\mathbf{f}) = \Psi(\mathbf{f}) . \tag{14b}$$

Our problem remains to implement $\Psi(\mathbf{f})$ in an analytic fashion for an arbitrary noise $N(\mathbf{f})$. To illustrate this regulation concept (14), Wiener's filtering of white noise $n(\mathbf{r})$ is reinterpreted as follows. Given the Wiener filter M_w, to be derived later, it follows readily, by factorization (14b), that Wiener's regulation function, Ψ appended by the subscript w,

$$\Psi_w(\mathbf{f}) = (1 + R(\mathbf{f}))^{-1} , \tag{15}$$

where R stands for the ratio of noise-to-signal,

$$R(\mathbf{f}) = \langle |N|^2 \rangle / (\langle |O(\mathbf{f})|^2 \rangle |S_0(\mathbf{f})|^2) . \tag{16}$$

The real, positive and monotonously decreasing function Ψ_w can also be called as Wiener's apodization function in the context of imaging, because the Wiener filter can be rewritten by (14) and (12) as

$$M_w(\mathbf{f}) = S_0(\mathbf{f})^{-1} \Psi_w(\mathbf{f}) = M_i(\mathbf{f}) \Psi_w(\mathbf{f}) . \tag{17}$$

The amount of broadening between (8) and (14) may be measured with their difference $B_w(\mathbf{f})$ in the Fourier domain defined by

$$B_w = 1 - \Psi_w , \tag{18}$$

which, using (15), equals

$$B_w = R(1+R)^{-1} , \tag{19}$$

or after a proper normalization, equals the noise-to-signal ratio

$$B_w / \Psi_w = R . \tag{20}$$

Evidently two limiting situations of (20) can happen simultaneously, i.e., $R(\mathbf{f}_l) \ll 1$ and $R(\mathbf{f}_h) \gg 1$; where l and h denote, respectively, the low and the high spatial frequencies. Because the optical imaging has the effect of a low-pass filter, a relative small noise $R(\mathbf{f}_l)$ occurs over a dominant object spectrum at low-frequency region, while a large $R(\mathbf{f}_h)$ appears at a high-frequency region. As a result, using binomial expansions of (15) in these limits,

$$\Psi_w(\mathbf{f}_l) \approx 1 - R(\mathbf{f}_l) , \quad R \ll 1 , \tag{21}$$

$$\Psi_w(\mathbf{f}_h) \sim 1/R(\mathbf{f}_h) , \quad R \gg 1 , \tag{22}$$

an approximate broadening near the low frequency,

$$M_i(\mathbf{f}_l) - M_w(\mathbf{f}_l) = B_w S_0^{-1} \approx (\langle |N|^2 \rangle / \langle |O|^2 \rangle) |S_0|^{-3} , \tag{23}$$

is inversely proportional to the cubic power of OTF, while in the high-frequency limit (22) the Wiener filter (17) is identical to the (North[7] or Van Vleck and Middleton[8]) matched filter implemented by Vander Lugt[12] for the optical character recognition

$$M_w(\mathbf{f}_h) = (\langle |O|^2 \rangle / \langle |N|^2 \rangle) S_0(\mathbf{f}_h)^* . \tag{24}$$

The effect of Wiener's restoration of the degraded optical imaging (1),

$$Q = M_w I = \Psi_w O + M_w N , \tag{25a}$$

becomes clear, viz.,

$$Q = O - R(1+R)^{-1} O + M_w N , \tag{25b}$$

if use is made of the inverse identity to replace Ψ_w (15),

$$\Psi_w = (1+R)^{-1} = 1 - R(1+R)^{-1} . \tag{26}$$

Thus, it is seen from (25) that the object spectrum O will always be subtracted off so long as $R \neq 0$. Furthermore, the object spectrum can never be restored in the limit $S_0(\mathbf{f}) = 0$ that $R = \infty$. The last fact can be restated as that a Wiener filter can never be able to have a superresolution beyond the diffraction limit where the OTF $S_0(\mathbf{f}) = 0$. But, as a sufficient result of no superresolution, a Wiener filter can suppress the output noise divergence in the mean square sense,

$$\langle |M_w N|^2 \rangle = (1+R)^{-2} R \langle |O|^2 \rangle , \tag{27}$$

which vanishes in the same limit $S_0(\mathbf{f}) = 0$ that $R = \infty$. However, while these sufficient proofs (25) and (27) are valid for all frequencies we left with the suspicion whether a global optimization procedure for all frequencies is necessary, and the possibility that Wiener might have indiscriminately and overly killed the ill-conditioned deconvolution (13) for all frequencies. This viewpoint will be elaborated as follows.

The Wiener filter results from the global performance criterion of minimum mean-square estimation error, $q(\mathbf{r}) - o(\mathbf{r})$,

$$\int \langle (q-o)^2 \rangle d\mathbf{r} = \text{minimum} , \tag{28}$$

or in the Fourier domain, using Parseval formula,

$$\int \langle |Q - O|^2 \rangle d\mathbf{f} = \text{minimum} , \tag{29}$$

where, by definition (5),

$$Q = MI . \tag{30}$$

The standard optimization procedure takes the variation of (29) with respect to M for an extremum which yields the Wiener filter

$$M_w = \langle I^* O \rangle / \langle I^* I \rangle . \tag{31}$$

The result (31) can be directly obtained from definition (30) after averaging it with $I^*(\mathbf{f}')$,

$$\langle I^*(\mathbf{f}') Q(\mathbf{f}) \rangle = M(\mathbf{f}) \langle I^*(\mathbf{f}') I(\mathbf{f}) \rangle , \tag{32}$$

and replacing $\langle I^* Q \rangle$ with $\langle I^* O \rangle$ by invoking the extremum principle of orthogonal projection, expressed in the inner product bracket notation (,) as

$$\langle (I, Q - O) \rangle = 0 , \tag{33}$$

i.e., the estimation error $Q - O$ becomes an extremum if it is orthogonal to the *input data* I.

Obviously, there exists, other than the input data I, viz., *a priori* knowledge (about the object denoted by $O' < O$ that is the subset of O which yields $I' < I$) such that the same principle of orthogonal projection might work better than (33) (cf. Fig. 1):

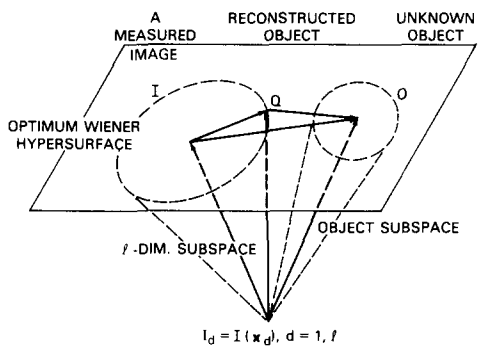

FIG. 1. Geometric construction for matched filtering. The triangle on the optimum Wiener hypersurface, which is orthogonal to the (lexicographically stacked) image vector I, guarantees that $\|O - I\|^2 = \|(O - Q) + (Q - I)\|^2 = \|O - Q\|^2 + \|Q - I\|^2 + 2(O - Q, Q - I)$ and the cross terms $(O - Q, Q - I)$ are positive and therefore must vanish because of Eq. (33) and $Q = MI$ as close as possible to O. However, the difference of the lengths, $\|O - I\|^2 - \|O - Q\|^2 < \|M - 1\|^2 \|I\|^2$ which is self-consistently determined, is exceedingly sensitive to the discrete sample space of I. The subsets I and O are those on the surface with smaller regions, which may be required by the *a priori* knowledge producing desirable restorations.

$$\langle (I', Q - O') \rangle = 0 . \tag{34}$$

Also the correlation of neighborhood frequency $\mathbf{f}'$ with $\mathbf{f}$ is not used in (32) because, according to the Wiener–Khinchine theorem, no mode coupling, $\delta(\mathbf{f}' - \mathbf{f})$, is possible for the stationary correlation $\langle I^*(\mathbf{f}')I(\mathbf{f}) \rangle$. Instead, the unknown object statistics $\langle O^*(\mathbf{f}')O(\mathbf{f}) \rangle$ is needed, but is impractical to be ascertained. Based on these criticisms for image applications, we would like to point out the possibility that, instead of using these unknown covariance alone, *a priori* knowledge, such as a local spatial frequency coupling through its curvature effects, may produce a convenience, and perhaps in some cases, better restoration. In fact, an alternative performance criterion is the minimum mean-square image spread $\Psi = MS_0$, which is to control mathematically the broadened regulation function that will allow us to restore imagery without relying implicitly upon the unknown object. This will lead to a local mode coupling theory in the spatial frequency domain as follows.

IV. A CONSTRAINED MOMENT EXPANSION METHOD

A real function $g(\mathbf{r})$ to be employed as a general performance measure for a constrained moment method will be introduced. In the previous example of a given psf $s_0(\mathbf{r})$ the performance criterion of inverse filtering was the minimum mean-square image spread (MMIS) $\psi(\mathbf{r}) = m(\mathbf{r}) * s_0(\mathbf{r})$, which stands for the function $g(\mathbf{r})$. In the present example of preinverse filtering, the lens psf $s_0(\mathbf{r})$ is either unknown or unused and thus the performance criterion needs to be accordingly modified. A straightforward modification will be replacing MMIS of the image filter function $m(\mathbf{r})$. Such is the ideal case of noisy signal processing without distortion $s_0(\mathbf{r}) = \delta(\mathbf{r})$, which suffers no divergence of inverse filtering because the inverse of OTF is simply the inverse of unity. However, either imaging with an unknown psf $s_0(\mathbf{r})$, or a preinverse filter $s_p(\mathbf{r})$, that can suppress an arbitrary noise $\tilde{n}(\mathbf{r})$ including background clutters, we will consider a combined system psf $s(\mathbf{r})$ and the regulation function $g(\mathbf{r})$. Thus, in viewing various possibilities, an arbitrary regulation function $g(\mathbf{r})$ is conveniently treated in the linear vector space and then specialized to $\psi(\mathbf{r})$ later. A 2-D time-independent real function $g(\mathbf{r})$ and its power series expansion are considered here, although a general spatiotemporal complex function can also be treated, using time-dependent expansion coefficients and spatial feature functions, $f_n(\mathbf{r})$, $n = 1,2,\ldots$. Moreover, the time independent and real functions are assumed to be square integrable. Then, such a class of functions defines, via the Schwarz inequality, a linear vector space, namely a Hilbert or inner-product space. Since noise is generally characterized by its real and positive power spectral density, the noise suppression to be traded with object distortion requires the square of a bipolar $g(\mathbf{r})$ as a real and positive energy density

$$\rho(\mathbf{r}) = g(\mathbf{r})^2 \geqslant 0 . \tag{35}$$

Since g is square-integrable, then a constant total energy follows:

$$\int \rho(\mathbf{r}) d\mathbf{r} = \int g(\mathbf{r})^2 \, d\mathbf{r} = \text{const} . \tag{36}$$

Since only lower orders of image moments are usually measured, then only lower orders of spatial frequency derivatives are known by the Parseval formula

$$\int |2\pi\mathbf{r}|^{2n} g(\mathbf{r})^2 \, d\mathbf{r} = (-)^n \int G^*(\mathbf{f}) \nabla_f^{2n} G(\mathbf{f}) d\mathbf{f} . \tag{37}$$

Furthermore, the estimated function $g(\mathbf{r})$ based on the partial series must be consistent with the constraints imposed upon the function $g(\mathbf{r})$ itself. Thus, the purpose of constrained moment method is to intersect the hypervolume representing the partial series with the hypersurfaces representing those constraints, so that the estimation uncertainty may be minimized in the Hilbert space. This is analytically fulfilled using the minimum energy concept of a Hamiltonian system. The partial series is associated with the kinetic energy of the form (37) while the constraints are casted in the form of potential energy. Before presenting this formalism, various constraints are discussed as follows.

There are two kinds of constraints. The first kind is intrinsic to the image. The second kind is external and pertinent to the imaging distortion and restoration system. The first kind is concerned about the symmetry, size, positivity, and other interesting features of the image known *a priori*. The second kind is generally known and related to the system that forms and processes the imagery. Both are useful for the 2-D information retrieval. The essence of a constrained moment method is to provide a convenient framework that can incorporate these constraints. This may be evident from the context of automatic pattern recognition where a moment expansion is often used. The expansion constitutes a pattern space which yields for the necessary reduction of information in the feature space, which in turn gives, via decision functions, the classification space. We expand the density function (35)

$$\rho(\mathbf{r}) = \sum c_n \zeta_n(\mathbf{r}) , \tag{38}$$

where

$$c_n = \int \rho(\mathbf{r}) \zeta_n(\mathbf{r}) W(\mathbf{r}) d\mathbf{r} \tag{39}$$

follows from the orthonormality assumption

$$\delta_{n,m} = \int \zeta_n(\mathbf{r}) \zeta_m(\mathbf{r}) W(\mathbf{r}) d\mathbf{r} , \tag{40}$$

with respect to the weighting (window) function $W(\mathbf{r})$. From the completeness relationship, obtained from (38) and (39),

$$\sum_n \sqrt{W}(\mathbf{r}) \zeta_n(\mathbf{r}) \zeta_n(\mathbf{r}') \sqrt{W}(\mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}') , \tag{41}$$

follows another representation

$$\sum_n f_n(\mathbf{r}) f_n(\mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}') , \tag{42}$$

where a new member is obtained by a linear transformation T

$$\int T(\mathbf{r},\mathbf{r}') \sqrt{W}(\mathbf{r}') \zeta_n(\mathbf{r}') d\mathbf{r}' = f_n(\mathbf{r}) . \tag{43}$$

A simple example of the moment expansion is the so-called raw moments based on the set

$$\zeta_n(\mathbf{r}) = 1,\mathbf{r},\ldots\,; \quad n = 0,1,\ldots\,, \tag{44}$$

which are complete, but not yet orthogonal, and can be used to obtain central moments and (algebraic) invariant moments.

The constraint upon $\rho(\mathbf{r})$ itself is based on *a priori* knowledge about the symmetry and smooth properties, etc. The other constraint is a constant output noise energy from the image processor

$$\int \langle (k \star n)^2 \rangle d\mathbf{r} = \text{const}\,, \tag{45}$$

where $k(\mathbf{r})$ is the processor filter function. Such a constrained moment method becomes the constrained image restoration when it is applied to regulate the inverse filtering introduced early. In that case, the processor filter function $k(\mathbf{r})$ is the regulated inverse filter $m(\mathbf{r})$. The real function $g(\mathbf{r})$ is identified with the regulation function $\psi(\mathbf{r})$ that replaced the Dirac $\delta(\mathbf{r})$,

$$m(\mathbf{r}) \star s(\mathbf{r}) \equiv \psi(\mathbf{r}) = g(\mathbf{r})\,. \tag{46}$$

Consequently $g(\mathbf{r})$ must be also localized and symmetric, and its positive energy density is likewise localized and symmetric,

$$\rho(\mathbf{r}) = \psi(\mathbf{r})^2\,. \tag{47}$$

Thus, all the odd moments vanish due to the symmetry property. There remain the even moments

$$c_{2n} = \int \rho(\mathbf{r}) \zeta_{2n}(\mathbf{r}) d\mathbf{r} = 0\,, \quad n = 0,1,2\,. \tag{48}$$

The lowest even moment gives a measure of the peak sharpness

$$c_2 = \int \rho(\mathbf{r}) \zeta_2(\mathbf{r}) d\mathbf{r} = \text{minimum}\,, \tag{49}$$

where

$$\zeta_2 = 4\pi^2 |\mathbf{r}|^2\,. \tag{50}$$

The extrinsic constraints are the constant total energy

$$c_0 = \int \rho(\mathbf{r}) d\mathbf{r} = \text{const}\,, \tag{51}$$

and the constant total output noise from the processor

$$\int (m \star n)^2 \, d\mathbf{r} = \text{const}\,. \tag{52}$$

Thus, it is clear that the measured moments can be used to formulate the optimization problem and the expansion coefficient can be used to incorporate the constraint.

V. OBJECT-INDEPENDENT IMAGE PROCESSING MMIS FILTER

Since the convenient performance criterion that replaces the estimation error $Q - O$ with the image-spread $\Psi = MS_0$ is no longer depending on the unknown object O, then we can choose, in analogy to Green's function, a test point object $\delta(\mathbf{r})$. Since $\langle o \rangle = \delta$, then one expects by (10) $\langle q \rangle \approx \delta$. Therefore one defines an output transfer (Green's) function

$$\langle q \rangle = m \star i = m \star s_0 \star \langle o \rangle = \psi \star \delta = \psi\,, \tag{53}$$

where, due to the low-pass imaging lens s_0 and arbitrary noise n, the ideal image δ is replaced by the point-spread function $\psi = m \star s_0$. Therefore, its spot size is conveniently measured by the radius of gyration

$$\int |2\pi \mathbf{r}|^2 (m \star s_0)^2 \, d\mathbf{r} = \text{minimum}\,, \tag{54}$$

which is demanded to be as small as possible. Such a performance criterion of minimum mean-square image spread must be constrained by a constant output noise in order to regulate the divergence problem [(13) and (27)]

$$\left\langle \int (m \star n)^2 \, d\mathbf{r} \right\rangle = \text{const} \tag{55}$$

and, moreover, constrained by a constant total image intensity throughput in order to be meaningful

$$\int (m \star s_0)^2 \, d\mathbf{r} = \text{const}\,. \tag{56}$$

This set of constrained least squares equations can be implemented in a slightly different and perhaps more flexible manner than those constrained least squares filters implemented in the coordinate domain. One obtains the result as follows:

$$-\nabla^2 \psi + V(\mathbf{f})\psi = E\psi\,. \tag{57}$$

The Laplacian operate, $\nabla^2 = |d/d\,\mathbf{f}|^2$, in the Fourier domain is derived from (54) by means of the Parseval formula and integrations by parts,

$$4\pi^2 \int |\mathbf{r}|^2 \, |\psi(\mathbf{r})|^2 \, d\mathbf{r} = -\int \psi^*(\mathbf{f}) \nabla^2 \psi(\mathbf{f}) d\,\mathbf{f}\,. \tag{58}$$

The potential energy $V(\mathbf{f})$ is derived from the constant of output noise by definition $\Psi = MS_0$,

$$\int |M|^2 \langle |N|^2 \rangle d\,\mathbf{f} = \int \Psi^* \frac{\langle |N|^2 \rangle}{|S_0|^2} \Psi \, d\,\mathbf{f}\,. \tag{59}$$

The total energy E is the Lagrangian multiplier that multiplies the quadratic normalization (55)

$$\int \psi^2 \, d\mathbf{r} = \int \Psi^* \Psi \, d\,\mathbf{f} = 1\,. \tag{60}$$

Since all integrands of Eqs. (58)–(60) are real and homogeneously quadratic, then the Lagrangian equation associated with the standard Hamilton variation principle gives that the sum of integrands must be vanished namely the result (57). Formally we can rewrite (57) as a Hamiltonian system

$$H\Psi_n = E_n \Psi_n\,, \tag{61}$$

where eigenvalues E_n and eigenfunctions Ψ_n are self-consistently determined by the condition of high-frequency boundedness. If a proper choice of total output intensity E_n is given, then an eigenfunction Ψ_n will be given a dynamic filter $M_n = \Psi_n S_0^{-1}$. It is clear that the ground state Ψ_0 is associated with the minimum image-spread provided that $\sqrt{E_0}$ is related to the output image location.

Solutions of the Schrödinger equation are abundant in quantum mechanics and well documented using analytic approximated and numerical methods for arbitrary potentials. Thus an analytic formula may be available for the trade-off between the relative noise energy $V(f)$ and the image-spread function $-\nabla^2 \Psi$. For example, such a resulting dynamic fil-

ter may behave like a high pass inverse filter $S_0(f)^{-1}$ multiplied by a Gaussian apodization function, $\Psi(f)$ being a typical ground state of the Schrödinger equation (57),

$$M(\mathbf{f}) = (a + b|\mathbf{f}|^2)\exp(-c|\mathbf{f}|^2). \tag{62}$$

Since it is similar to a model of the bandpass filter for visual perception, it is tempting to conjecture that this MMIS, minimum mean-square image spread, might be actually corresponding to a physiological model of visual processing. Although any further study[22] would require the material beyond the present scope, we would like to support the conjecture with some observations in image processing[23] which shows empirical tests using a similar functional form[24] giving better results[25] than those of Wiener's.

VI. OBJECT-MATCHED PATTERN RECOGNITION MMIS FILTER

In order to illustrate that *a priori* knowledge of the object can also be incorporated, we apply the constrained moment method to derive an object-matched MMIS filter. Instead of minimizing the image spread of the output transfer function ψ due to a point object, we can minimize the spread of the restored object $q(r)$. Although two are identical for a point object, for an extended object the generalization allows us to derive a new object-matched dynamic filter as follows. The Parseval formula gives

$$\left\langle 4\pi^2 \int |\mathbf{r}|^2 q^2(\mathbf{r})d\mathbf{r} \right\rangle = \left\langle \int |\nabla Q(\mathbf{f})|^2 \, d\mathbf{f} \right\rangle = \text{minimum}. \tag{63}$$

Note that the zeroth moment of g is related to Wiener's performance criteria (28).

We will assume a piecewise space invariant image for the general (object) space variant image. This will allow us to use Fourier deconvolution of linear image equation within the large piecewise region. Similarly, we obtain, within the piecewise region, respectively,

$$Q \equiv MI = MSO + MN \equiv \Psi O + \hat{N}. \tag{64}$$

We minimize the spread of restored object under two constraints: (i) a constant output noise energy

$$\int \langle \hat{n}(\mathbf{r}) \rangle d\mathbf{r} = \int \langle |\hat{N}(\mathbf{f})|^2 \rangle d\mathbf{r}$$

$$= \int \left\langle Q^* \frac{|N|^2}{|SO + N|^2} Q \right\rangle d\mathbf{f}, \tag{65}$$

where use is made of

$$\hat{N} \equiv MN \equiv QN/I;$$

and (ii) a constant restored object energy, by using the Lagrange multiplier E,

$$\int \langle q^2(\mathbf{r}) \rangle d\mathbf{r} = \int \langle |Q(\mathbf{f})|^2 \rangle d\mathbf{f}. \tag{66}$$

Then we use the identical technique in obtaining the Schrödinger equation of one realization of the restored object ensemble in Fourier domain denoted by the overhead bar

$$-\nabla^2 \overline{Q}(\mathbf{f}) + U(\mathbf{f})\overline{Q}(\mathbf{f}) = E\overline{Q}(\mathbf{f}). \tag{67}$$

Given the noise-to-signal ratio, Eq. (16),

$$R(\mathbf{f}) = \langle |N|^2/|S|^2|O|^2 \rangle, \tag{68}$$

the Wiener output transfer function Ψ_w [Eq. (15)] enters naturally into Eq. (65) and becomes the difference of the Schrödinger potential from the unit output noise energy

$$U(f) = \left\langle \frac{|N|^2}{|SO + N|^2} \right\rangle = \frac{R}{1 + R} \equiv 1 - \Psi_w. \tag{69}$$

We arrive at the object matched dynamic filter

$$M(\mathbf{r}) = \int \frac{\overline{Q}(\mathbf{f})}{\overline{I}(\mathbf{f})} \exp(i2\pi \mathbf{f} \cdot \mathbf{r}) d\mathbf{f}. \tag{70}$$

VII. DISCUSSION

Since a blurred object tends to be diffusively vanishing over a larger spatial frequency domain, then the scheme of minimizing the radius of gyration of the object in the coordinate domain will produce in the spatial frequency domain the desired frequency spread under a constant output noise energy. It reduces for a point object, $O = 1$ to the previously derived Schrödinger equation of the output transfer function. We note that from Eq. (64) follows

$$Q(\mathbf{f}) = \Psi(\mathbf{f}) + \hat{N} \simeq \Psi(\mathbf{f}); \tag{71}$$

from Eqs. (68) and (41) follow

$$R(\mathbf{f}) = V(\mathbf{f}); \tag{72}$$

and from Eqs. (69) and (71) follow, for a weak signal-to-noise ratio,

$$U(\mathbf{f}) = V(\mathbf{f})(1 + V(\mathbf{f}))^{-1} \simeq V(\mathbf{f}). \tag{73}$$

We expect that the general object-matched MMIS filter should work better than the previous one. Since we have included the *a priori* knowledge of the expected object spectral density $|O|^2$, the optimization scheme gives

$$-\nabla^2 \overline{Q} + (1 - \Psi_w)\overline{Q} = E\overline{Q}, \tag{74}$$

where the Wiener apodization function

$$\Psi_w = (1 + \langle |N|^2 \rangle / |S|^2 |O|^2)^{-1} \tag{75}$$

is the difference potential with respect to the uniform case $U_0 = 1$:

$$-\nabla^2 \overline{Q}(\mathbf{f}) = (E - 1)\overline{\overline{Q}}(\mathbf{f}). \tag{76}$$

We shall discuss the implications in the sequel paper,[21] where we give the Abraham–Moses application of the Gel'fand–Levitan transformation used for the reconstruction of difference potentials. Also, the coherent and incoherent image restoration will be exemplified.

[1] H. H. Szu and J. A. Blodgett, J. Opt. Soc. Am. **72**, 1666 (1982).
[2] H. H. Szu, J. A. Blodgett, and L. Sica, Opt. Commun. **35**, 317 (1980).
[3] H. H. Szu, J. S. Lee, and J. A. Blodgett, Proc. Soc. Photo. Opt. Instrum. Eng. **186**, 148 (1979).
[4] H. H. Szu, Opt. Commun. **32**, 229 (1980).
[5] H. H. Szu, J. Opt. Soc. Am. **72**, A1788 (1982).
[6] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series* (Wiley, New York, 1949), p. 59; A. Kolmogoroff, Bull. Acad. Sci. U.R.S.S. Math. **5**, 3 (1941).
[7] D. O. North, "Analysis of the Factors which determine signal-to-noise Discrimination in Radar," RCA Laboratories, Princeton, NJ, Classified Technical Rep. PTR-6C, June 1943; D. O. North, Proc. IEEE **51**, 1016 (1963).

[8]J. H. Van Vleck and David Middleton, J. Appl. Phys. **17**, 940 (1946).
[9]"Special Issue on Matched Filters," IRE Trans. Inf. Theory **IT-6**, 310–408, June 1960.
[10]*Benchmark Papers in Electrical Engineering and Computer Science*, series edited by J. B. Thomas, *Detection and Estimation: Applications to Radar*, edited by S. S. Haykin (IEEE, New York, 1976).
[11]J. Herrmann, J. Opt. Soc. Am. **70**, 28 (1980).
[12]A. B. Vander Lugt, IEEE Trans. Inf. Theory **IT:10**, 139 (1964); J. W. Goodman, *Introduction to Fourier Optics* (McGraw–Hill, San Francisco, 1968) Chap. 7.
[13]A. Papoulis, J. Opt. Soc. Am. **62**, 1423 (1972).
[14]H. A. Smith, J. SIAM Appl. Math. **14**, 23 (1966).
[15]H. J. Caulfield, R. Haimes, and D. Casasent, Opt. Eng. **19**, 152 (1980).
[16]A. A. Q. Kadkly, Opt. Acta **26**, 461 (1979).
[17]B. Roy Frieden, J. Opt. Soc. Am. **64**, 682 (1974); B. Roy Frieden, Comp. Graphics Image Proc. **12**, 40 (1980).
[18]B. E. A. Saleh, Appl. Opt. **13**, 1833 (1974).
[19]M. J. Lahart, J. Opt. Soc. Am. **69**, 1333 (1979).
[20]H. H. Szu, "The foundation of single frame image processing," in *Optics in 4-dimension* (AIP, New York, 1980).
[21]H. H. Szu, J. Math. Phys. **25**, 2767 (1984).
[22]R. A. Messner and H. H. Szu, *Proceedings Computer Vision and Pattern Recognition* (IEEE Computer Society, New York, 1983), p. 522.
[23]B. R. Hunt, Proc. IEEE **63**, 693 (1975).
[24]B. R. Hunt, IEEE Trans. **AU-20**, 94 (1972).
[25]B. R. Hunt, IEEE Trans. **C-22**, 805 (1973).

Resonance fluorescence spectra of a strong bichromatic field interacting with a three-level atom in the "V" configuration[a)]

Constantine Mavroyannis
Division of Chemistry, National Research Council of Canada, Ottawa, Ontario, Canada K1A OR6

(Received 22 December 1983; accepted for publication 20 April 1984)

We have considered the fluorescence spectra arising from the interaction of a three-level atom in the "V" configuration with two strong electromagnetic fields whose initially populated modes ω_a and ω_b are equal to the two atomic transition frequencies, respectively. The excitation spectra are calculated by making use of the Green function formalism where the hierarchy of the Green functions is truncated by a decoupling scheme which takes into account photon–photon correlations from each laser field in the limit of high photon densities. The excitation spectra consist of the central peak of the excitation frequency ω_a and four pairs of sidebands, which are peaked at the frequencies $\omega_a \pm \eta_1\sqrt{2}$, $\omega_a \pm \eta_2/\sqrt{2}$, $\omega_a \pm (\eta_1^2+\eta_2^2)^{1/2}/\sqrt{2}$, and $\omega_a \pm 2(\eta_1^2+\eta_2^2)^{1/2}/\sqrt{2}$, where η_1 and η_2 are the Rabi frequencies of the two laser fields, respectively. Similar spectra are exhibited near the excitation frequency ω_b. Numerical calculations for selected values of the Rabi frequencies are graphically presented and compared with those known in the literature. In the Appendix expressions for the excitation spectra are derived first when the two laser fields are treated classically, and second, when the fields are quantized but photon–photon correlations are neglected. It is shown that when the two laser fields are treated classically, the derived results are identical to those obtained when the fields are quantized only when photon–photon correlations arising from each laser field are discarded. The derived results when the fields are quantized but without photon–photon correlations taken into account are also identical to those obtained by the use of the dressed-atom approach. The appearance of the new sidebands reveals the boson character of the photon fields in question. The merit of the method is that it yields results describing both the classical as well as the quantum nature of the photon fields involved.

PACS numbers: 42.50. + q, 42.65.Cq, 32.80.Kf

I. INTRODUCTION

Tunable lasers have been extremely useful tools for the investigation of the interactions of strong resonant electromagnetic fields with atoms and molecules. Considerable theoretical and experimental studies have been done upon two-[1–3] and three-level[14–29] atoms interacting with strong laser fields. Cohen-Tannoudji and Reynaud[17] have considered the fluorescence spectra arising from the interaction of two strong laser fields with a three-level atom in the "V" configuration shown in Fig. 1. The two laser fields a and b, whose frequency modes ω_a and ω_b are highly populated, are in resonance with the two atomic transitions $|1\rangle\leftrightarrow|2\rangle$ and $|1\rangle\leftrightarrow|3\rangle$, respectively. Following the dressed-atom method,[16] which has been developed by the same authors, Cohen-Tannoudji and Reynaud[17] have found that the fluorescence spectra for the $|1\rangle\leftrightarrow|2\rangle$ transition consist of five components[30] at the frequencies $\omega_a, \omega_a \pm (\eta_1^2+\eta_2^2)^{1/2}$ and $\omega_a \pm \frac{1}{2}(\eta_1^2+\eta_2^2)^{1/2}$. Similarly, the fluorescence spectra for the $|1\rangle\leftrightarrow|3\rangle$ transition exhibit five components[30] at the frequencies $\omega_b, \omega_b \pm (\eta_1^2+\eta_2^2)^{1/2}$, and $\omega_b \pm \frac{1}{2}(\eta_1^2+\eta_2^2)^{1/2}$, where η_1 and η_2 are the Rabi frequencies for the transitions $|1\rangle\leftrightarrow|2\rangle$ and $|1\rangle\leftrightarrow|3\rangle$, respectively.

Sharma, Villaverde, and Mavroyannis,[24] hereafter referred to as I, have recently calculated the resonance fluorescence spectra for the system shown in Fig. 1. In I, the fluorescence spectra have been found to consist of either one pair or two pairs or three pairs of sidebands, in addition to the central peak, depending upon the relative strengths of the Rabi frequencies η_1 and η_2 of the two laser fields. In I, use has been made of the Green function formalism, where the hierarchy of the Green functions in question has been truncated by using decoupling approximations of the form

$$\langle\langle \alpha_i^\dagger \alpha_j \beta_a^\dagger \beta_a \beta_a ; \alpha_k^\dagger \alpha_l \rangle\rangle \approx 2\bar{n}_a \langle\langle \alpha_i^\dagger \alpha_j \beta_a ; \alpha_k^\dagger \alpha_l \rangle\rangle, \quad (1)$$

$$\langle\langle \alpha_i^\dagger \alpha_j \beta_b^\dagger \beta_b \beta_b ; \alpha_k^\dagger \alpha_l \rangle\rangle \approx 2\bar{n}_b \langle\langle \alpha_i^\dagger \alpha_j \beta_b ; \alpha_k^\dagger \alpha_l \rangle\rangle, \quad (2)$$

$$\langle\langle \alpha_i^\dagger \alpha_j \beta_a^\dagger \beta_a \beta_b ; \alpha_k^\dagger \alpha_l \rangle\rangle \approx \bar{n}_a \langle\langle \alpha_i^\dagger \alpha_j \beta_b ; \alpha_k^\dagger \alpha_l \rangle\rangle, \quad (3)$$

$$\langle\langle \alpha_i^\dagger \alpha_j \beta_b^\dagger \beta_b \beta_a ; \alpha_k^\dagger \alpha_l \rangle\rangle \approx \bar{n}_b \langle\langle \alpha_i^\dagger \alpha_j \beta_a ; \alpha_k^\dagger \alpha_l \rangle\rangle, \quad (4)$$

where $\bar{n}_a = \langle \beta_a^\dagger \beta_a \rangle$ and $\bar{n}_b = \langle \beta_b^\dagger \beta_b \rangle$ denote the average values of the photon number operators of the laser fields a and b, respectively, and i, j, k, and l enumerate the energy levels of the atom. Here $\beta_a^\dagger(\beta_a)$ and $\beta_b^\dagger(\beta_b)$ are the photon creation (annihilation) operators of the laser fields and $\alpha_i^\dagger$ and α_i are the Fermi creation and annihilation operators describing the electron states of the atom. Unless otherwise stated, the notation here is identical to that of I.

Physically, the decoupling approximations (1) and (2) describe photon–photon correlations from each laser field a and b, namely, $\beta_a^\dagger \beta_a \beta_a \to 2\langle \beta_a^\dagger \beta_a \rangle \beta_a$ and $\beta_b^\dagger \beta_b \beta_b \to 2\langle \beta_b^\dagger \beta_b \rangle \beta_b$, respectively. Photon correlations between the two laser fields of the form $\langle \beta_a^\dagger \beta_b \rangle$ and $\langle \beta_b^\dagger \beta_a \rangle$

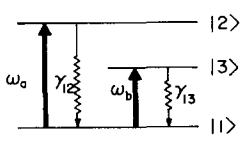

FIG. 1. Energy-level diagram of a three-level atom in the "V" configuration. Thick lines represent laser fields operating between the levels $|1\rangle\leftrightarrow|2\rangle$ and $|1\rangle\leftrightarrow|3\rangle$, respectively. Wiggle lines represent radiative decays.

[a)] Issued as NRCC No. 23275.

 0022-2488/84/092780-08$02.50

have not been considered. The decoupling schemes of the form of Eqs. (3) and (4) imply that the photon operators $n_a = \beta_a^\dagger \beta_a$ and $n_b = \beta_b^\dagger \beta_b$ do not interact with the operators $\alpha_i^\dagger \alpha_j \beta_b$ and $\alpha_i^\dagger \alpha_j \beta_a$ of the system, respectively. Hence, the systems in question are decoupled by replacing the photon number operators n_a and n_b by their average numbers $\bar{n}_a$ and $\bar{n}_b$, respectively. It is obvious that decouplings of the form of Eqs. (3) and (4) do not take into account photon–photon correlations.

In a very recent study on the physical process of hyper-Raman scattering,[31] hereafter referred to as II, it has been shown that the decoupling approximations of the form (1) and (2) are the appropriate ones describing photon–photon correlations from each laser field, respectively, while the Green functions in Eqs. (3) and (4) have been decoupled prematurely. The correct approach is to calculate the equations of motion of the Green functions in question, like those appearing on the left-hand side (lhs) of Eqs. (3) and (4), by means of the Hamiltonian of the system and then the resulting higher-order Green functions should be decoupled in accordance with Eqs. (1) and (2). In this approach, photon–photon correlations have been treated on the same footing. It has been shown in II that the results derived by making use of the premature decouplings of the form of Eqs. (3) and (4) are equivalent to those obtained when both laser fields are treated classically. In II, the neglect of photon–photon correlations between the two laser fields of the form $\langle \beta_a^\dagger \beta_b \rangle$ and $\langle \beta_b^\dagger \beta_a \rangle$ has been justified. Since in I use has been made of Eqs. (1)–(4), hence, photon–photon correlations from each laser field have been taken only partly into account. The purpose of the present is to repeat the calculation for the excitation spectra considered in I by making use only of the decoupling scheme given by Eqs. (1) and (2) and to compare the derived results with those obtained in I as well as with those derived by Cohen-Tannoudji and Reynaud.[17]

The equations of motion for the hierarchy of the Green functions of the system are derived in Sec. II by means of the Hamiltonian (1) of I and Eqs. (1) and (2). It is shown that the whole system is described by a set of fourteen coupled equations instead of those nine equations derived in I. However, the final expression for the Green function $G_{12}(\omega) = \langle\langle \alpha_1^\dagger \alpha_2; \alpha_2 \alpha_1^\dagger \rangle\rangle$ is derived in a closed form, the poles of which can be studied analytically while the corresponding expression in I requires the use of a computer. The excitation spectra are considered in Sec. III, where the spectral function of the system is derived and discussed. The results of numerical calculations are graphically presented and compared with those derived in I. In the Appendix, expressions for the Green function $G_{12}(\omega)$ are derived when both laser fields are treated classically and when the fields are quantized but use is made of the decoupling scheme given by Eqs. (3) and (4), namely, when photon–photon correlations are discarded.

II. EQUATIONS OF MOTION FOR THE GREEN FUNCTIONS

We shall calculate here the equation of motion for the Green function $G_{12}(\omega) = \langle\langle \alpha_1^\dagger \alpha_2; \alpha_2^\dagger \alpha_1 \rangle\rangle$ by means of the Hamiltonian (1) of I. To avoid repetitions, we adopt exactly the same system and notation as in I, hence, the reader is referred to I for details. Thus equations from I will be referred to I and will not be repeated here. Only new results and those differing from I will be quoted here.

The equations of motion for the Green functions $G_{12}(\omega)$, $F_{21a}(\omega) = \langle\langle (n_2 - n_1)\beta_a; \alpha_2^\dagger \alpha_1 \rangle\rangle$, and $G_{32b}(\omega) = \langle\langle \alpha_3^\dagger \alpha_2 \beta_b; \alpha_2^\dagger \alpha_1 \rangle\rangle$ have been derived in I and are given by Eqs. (6), (8), and (9) of I, respectively. Instead of Eqs. (10)–(14) of I, we derive the corresponding set of equations

$$(\omega - \omega_a - \tfrac{1}{2}\gamma_{12})\langle\langle \alpha_1^\dagger \alpha_2 n_a; \alpha_2^\dagger \alpha_1 \rangle\rangle = [(\bar{n}_1 - \bar{n}_2)/2\pi]\bar{n}_a + i\eta_1 \sqrt{\bar{n}_a} F_{21a}(\omega) + \tfrac{1}{2} i\omega_p \sqrt{f_2} G_{32bn_a}(\omega), \tag{5}$$

$$(\omega - \omega_a - \tfrac{1}{2}\gamma_{12})\langle\langle \alpha_2^\dagger \alpha_1 \beta_a \beta_a; \alpha_2^\dagger \alpha_1 \rangle\rangle = i\eta_1 \sqrt{\bar{n}_a} F_{21a}(\omega) + i\omega_p \sqrt{f_2} G_{23b^\dagger aa}(\omega), \tag{6}$$

$$(\omega - \omega_a - \tfrac{1}{2}\gamma_{13})\langle\langle \alpha_1^\dagger \alpha_2 \beta_b^\dagger \beta_a; \alpha_2^\dagger \alpha_1 \rangle\rangle = \tfrac{1}{2}\omega_p \sqrt{f_2} F_{31an_b}(\omega) + \tfrac{1}{2} i\omega_p \sqrt{f_1} G_{23b^\dagger aa}(\omega), \tag{7}$$

$$(\omega - \omega_a - \tfrac{1}{2}\gamma_{13})\langle\langle \alpha_3^\dagger a_1 \beta_a \beta_b; \alpha_2^\dagger \alpha_1 \rangle\rangle = \tfrac{1}{2} i\omega_p \sqrt{f_1} G_{32bn_a}(\omega) + \tfrac{1}{2} i\omega_p \sqrt{f_2} F_{31an_b}(\omega), \tag{8}$$

$$(\omega - \omega_a - \tfrac{1}{2}\gamma_{12})\langle\langle \alpha_1^\dagger \alpha_2 n_b; \alpha_2^\dagger \alpha_1 \rangle\rangle = [(\bar{n}_1 - \bar{n}_2)/2\pi]\bar{n}_b + \tfrac{1}{2} i\omega_p \sqrt{f_1} F_{21an_b}(\omega) + i\eta_2 \sqrt{\bar{n}_b} G_{32b}(\omega), \tag{9}$$

where $n_a = \beta_a^\dagger \beta_a$ and $n_b = \beta_b^\dagger \beta_b$ are the photon number operators of the two laser fields, respectively, while η_1 and η_2 are the corresponding Rabi frequencies defined by

$$\eta_1^2 = \omega_p^2 f_1 \bar{n}_a, \quad n_2^2 = \omega_p^2 f_2 \bar{n}_b.$$

The rhs of Eqs. (5)–(9) differ from the corresponding Eqs. (10)–(14) of I in that the Green functions

$$G_{32bn_a}(\omega) = \langle\langle \alpha_3^\dagger \alpha_2 \beta_b n_a; \alpha_2^\dagger \alpha_1 \rangle\rangle,$$

$$F_{31an_b}(\omega) = \langle\langle (n_3 - n_1)\beta_a n_b; \alpha_2^\dagger \alpha_1 \rangle\rangle,$$

and

$$F_{21bn_a}(\omega) = \langle\langle (n_2 - n_1)\beta_b n_a; \alpha_2^\dagger \alpha_1 \rangle\rangle$$

have been decoupled in I by applying the decoupling approximations described by Eqs. (15) and (17) of I, which are of form of Eqs. (3) and (4). The Green function $G_{23b^\dagger aa}(\omega) = \langle\langle \alpha_2^\dagger \alpha_3 \beta_b^\dagger \beta_a \beta_a; \alpha_2^\dagger \alpha_1 \rangle\rangle$ appearing on the rhs of Eqs. (6) and (7) is also missing in the corresponding Eqs. (11) and (12) of I. However, such an omission is consistent with the use of the decoupling scheme given by Eqs. (15) and (17) of I. In the present study, instead of decoupling the Green functions in question, we derive their equations of motion:

$$(\omega - \omega_a - \tfrac{1}{2}\gamma_{12} - \tfrac{1}{2}\gamma_{13})G_{32bn_a}(\omega) = -i\eta_1 \sqrt{\bar{n}_a}\langle\langle \alpha_3^\dagger \alpha_1 \beta_a \beta_b; \alpha_2^\dagger \alpha_1 \rangle\rangle - \tfrac{1}{2} i\omega_p \sqrt{f_2} G_{12n_an_b}(\omega), \tag{10}$$

$$(\omega - \omega_a - \gamma_{13})F_{31an_b}(\omega) = -\tfrac{1}{2} i\omega_p \sqrt{f_1}\,[G_{12n_an_b}(\omega) + G_{21aan_b}(\omega)] - 2i\eta_2 \sqrt{\bar{n}_b}\langle\langle \alpha_1^\dagger \alpha_3 \beta_b^\dagger \beta_a; \alpha_2^\dagger \alpha_1 \rangle\rangle$$

$$+\langle\langle\alpha_3^\dagger\alpha_1\beta_a\beta_b;\alpha_2^\dagger\alpha_1\rangle\rangle), \tag{11}$$

$$(\omega-\omega_a-\gamma_{12})F_{21an_b}(\omega)$$
$$=-\tfrac{1}{2}i\omega_p\sqrt{f_1}\,[G_{12n_an_b}(\omega)+G_{21aan_b}(\omega)]$$
$$-i\eta_2\sqrt{\bar{n}_b}(\langle\langle\alpha_1^\dagger\alpha_3\beta_a^\dagger\beta_b;\alpha_2^\dagger\alpha_1\rangle\rangle$$
$$+\langle\langle\alpha_3^\dagger\alpha_1\beta_a\beta_b;\alpha_2^\dagger\alpha_1\rangle\rangle), \tag{12}$$

$$(\omega-\omega_a-\tfrac{1}{2}\gamma_{12}-\tfrac{1}{2}\gamma_{13})G_{23b^\dagger aa}(\omega)$$
$$=-i\eta_1\sqrt{\bar{n}_a}\langle\langle\alpha_1^\dagger\alpha_3\beta_b^\dagger\beta_a;\alpha_2^\dagger\alpha_1\rangle\rangle$$
$$-\tfrac{1}{2}i\omega_p\sqrt{f_2}G_{21aan_b}(\omega), \tag{13}$$

$$(\omega-\omega_a-\tfrac{1}{2}\gamma_{12})G_{12n_an_b}(\omega)$$
$$=[(\bar{n}_1-\bar{n}_2)/2\pi]\bar{n}_a\bar{n}_b+i\eta_1\sqrt{\bar{n}_a}F_{21an_b}(\omega)$$
$$+i\eta_2\sqrt{\bar{n}_b}G_{32bn_a}(\omega), \tag{14}$$

$$(\omega-\omega_a-\tfrac{1}{2}\gamma_{12})G_{21aan_b}(\omega)$$
$$=i\eta_1\sqrt{\bar{n}_a}F_{21an_b}(\omega)+i\eta_2\sqrt{\bar{n}_b}G_{23b^\dagger aa}(\omega), \tag{15}$$

where $G_{12n_an_b}(\omega)=\langle\langle\alpha_1^\dagger\alpha_2n_an_b;\alpha_2^\dagger\alpha_1\rangle\rangle$ and $G_{21aan_b}(\omega)=\langle\langle\alpha_2^\dagger\alpha_1\beta_a\beta_an_b;\alpha_2^\dagger\alpha_1\rangle\rangle$. In deriving Eqs. (10)–(15), only decouplings of the form of Eqs. (1) and (2) have been performed.

Thus in I, the system is described by the set of nine coupled equations, namely, Eqs. (6), (8)–(14), and (18) of I, while here we have a set of 14 coupled equations, Eqs. (6), (8), and (9) of I and Eqs. (5)–(15). However, the form of the fourteen coupled equations is such that they can be solved much more easily than those in I. From the physical point of view, the Green functions $G_{32bn_a}(\omega)$, $F_{31an_b}(\omega)$, $F_{21an_b}(\omega)$, and $G_{21aan_b}(\omega)$, which have been prematurely decoupled in I, describe physical processes where the photon density operators n_a and n_b interact with the operators $\alpha_3^\dagger\alpha_2\beta_b$ and $(n_3-n_1)\beta_a$, $(n_2-n_1)\beta_a$, $\alpha_2^\dagger\alpha_1\beta_a\beta_a$, respectively. The Green function $G_{12n_an_b}(\omega)$ describes the physical process where the electronic transition $|1\rangle\rightarrow|2\rangle$ takes place in the presence of the photon density operators n_an_b of both laser fields. Hence, the set of the 14 coupled equations describe all the possible physical processes which contribute to the spectrum arising from the $|1\rangle\rightarrow|2\rangle$ electronic transition in the limit of high photon densities of both laser fields.

We use the notation of I for $X=(\omega-\omega_a)/\gamma$, where $\gamma=\mathrm{Im}\,\gamma_{12}\approx\mathrm{Im}\,\gamma_{13}$ and we define

$$d_0=X+\tfrac{1}{2}i,\quad d=X+i, \tag{16}$$

$$\eta_a=\eta_1/\gamma,\quad \eta_b=\eta_2/\gamma,\quad \eta^2=(\eta_a^2+\eta_b^2)/2, \tag{17}$$

X being the reduced frequency and η_a and η_b the reduced Rabi frequencies of the laser fields a and b, respectively. Inspection of Eqs. (5)–(15) indicates that because of the inclusion of Eq. (14) for the Green function $G_{12n_an_b}(\omega)$, the expressions for the Green functions $G_{32bn_a}(\omega)$,$F_{31an_b}(\omega)$,$F_{21an_b}(\omega)$, and $G_{21aan_b}(\omega)$ can be easily derived in a closed form. Using Eqs. (16) and (17) we obtain from Eqs. (8) and (9) of I and Eqs. (5)–(15) the following expressions:

$$F_{21an_b}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)(-i\eta_a\sqrt{\bar{n}_a})\bar{n}_b}{2\pi\gamma(dd_0-4\eta^2)}\left(1-\frac{3\eta_b^2/4}{dd_0-\eta^2}\right), \tag{18}$$

$$F_{31an_b}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)(i\eta_a\sqrt{\bar{n}_a})\bar{n}_b}{2\pi(dd_0-4\eta^2)}\left(\frac{1}{2}+\frac{3\eta_b^2/4}{dd_0-\eta^2}\right), \tag{19}$$

$$G_{32bn_a}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)(i\eta_b\sqrt{\bar{n}_b})\bar{n}_a}{2\pi\gamma(dd_0-\eta^2)}\left(\frac{1}{2}+\frac{3\eta_a^2/4}{dd_0-4\eta^2}\right), \tag{20}$$

$$G_{23b^\dagger aa}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)(-i\eta_b\sqrt{\bar{n}_b})\bar{n}_a}{2\pi\gamma(dd_0-\eta^2)(dd_0-4\eta^2)}\left(\frac{3}{4}\eta_a^2\right), \tag{21}$$

$$G_{21aan_b}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)\eta_a^2(\bar{n}_a\bar{n}_b)}{2\pi\gamma d_0(dd_0-4\eta^2)}. \tag{22}$$

Using Eqs. (18)–(22), we derive the expressions for $F_{21a}(\omega)$ and $G_{32b}(\omega)$ as

$$F_{21a}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)(-i\eta_a\sqrt{\bar{n}_a})}{2\pi\gamma(dd_0-2\eta_a^2)}$$
$$\times\left[1+\frac{\eta_b^2(\eta^2-2\eta_a^2)/8\eta^2}{dd_0-\eta^2}+\frac{\eta_b^2(2\eta^2+\eta_a^2)/4\eta^2}{dd_0-4\eta^2}\right], \tag{23}$$

$$G_{32b}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)(-i\eta_b\sqrt{\bar{n}_b})}{4\pi\gamma(dd_0-\frac{1}{2}\eta_b^2)}$$
$$\times\left[1+\frac{\eta_a^2\,\eta_b^2/8}{dd_0-\eta^2}+\frac{\eta_a^2(6\eta^2+\eta_a^2)/8\eta^2}{dd_0-4\eta^2}\right]. \tag{24}$$

Substitution of Eqs. (23) and (24) into Eq. (6) of I yields

$$G_{12}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)}{2\pi\gamma d_0}\left[1+\frac{\eta_a^2/4}{dd_0-2\eta_a^2}+\frac{\eta_b^2/8}{dd_0-\frac{1}{2}\eta_b^2}\right.$$
$$\left.+\frac{\eta_b^2/8}{dd_0-\eta^2}+\frac{\eta_a^2/4}{dd_0-4\eta^2}\right]. \tag{25}$$

Equation (25) for $G_{12}(\omega)$ is the required expression describing the $|1\rangle\rightarrow|2\rangle$ electronic transition in the limit of high photon densities of both laser fields. The first term describes the central peak at $X=0$ while the remaining ones represent pairs of sidebands peaked at the reduced frequencies $X=\pm\eta_a\sqrt{2}$, $\pm\eta_b/\sqrt{2}$, $\pm\eta$, and $\pm2\eta$, respectively. In the absence of the laser field b, namely, in the limit when $\eta_b\rightarrow0$, the third and fourth terms in Eq. (25) vanish while the last term becomes equal to the second one with the result

$$G_{12}^{(a)}(\omega)=\frac{(\bar{n}_1-\bar{n}_2)}{2\pi\gamma d_0}\left[1+\frac{\eta_a^2/2}{dd_0-2\eta_a^2}\right], \tag{26}$$

which is identical to Eq. (29) of I and describes the spectra of a two-level system. Thus the laser field b results in splitting the last term in Eq. (26) into two pairs of sidebands described by the second and last terms in Eq. (25), which appear at the reduced frequencies $X=\pm\eta_a\sqrt{2}$ and $\pm2\eta$, respectively. The third and fourth terms in Eq. (25) describe two pairs of sidebands, which are induced by the laser field b, and they are centered at $X=\pm\eta_b/\sqrt{2}$ and $\pm\eta$, respectively. The relative intensity and the position of the pair of peaks at $X=\pm\eta$ depends on the Rabi frequency η_a as well. Equation (25) for $G_{12}(\omega)$ is much simpler than that of Eq. (25) of I which can be studied only through computation. The expression (25) for $G_{12}(\omega)$ will be used in the next section to discuss the excitation spectra.

III. EXCITATION SPECTRA

To study the excitation spectra, we derive the imaginary part of Eq. (25) as

$$
\begin{aligned}
-2\,\mathrm{Im}\,G_{12}(\omega) = \frac{(\bar{n}_1-\bar{n}_2)}{2\pi\gamma}\Bigg\{&\left(1-\frac{\eta_b^2}{8\eta^2}\right)\frac{\frac{1}{2}}{X^2+\frac{1}{4}}+\frac{1}{8(1+1/32\eta_a^2)}\left[\frac{\frac{3}{4}-(X-\eta_a\sqrt{2})(\sqrt{2}/8\eta_a)}{(X-\eta_a\sqrt{2})^2+\frac{9}{16}}+\frac{\frac{3}{4}+(X+\eta_a\sqrt{2})(\sqrt{2}/8\eta_a)}{(X+\eta_a/\sqrt{2})^2+\frac{9}{16}}\right]\\
&+\frac{1}{4(1+1/8\eta_b^2)}\left[\frac{\frac{3}{4}-(X-\eta_b/\sqrt{2})(\sqrt{2}/4\eta_b)}{(X-\eta_b/\sqrt{2})^2+\frac{9}{16}}+\frac{\frac{3}{4}+(X+\eta_b/\sqrt{2})(\sqrt{2}/4\eta_b)}{(X+\eta_b/\sqrt{2})^2+\frac{9}{16}}\right]\\
&+\frac{\eta_b^2}{8\eta^2(1+1/16\eta^2)}\left[\frac{\frac{3}{4}-(X-\eta)(1/4\eta)}{(X-\eta)^2+\frac{9}{16}}+\frac{\frac{3}{4}+(X+\eta)(1/4\eta)}{(X+\eta)^2+\frac{9}{16}}\right]\\
&+\frac{\eta_a^2}{16\eta^2(1+1/64\eta^2)}\left[\frac{\frac{3}{4}-(X-2\eta)(1/8\eta)}{(X-2\eta)^2+\frac{9}{16}}+\frac{\frac{3}{4}+(X+2\eta)(1/8\eta)}{(X+2\eta)^2+\frac{9}{16}}\right]\Bigg\}. \qquad (27)
\end{aligned}
$$

The spectral function (27) describes the excitation spectra for the electronic transition $|1\rangle \rightarrow |2\rangle$ in the limit of high photon densities of both laser fields. The first term describes the central peak, which is a Lorentzian line peaked at $X=0$, and has a spectral width of the order of $\gamma/2$. The intensity of the central peak depends on the factor $1-(\eta_b^2/8\eta^2)$, where $\eta_b^2/8\eta^2$ is due to the presence of the laser field b. The remaining terms in Eq. (27) describe four pairs of Lorentzian peaks centered at $X=\pm\eta_a\sqrt{2}$, $\pm\eta_b/\sqrt{2}$, $\pm\eta$, and $\pm 2\eta$ and having spectral widths of the order of $3\gamma/4$. The spectral function (27) implies that in the limit of high photon densities and strong fields where $\eta_a^2 \gg 1$ and $\eta_b^2 \gg 1$, the intensities of the pairs of sidebands at $X=\pm\eta_a\sqrt{2}$ and $\pm\eta_b/\sqrt{2}$ are constant and independent of the values of η_a and η_b while the intensities of those at $X=\pm\eta$ and $\pm 2\eta$ depend on the values of the ratios η_b^2/η^2 and η_a^2/η^2, respectively. In the limiting case when both Rabi frequencies are equal, i.e., when $\eta_a=\eta_b=\eta$ then for $\eta^2>1$, the maximum intensities of the peaks described by Eq. (27) are independent of η provided that η is large enough so that overlap between the peaks of the system can be avoided.

Numerical results derived from Eq. (27) are illustrated in Figs. 2–4 for selective values of the reduced Rabi frequencies η_a and η_b. The relative intensity $I(\omega)$, which is defined as

$$I(\omega) = -2\pi\gamma\,\mathrm{Im}\,G_{12}(\omega)/(\bar{n}_1-\bar{n}_2), \qquad (28)$$

is plotted in Figs. 2–4 versus the relative frequency $X=(\omega-\omega_a)/\gamma$. The expression (28) for the relative intensity $I(\omega)$, apart from a constant factor, defines the absorption coefficient describing the physical processes of the system under investigation.

As stated in I, Eq. (29) of I for $-2\,\mathrm{Im}\,G_{12}(\omega)$ cannot be studied analytically and the excitation spectra are obtained only by numerical computation using a computer. Therefore, comparison between Eqs. (27) and (29) of I has to be done graphically. In order to compare graphically the numerical results derived from Eq. (27) with those obtained in I, all the peaks in Figs. 2–4 are normalized with respect to the relative intensity of the central peak ($X=0$) taken as equal to 1. In Figs. 2–4, solid lines refer to the results derived from Eq. (27) while dashed lines refer to those obtained in I.

Figures 2(a) and 2(b) illustrate the spectra for values $\eta=\eta_a=\eta_b=5$ and 20, respectively. The peaks depicted by dashed lines in Figs. 2(a) and 2(b) are taken from Figs. 2(b) and 2(c) of I, respectively. In Fig. 2(a), the value of $\eta=5$ is not large enough to separate the peaks so that there is an overlap in the intensities of the peaks, while for $\eta=20$ in Fig. 2(b), the peaks are well separated and their intensities remain constant and independent of η for values of $\eta \geq 20$. There are four pairs of sidebands, which are peaked at the frequencies $X=\pm\eta/\sqrt{2}$, $\pm\eta$, $\pm\eta\sqrt{2}$, and $\pm 2\eta$ as illustrated by the solid lines while the dashed lines give only two pairs of sidebands. In Fig. 2(b), the heights of the sidebands

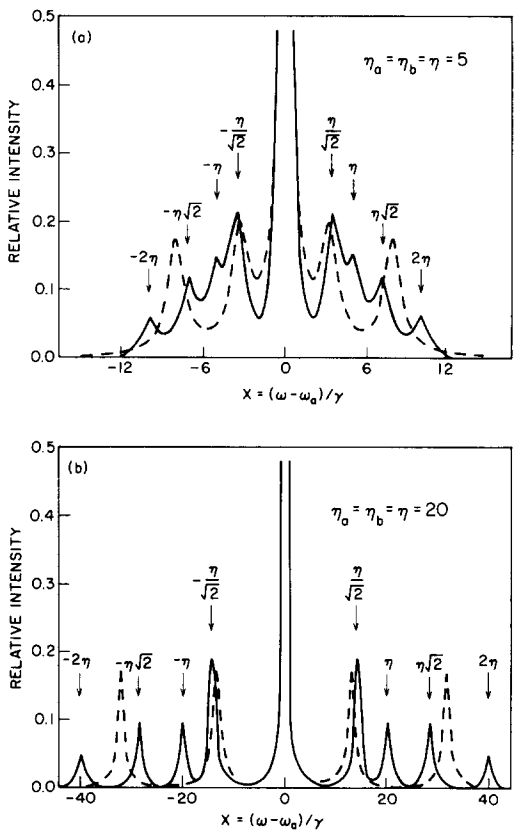

FIG. 2. The relative intensity of the fluorescent light is plotted versus $X=(\omega-\omega_a)/\gamma$ for different values of η. (a) $\eta=\eta_a=\eta_b=5$ and (b) $\eta=\eta_a=\eta_b=20$. Solid lines are derived from Eq. (27) while dashed lines are obtained from Figs. 2(b) and 2(c) of the paper by Sharma *et al.*[21]

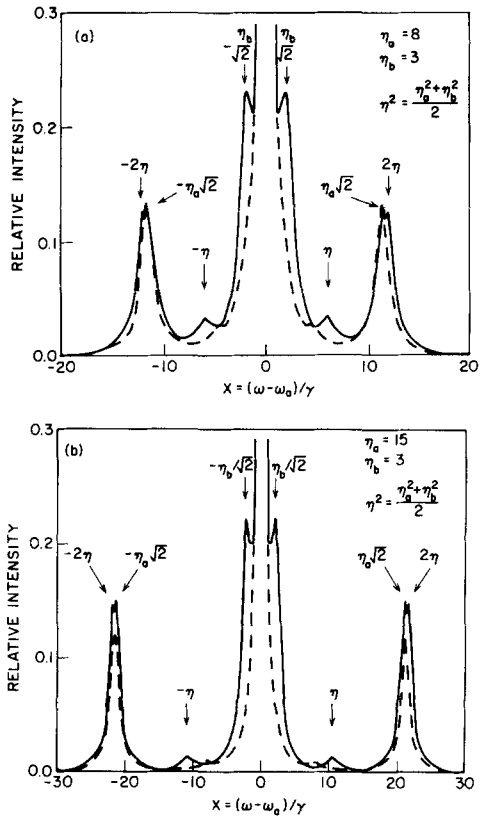

FIG. 3. As in Fig. 2 but $\eta_b = 3$ is kept constant while η_a is varied. (a) $\eta_a = 8$ and (b) $\eta_a = 15$. Solid lines are derived from Eq. (27) while dashed lines are obtained from Figs. 3(b) and 3(c) of the paper by Sharma *et al.*[24]

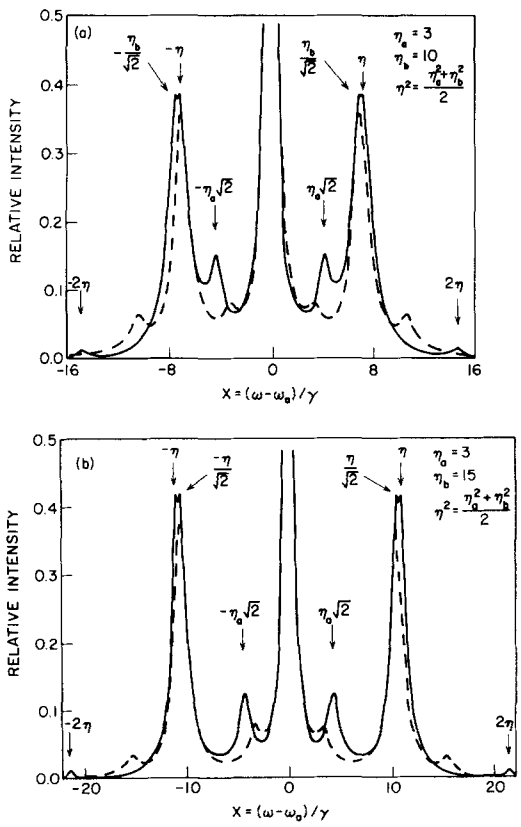

FIG. 4. As in Fig. 2 but $\eta_a = 3$ is kept constant while η_b is varied. (a) $\eta_b = 10$ and (b) $\eta_b = 15$. Solid lines are derived from Eq. (27) while dashed lines are obtained from Figs. 4(b) and 4(c) of the paper by Sharma *et al.*[24]

at the frequencies $X = \pm\eta/\sqrt{2}, \pm\eta, \pm\eta\sqrt{2}$, and $\pm 2\eta$ are found to be $\frac{1}{5}, \frac{1}{10}, \frac{1}{10}$, and $\frac{1}{20}$ of that of the central peak while the corresponding height for the sidebands given by the dashed lines is $\frac{1}{8}$. The linewidths[32] of all sidebands are equal to $3\gamma/4$.

In Figs. 3(a) and 3(b), we consider the case where the reduced Rabi frequency $\eta_b = 3$ is kept constant while η_a takes the values of $\eta_a = 8$ and 15, respectively. The calculated results from Eq. (27) are depicted in Figs. 3(a) and 3(b) by solid lines while dashed lines indicate the peaks obtained from Figs. 3(b) and 3(c) of I, respectively. Figures 3(a) and 3(b) indicate that as the value of η_a increases for a given value of η_b, the two pairs of peaks at $X = \pm\eta_a\sqrt{2}$ and $\pm 2\eta$ coalesce rapidly since $\eta_a\sqrt{2}$ and 2η are differing only by a small amount. Equation (27) implies that the relative intensity of the pair of sidebands at $X = \pm\eta$ depends on the value of $(\eta_b/\eta)^2$, which goes to zero as η_a increases, and hence, the pair of sidebands at $X = \pm\eta$ vanishes. Thus for a given value of η_b and for large values of η_a, $\eta_a \gg \eta_b$, there will be only two pairs of sidebands peaked at $X = \pm\eta_a\sqrt{2}$ and $\pm\eta_b/\sqrt{2}$, provided that η_b is large enough so that the peaks at $X = \pm\eta_b/\sqrt{2}$ will be well separated from the central peak at $X = 0$. In this case, the sidebands at $X = \pm\eta_a\sqrt{2}$ become identical to those of the fluorescence spectrum of a two-level system.[1-13]

The fluorescence spectra for $\eta_b > \eta_a$ are illustrated in Figs. 4(a) and 4(b), where $\eta_a = 3$ is kept constant while η_b takes the values of $\eta_b = 10$ and 15, respectively. In Figs. 4(a) and 4(b), the spectra depicted by dashed lines are obtained from Figs. 4(b) and 4(c) of I. As η_b increases for a given value of η_a, the values of $\eta_b/\sqrt{2}$ and η differ by a small amount, and therefore, the two pairs of sidebands at $X = \pm\eta_b/\sqrt{2}$ and $\pm\eta$ tend to coalesce. The relative intensity of the pair of sidebands at $X = \pm 2\eta$ depends on the value of $(\eta_a/\eta)^2$, which becomes very small for large η_b. Thus in the case, the intensity of the pair of sidebands at $X = \pm\eta_a\sqrt{2}$ remains constant while that of the pair at $X = \pm 2\eta$ diminishes as η_b increases. The two pairs of peaks at $X = \pm\eta_b/\sqrt{2}$ and $\pm\eta$ will coalesce for large values of η_b. Hence, the case of $\eta_b > \eta_a$ describes interference spectra, which are induced by the laser field b into the spectra of the $|1\rangle \leftrightarrow |2\rangle$ transition. Figures 2–4 indicate that there are characteristic differences between the results derived from Eq. (27) (solid lines) and those obtained from I (dashed lines). The spectral function given by Eq. (29) of I is a complicated polynomial of the parameters involved and it can be studied accurately only numerically through a computer and, hence, the comparison between Eqs. (27) and (29) of I has to be done graphically.

The fluorescence spectra due to the transition $|1\rangle \leftrightarrow |3\rangle$ are described by the imaginary part of the Green function $G_{13}(\omega) = \langle\langle \alpha_1^\dagger \alpha_3; \alpha_3^\dagger \alpha_1 \rangle\rangle$, which can be obtained from Eq. (27) if ω_a and η_a are replaced everywhere by ω_b and η_b ($\omega_a \leftrightarrow \omega_b, \eta_a \leftrightarrow \eta_b$), respectively, and $\bar{n}_2$ is replaced by $\bar{n}_3$ ($\bar{n}_2 \leftrightarrow \bar{n}_3$) as well. The fluorescence spectra for the transition $|1\rangle \leftrightarrow |3\rangle$ consist of nine Lorentzian lines peaked at the frequencies ω_b, $\omega_b \pm \eta_2\sqrt{2}$, $\omega_b \pm \eta_1/\sqrt{2}$, $\omega_b \pm \eta\gamma$, and $\omega_b \pm 2\eta\gamma$, respectively. The spectral width for the central peak is $\frac{1}{2}\gamma$ while those for the sidebands are equal to $3\gamma/4$.

IV. SUMMARY

We have calculated the resonant fluorescent spectra arising from the interaction of a three-level atom with two strong laser fields as shown in Fig. 1. Using the Hamiltonian (1) of I, the Green function formalism and the decoupling approximations defined by Eqs. (1) and (2), the expression (25) for the Green function $G_{12}(\omega)$ is derived, which describes the electronic transition $|1\rangle \leftrightarrow |2\rangle$ in the limit of high photon densities.

Using the expression (25), the spectral function (27) is derived, which describes the excitation spectra for the $|1\rangle \leftrightarrow |2\rangle$ transition under consideration. Apart from the central peak, the spectra consist of four pairs of Lorentzian lines peaked at the reduced frequencies $X = \pm \eta_a\sqrt{2}$, $\pm \eta_b/\sqrt{2}$, $\pm \eta$, and $\pm 2\eta$, and having spectral widths of the order of $3\gamma/4$. Numerical results for the relative intensities are graphically presented in Figs. 2–4 for different values of the parameters involved and compared with those derived in I.

The spectral function for the $|1\rangle \leftrightarrow |3\rangle$ transition can be obtained from Eq. (27) after the replacements $\omega_a \leftrightarrow \omega_b$, $\eta_a \leftrightarrow \eta_b$, and $\bar{n}_2 \leftrightarrow \bar{n}_3$ have been made. The sidebands due to the transition $|1\rangle \leftrightarrow |3\rangle$ are described by four pairs of Lorentzian lines centered at the frequencies $\omega_b \pm \eta_2\sqrt{2}$, $\omega_b \pm \eta_1/\sqrt{2}$, $\omega_b \pm \eta\gamma$, and $\omega_b \pm 2\eta\gamma$, respectively, and having spectral widths of the order of $3\gamma/4$.

In the Appendix, the expression (A12) for the Green function $G_{12}(\omega)$ is derived when both laser fields are treated classically. Also, the expression (A13) for $G_{12}(\omega)$ is calculated when both laser fields are quantized but use has been made of the decoupling approximations given by Eqs. (3) and (4), where photon–photon correlations from each laser field are neglected. Comparison between Eqs. (A12) and (A13) indicates that both results become identical only when the classical expressions for the Rabi frequencies g_a and g_b are replaced by their quantized field counterparts η_a and η_b, respectively, and vice versa. The excitation spectra described by the expression (A13) for $G_{12}(\omega)$ are identical to those derived by Cohen-Tannoudji and Reynaud[30] by means of the dress-atom method.[16,17]

Expressions (25) and (A13) for $G_{12}(\omega)$ have been calculated with and without photon–photon correlations from each laser field taken into account, respectively. Comparison between them implies that the second and third terms in Eq. (25), which do not appear in Eq. (A13), are due entirely to photon–photon correlations arising from each of the laser fields a and b, respectively; this is in agreement with the Dirac's definition of interference of photons.[33] The last two terms in Eq. (25) and (A13) may be attributed to cooperative effects, where both laser fields contribute for the position of the two pairs of peaks in question while the relative intensities of these peaks depend upon the relative strength of the Rabi frequencies involved.

The merits of the methods in question can be described as follows: In problems where there is evidence that the laser field is very strong to the extent that any fluctuations due to photon–photon interactions are very small compared to that of the classical field, the treatment of the radiation field as a classical entity is a good approximation. In this case the dress-atom approach[16,17] is the proper method and, for our problem, the excitation spectra derived by Cohen-Tannoudji[30] are applicable. The present method reveals the appearance of new sidebands in addition to the ones derived when the laser field is treated classically. These new sidebands are due entirely to photon–photon correlation effects where photons correlate while obeying Bose statistics as it is easily seen from Eqs. (1) and (2). Therefore, the appearance of the new sidebands reveals the boson character of the photons for the laser fields in question. Thus, the merit of the present method is that it describes both the classical results as well as the quantized field ones which represent the quantum nature of the photon that is lost in the classical treatment; the latter effects are very important in photon correlation experiments.[7]

As mentioned in the Introduction, the decoupling schemes determined by Eqs. (1) and (2) describe photon correlations where photons from each laser field correlate according to Bose statistics while those given by Eqs. (3) and (4) indicate the lack of any photon correlations. It is shown in Sec. 2 of the Appendix that when only decoupling approximations of the form of Eqs. (3) and (4) are used then are the final results identical to those obtained when the laser fields are treated classically. In order to truncate the hierarchy of the Green functions in I, use has been made of Eqs. (1)–(4). In the present study, only decoupling approximations given by Eqs. (1) and (2) have been used while Green's functions in the form of those appearing on the lhs of Eqs. (3) and (4) have been calculated to higher order by means of the Hamiltonian (1) instead of being decoupled. As was shown in Sec. II, the system is described by a set of 14 coupled equations while in I the corresponding set consists of nine coupled equations. However, the final result determined by Eq. (25) or by Eq. (27) is derived in a closed form and is much simpler to study than the corresponding Eq. (29) of I. Thus by making use of Eqs. (1) and (2), both laser fields have been treated on the same footing as far as photon correlations are concerned and, therefore, the present method may be considered as an improvement to that of I, where some Green's functions have been premturely decoupled.

APPENDIX: EXPRESSIONS FOR THE GREEN FUNCTION

We shall calculate here expressions for the Green function $G_{12}(\omega)$: (i) when both laser fields are treated classically, and (ii) when photon–photon correlations from each laser field are ignored.

1. Classical field treatment

To treat the laser fields a and b classically, we consider the Hamiltonian (1) of I, where the terms $\omega_a\beta_a^\dagger\beta_a$ and $\omega_b\beta_b^\dagger\beta_b$ are absent while the fifth and sixth terms are replaced by

$$(i/2)g_1(\alpha_1^\dagger\alpha_2 e^{-i\omega_a t} - \alpha_2^\dagger\alpha_1 e^{i\omega_a t}) + (i/2)g_2(\alpha_1^\dagger\alpha_3 e^{-i\omega_b t} - \alpha_3^\dagger\alpha_1 e^{i\omega_b t}), \quad \text{(A1)}$$

where g_1 and g_2 are the corresponding classical counterparts

for the Rabi frequencies η_1 and η_2, respectively. The two terms in (A1) are analogous to those given by Eq. (A6) of II.[34] Using this Hamiltonian, we derive the following equations of motion:

$$d_0 G_{12}(\omega) = (\bar{n}_1 - \bar{n}_2)/2\pi\gamma + \tfrac{1}{2}ig_a F^c_{21a}(\omega) + \tfrac{1}{2}ig_b G_{32b}(\omega), \quad \text{(A2)}$$

$$dF^c_{21a}(\omega) = -ig_a\left[G_{12}(\omega) + G^c_{21aa}(\omega)\right] - \tfrac{1}{2}ig_b\left[G^c_{13b^\dagger a}(\omega) + G^c_{31ab}(\omega)\right], \quad \text{(A3)}$$

$$dG^c_{32b}(\omega) = -\tfrac{1}{2}ig_a G^c_{31ab}(\omega) - \tfrac{1}{2}g_b G_{12}(\omega), \quad \text{(A4)}$$

$$d_0 G^c_{21aa}(\omega) = \tfrac{1}{2}ig_a F^c_{21a}(\omega) + \tfrac{1}{2}ig_b G^c_{23b^\dagger aa}(\omega), \quad \text{(A5)}$$

$$d_0 G^c_{13b^\dagger a}(\omega) = \tfrac{1}{2}ig_b F^c_{31a}(\omega) + \tfrac{1}{2}ig_a G^c_{23b^\dagger aa}(\omega), \quad \text{(A6)}$$

$$d_0 G^c_{31ab}(\omega) = \tfrac{1}{2}ig_a G^c_{32b}(\omega) + \tfrac{1}{2}ig_b F^c_{31a}(\omega), \quad \text{(A7)}$$

$$dF^c_{31a}(\omega) = -\tfrac{1}{2}ig_a\left[G_{12}(\omega) + G^c_{21aa}(\omega)\right] - ig_b\left[G^c_{13b^\dagger a}(\omega) + G^c_{31ab}(\omega)\right], \quad \text{(A8)}$$

$$dG^c_{23b^\dagger aa}(\omega) = -\tfrac{1}{2}ig_a G^c_{13b^\dagger a}(\omega) - \tfrac{1}{2}ig_b G^c_{21aa}(\omega), \quad \text{(A9)}$$

where

$$g_a = g_1/\gamma, \quad g_b = g_2/\gamma. \quad \text{(A10)}$$

The superscript c on the G^c's designates classical Green functions with respect to the laser fields, which can be obtained from their corresponding field-quantized counterparts after replacing the operators $\beta_a(\beta^\dagger_a)$ and $\beta_b(\beta^\dagger_b)$ by the exponentials $e^{-i\omega_a t}(e^{i\omega_a t})$ and $e^{-i\omega_b t}(e^{i\omega_b t})$, respectively. For instance

$$F^c_{21a}(\omega) = \langle\langle (n_2 - n_1)e^{-i\omega_a t}; \alpha^\dagger_2\alpha_1\rangle\rangle,$$
$$G^c_{32b}(\omega) = \langle\langle \alpha^\dagger_3\alpha_2 e^{-\omega_b t}; \alpha^\dagger_2\alpha_1\rangle\rangle. \quad \text{(A11)}$$

In deriving Eqs. (A2)–(A9), we have taken as before, $\gamma = \mathrm{Im}\,\gamma_{12} \approx \mathrm{Im}\gamma_{13}$ and the expressions for d and d_0 are given by Eq. (16).

Solving the set of Eqs. (A2)–(A9), we obtain

$$G_{12}(\omega) = \frac{(\bar{n}_1 - \bar{n}_2)}{2\pi\gamma d_0}\left[1 + \frac{g_b^2/4}{dd_0 - \frac{1}{4}(g_a^2 + g_b^2)} + \frac{g_a^2/2}{dd_0 - (g_a^2 + g_b^2)}\right]. \quad \text{(A12)}$$

2. Quantized field treatment without photon–photon correlations

We consider Eqs. (6), (8), (9), (13), and (18) of I and Eqs. (6), (7), and (13) and then we apply the decoupling approximations defined by Eqs. (15) and (17) of I or, equivalently, Eqs. (3) and (4). The final result consists of a closed set of eight coupled equations, which can be solved easily and yield the expression

$$G_{12}(\omega) = \frac{(\bar{n}_1 - \bar{n}_2)}{2\pi\gamma d_0}\left[1 + \frac{\eta_b^2/4}{dd_0 - \frac{1}{4}(\eta_a^2 + \eta_b^2)} + \frac{\eta_a^2/2}{dd_0 - (\eta_a^2 + \eta_b^2)}\right]. \quad \text{(A13)}$$

Equation (A12) is identical to (A13) if the classical expressions g_a and g_b for the reduced Rabi frequencies are replaced by their quantized field counterparts η_a and η_b, respectively. The expression (A13) for $G_{12}(\omega)$ describes the spectrum of the system when photon–photon correlations from each laser field have been neglected. The first term in Eq. (A13) is identical to that of Eq. (25), while the last two terms in both equations differ only by the numerical factor of $\frac{1}{2}$. As it has been discussed in II, the factor of $\frac{1}{2}$ is due to the fact that in deriving Eq. (25) photon–photon correlations have been considered through the use of the decoupling approximations given by Eqs. (1) and (2). The second and third terms in Eq. (25), which are peaked at the reduced frequencies $X = \pm\eta_a\sqrt{2}$ and $\pm\eta_b/\sqrt{2}$, do not appear in Eq. (A13) because in calculating Eq. (A13) use has been made of the decoupling approximation (3) and (4), where photon–photon correlations from each laser field are absent. Thus the second and third terms in Eq. (25) are due entirely to effects arising from photon–photon correlations from each laser field, respectively.

The expression (A13) is in agreement with the results derived by Cohen-Tannoudji and Reynaud,[17] where the dressed-atom method[16] has been used. The two pairs of sidebands described by the last two terms of Eq. (A13) are peaked at the frequencies $\omega_a \pm \frac{1}{2}(\eta_1^2 + \eta_2^2)^{1/2}$ and $\omega_a \pm (\eta_1^2 + \eta_2^2)^{1/2}$, respectively, which are identical to those obtained by Cohen-Tannoudji and Reynaud.[30]

[1] B. R. Mollow, Phys. Rev. A **188**, 1969 (1969); **5**, 1522 (1972); **12**, 1919 (1975); **13**, 758 (1976); **15**, 1023 (1977).
[2] F. Schuda, C. R. Stroud, Jr., and M. Hercher, J. Phys. B **7**, L198 (1974).
[3] H. J. Carmichael and D. F. Walls, J. Phys. B **8**, L77 (1975); **9**, 1199 (1976).
[4] S. S. Hassan and R. K. Bullough, J. Phys. B **8**, L147 (1975).
[5] G. S. Agarwal, Phys. Rev. Lett. **37**, 1383 (1976).
[6] J. H. Eberly, Phys. Rev. Lett. **37**, 1387 (1976).
[7] H. J. Kimble and L. Mandel, Phys. Rev. A **13**, 2123 (1976); **15**, 689 (1977).
[8] C. Cohen-Tannoudji, *Frontiers in Laser Spectroscopy* (North-Holland, Amsterdam, 1977), Vol. 1, Sec. 2.7, pp. 1–104.
[9] F. Y. Wu, R. E. Grove, and S. Ezekiel, Phys. Rev. Lett. **35**, 1426 (1975).
[10] R. E. Grove, F. Y. Wu, and S. Ezekiel, Phys. Rev. A **15**, 227 (1977).
[11] C. Mavroyannis, Phys. Rev. A **18**, 185 (1978); Opt. Commun. **26**, 453 (1978); **27**, 371 (1978); **33**, 42 (1980).
[12] D. A. Hutchinson and C. Mavroyannis, Opt. Commun. **38**, 314 (1981).
[13] D. A. Hutchinson, C. Downie, and C. Mavroyannis, Opt. Commun. **40**, 391 (1982).
[14] R. M. Whitley and C. R. Stroud, Phys. Rev. A **14**, 1498 (1976).
[15] B. Subolewska, Opt. Commun. **19**, 185 (1976).
[16] C. Cohen-Tannoudji and S. Reynaud, J. Phys. B **10**, 345 (1977).
[17] C. Cohen-Tannoudji and S. Reynaud, J. Phys. B **10**, 2311 (1977).
[18] R. Salomaa, J. Phys. B **10**, 3005 (1977).
[19] P. Avan and C. Cohen-Tannoudji, J. Phys. B **10**, 171 (1977).
[20] G. S. Agarwal, Phys. Rev. A **18**, 1490 (1978).
[21] R. Kornblith and J. H. Eberly, J. Phys. B **11**, 1545 (1978).
[22] C. Mavroyannis, Mol. Phys. **37**, 1175 (1979); Opt. Commun. **29**, 80 (1979).
[23] C. Mavroyannis and M. P. Sharma, Physica A **102**, 431 (1980).
[24] M. P. Sharma, A. B. Villaverde, and C. Mavroyannis, Can. J. Phys. **58**, 1570 (1980).
[25] C. Mavroyannis, Can. J. Phys. **58**, 957 (1980); *ibid.* **59**, 1917 (1981).
[26] C. Mavroyannis, K. J. Woloschuk, D. A. Hutchinson, and C. Downie, Can. J. Phys. **60**, 245 (1982).
[27] C. Mavroyannis and K. J. Woloschuk, Physica C **115**, 77 (1982).
[28] K. J. Woloschuck, S. Hontzeas, and C. Mavroyannis, Can. J. Phys. **60**, 968 (1982); Opt. Commun. **42**, 77 (1982).
[29] D. A. Hutchinson, C. Downie, and C. Mavroyannis, Can. J. Phys. **61**, 15 (1983).

[30]Ref. 17, p. 2322.
[31]C. Mavroyannis, Phys. Rev. A **27**, 1414 (1983).
[32]There is a printing error on page 1578 of Ref. 24, second line from the bottom, it should read "equal to $3\gamma/4$" instead of "equal to 1.5γ."
[33]P. A. M. Dirac, *The Principles of Quantum Mechanics* (Oxford U. P. , Oxford, 1958), 4th ed., pp. 7–10.
[34]There are printing errors in Eqs. (A6)–(A9) in Ref. 31, it should read "g_a" instead of "$\sqrt{g_a}$" and "g_b" instead of "$\sqrt{g_b}$."

Extension of Ruelle's stability criteria to continuous systems

Georges Abi-Ghanem[a] and V. V. Nguyen
Program in Solid Mechanics, Department of Civil Engineering, Princeton University, Princeton, New Jersey 08544

(Received 30 March 1982; accepted for publication 10 February 1984)

For a compatible description of state transition in discrete and continuous representations of a one material system, an extension of Ruelle's stability criteria is now given.

PACS numbers: 46.20. + e, 03.40.Dz, 05.70.Fh, 05.20. − y

I. INTRODUCTION

Constitutive principles of continuum mechanics are useful in describing the behavior of "nice" deformation of material systems. However, to describe the response of the system to extreme excitations, such formalism encounters technical difficulty in accounting for the various modes of critical structural changes. On the other hand, statistical mechanics, by conceptualizing the system as an assembly of a large number of discrete subsystems which are responsible for the macroscopic behavior, generically circumvents similar difficulties. The thermodynamic limits resultant of the discrete system assemblage formalism do have an equivalent basis in the continuum description. Once such an equivalence statement is found, structural stability criteria of the physical system may be conveniently extended from one mathematical framework to another. Recently, Abi-Ghanem and Nguyen[1,2] proposed a method to equate the discrete and the continuous representations of modes of structural changes. A compatible connection between the discrete and continuous approaches will be valid only when macroscopic observations are operationally definable. In statistical mechanics terms, the system must exhibit thermodynamic behavior which obeys specific stability criteria, e.g., those given by Ruelle.[3]

In this communication, we show how to extend certain stability criteria to a continuous system. Our continuum extension of these criteria employ a measure capable of accounting for critical structural behavior extant in a discrete system. Immediate consequences implicative to the classical constitutive theory arise naturally from these extended criteria, and have been discussed elsewhere.[2]

II. GENERALIZED RUELLE'S STABILITY CRITERIA FOR A CONTINUOUS SYSTEM

The possible states of our system form a set $\mathscr{G}$ of positive nonzero measures with compact support on R^m. We assume that $\mathscr{G}$ satisfies the following conditions:

R 1: (translation invariance),

$$\forall X\in R^m,\quad \forall\mu\in\mathscr{G},\quad \mu^X\in\mathscr{G},$$
$$\text{where } \mu^X(Y)=\mu(Y+X); \tag{1}$$

R 2: (superposition principle)

$$\forall\mu\in\mathscr{G},\quad \forall\mu'\in\mathscr{G},\quad \mu+\mu'\in\mathscr{G}. \tag{2}$$

Let the pair potential ϕ be a real-valued continuous function on R^m such that $\phi(X)=\phi(-X)$ and $\phi(0)\geqslant 0$, and define

$$H(\mu,\mu')=\int_{R^{2m}}\phi(X-Y)\,d\mu(X)\,d\mu'(Y),\quad \forall\mu,\mu'\in\mathscr{G}, \tag{3}$$

$$K(\mu)=H(\mu,\mu),\quad \forall\mu\in\mathscr{G}, \tag{4}$$

$$U(\mu)=\tfrac{1}{2}[K(\mu)-\phi(0)\mu(R^m)],\quad \forall\mu\in\mathscr{G}. \tag{5}$$

Here $U(\mu)$ represents the total potential energy of the system in the state μ, excluding the self-energy $\phi(0)\mu(R^m)$. We also remark that the subsequent analysis with ϕ upper-semicontinuous will not change provided representations(3) are finite.

Remark: In definition (5), the self-energy $\phi(0)\mu(R^m)$ is assumed to be well defined for a continuous material system, and has its equivalent counterpart in the discrete system. This assumption implies the following behavior of $\phi(X)$ as X approaches 0:

$$\phi(0)=\lim_{r\to 0}\,[3/4\pi r^3\mu(R^m)]$$
$$\times\int_{R^m}\int_{\|X-Y\|<r}\phi(X-Y)\,d\mu(X)\,dV(Y)<\infty,$$

where $dV(Y)$ is the regular Lebesgue measure. The above constraint may be regarded as a stability condition for the pairwise interaction.

Definition 1: We say that ϕ is stable with respect to $\mathscr{G}$ if and only if

$$K(\mu)\geqslant 0,\quad \forall\mu\in\mathscr{G}. \tag{6}$$

We now proceed to show that the stability in this sense is equivalent to convergence of the grand partition function. Since we do not want our system to describe necessarily an integer number of particles, we replace the series in the usual definition of the grand partition function by an integral. We also need to introduce an *a priori* measure on the states of our system which is sufficiently regular with the size of the system. More precisely, we have the following definition.

Definition 2: For all compact volume V in R^m, let

$$\mathscr{G}_V=\{\mu\in\mathscr{G}:\operatorname{supp}\mu\subseteq V\}.$$

Let N^* be the set of strictly positive integers. For every subset S of $\mathscr{G}_V$ and every $p\in N^*$, we denote the subsets of $\mathscr{G}_V$:

$$pS=\left\{\mu\in\mathscr{G}_V\,\middle|\,\mu=\sum_{i=1}^{p}\mu_i,\quad \mu_i\in S,\quad 1\leqslant i\leqslant p\right\},$$

and

$$S_p=\{\mu\in\mathscr{G}_V\,|\,\mu(V)=\mu'(V),\quad \mu'\in pS\}.$$

Here $\mathscr{G}_V$ is locally compact for the induced weak topology.[4] A positive measure $\mathscr{D}_V$ on $\mathscr{G}_V$ is said to be admissible if it satisfies the following condition: "For any bounded open nonempty set S of $\mathscr{G}_V$ (in the weak topology) there exist two

[a] Present address: ARDI Corporation, Research Division, P. O. Box 27113, Minneapolis, MN 55427.

 0022-2488/84/092788-03$02.50

strictly positive numbers $K_1(S)$ and $K_2(S)$, such that for any $p\in N^*$:

$$[K_1(S)]^p \leqslant \mathscr{D}_V(pS) \leqslant \mathscr{D}_V(S_p) \leqslant [K_2(S)]^p." \tag{7}$$

We now proceed to state the extension statement of Ruelle's criteria of stability.

Theorem: The following properties are equivalent.

(a) ϕ is stable with respect to $\mathscr{G}$, hence $K(\mu)\geqslant 0, \quad \forall \mu\in\mathscr{G}$.

(b) There exists $B\geqslant 0$ such that $U(\mu)\geqslant -B\mu(R^m)$, $\forall\mu\in\mathscr{G}$.

(c) For all compact $V\subseteq R^m$, all $\beta, z>0$, and all admissible measures $\mathscr{D}_V$ on $\mathscr{G}_V$, the following integral converges:

$$Z_V = 1 + \int_{\mathscr{G}_V} \mathscr{D}_V(\mu)\left(\frac{z^{\mu(V)}}{\Gamma(\mu(V)+1)}\exp[-\beta U(\mu)]\right), \tag{8}$$

where $\Gamma(\cdot)$ is the usual gamma function.

Proof: (a)$\Rightarrow$(b) with $B=\frac{1}{2}\phi(0)$ is a mere consequence of (5). To prove (b)$\Rightarrow$(c), let V be a compact volume such that $\mathscr{G}_V$ is not empty [otherwise (8) converges trivially]. Let $\mu^*\in\mathscr{G}_V$ and

$$S=\{\mu\in\mathscr{G}_V \mid L=\tfrac{1}{2}\mu^*(V)<\mu(V)<M=\mu^*(V)+1\}.$$

Here, S is nonempty, bounded, and weakly open in $\mathscr{G}_V$. Therefore, applying condition (7) gives

$$\begin{aligned}
&\int_{\mathscr{G}_V} \mathscr{D}_V(\mu)\left(\frac{z^{\mu(V)}}{\Gamma(\mu(V)+1)}\exp[-\beta U(\mu)]\right)\\
&\leqslant \int_{\mathscr{G}_V} \mathscr{D}_V(\mu)\frac{[z\exp(\beta B)]^{\mu(V)}}{\Gamma(\mu(V)+1)}\\
&\leqslant \sum_{p=1}^{\infty}\frac{[\sup\{1,z\exp(\beta B)\}]^{pM}}{\Gamma(pL+1)}\mathscr{D}_V(S_p)\\
&\quad + [\sup\{1,z\exp(\beta B)\}]^L\\
&\leqslant \sum_{p=1}^{\infty}\frac{[\sup\{1,z\exp(\beta,B)\}]^{pM}}{\Gamma(pL+1)}[K_2(S)]^p\\
&\quad + [\sup\{1,z\exp(\beta B)\}]^L < +\infty.
\end{aligned}$$

We now show that if (a) does not hold, then (c) does not hold. Indeed suppose there exists $\mu^{**}\in\mathscr{G}$ and $K(\mu^{**})=-2\epsilon<0$. Let V contain the support of μ^{**}. The function H defined by (3) is continuous for the product of weak topologies in $\mathscr{G}_V\times\mathscr{G}_V$.[4] Since it takes the value -2ϵ at (μ^{**},μ^{**}), there exists an open neighborhood S^* of μ^{**} in $\mathscr{G}_V$ such that

$$(\mu,\mu')\in S^*\times S^* \Rightarrow H(\mu,\mu') < -\epsilon.$$

If $\mu\in pS^*$, we have $\mu=\Sigma_{1<i<p}\mu_i$, $\mu_i\in S^*$, and therefore

$$K(\mu)=\sum_{1\leqslant i<p}\sum_{1\leqslant j<p}H(\mu_i,\mu_j) < -p^2\epsilon. \tag{9}$$

If necessary we restrict S^* so that for all $\mu\in S^*$,

$$\mu(V)\leqslant M=\mu^{**}(V)+1. \tag{10}$$

Since $U(\mu)\leqslant\frac{1}{2}K(\mu)$, and by using (7), (9), and (10) we obtain

$$\begin{aligned}
Z_V &\geqslant \sup_{p\in N^*}\left\{\int_{pS}\mathscr{D}_V(\mu)\frac{\mu(V)}{\Gamma(\mu(V)+1)}\exp[-\beta K(\mu)]\right\}\\
&\geqslant \sup_{p\in N^*}\left\{\frac{\exp(p^2\beta\epsilon)}{\Gamma(pM+1)}\{[\inf(1,z)]^M[K_1(S)]\}^p\right\} > +\infty.
\end{aligned}$$

This concludes the proof.

III. SPECIAL CASES

A. Example 1: "pointwise particles"

Our first example shows that the theorem is a generalization of Ruelle's Proposition 3.2.2. (Ref. 3).

Let $\mathscr{G}^1$ be the set of all finite sums of Dirac distributions with integer coefficients. The $\mathscr{G}^1_V$ is the disjoint union of closed subsets isomorphic to V^i, for $i\in N^*$. It is therefore natural to choose $\mathscr{D}_V$ to be the Lebesgue measure $\Pi_{j=1}^i dX^j$ on each component V^i of $\mathscr{G}^1_V$. Since any nonempty bounded open subset S of $\mathscr{G}^1_V$ contains an $A\subseteq V^{k_1}$ which is a product of nonempty balls, i.e., $A=\Pi_{i=1}^{k_1}A_i$, $A_i\subseteq V$, and since it is contained in $\cup_{i<k_2}V^i$ for some k_2, we have

$$\begin{aligned}
\left(\prod_{i=1}^{k_1}\mathrm{vol}(A_i)\right)^p &\leqslant \mathscr{D}_V(pS)\\
&\leqslant \sum_{i=1}^{k_2}[\mathrm{vol}(V)]^{ip}\\
&\leqslant [\sup\{1,\mathrm{vol}(V)\}]^{2k_2p}.
\end{aligned}$$

Therefore $\mathscr{D}_V$ is admissible.

The conditions (a), (b), and (c) in the main theorem then reduce, respectively, to

$$\text{(a')}\quad \sum_{1\leqslant i\leqslant n}\sum_{1\leqslant j\leqslant n}\phi(X^i-X^j)\geqslant 0, \quad \forall X^1,\ldots,X^n; \tag{11}$$

$$\text{(b')}\quad \sum_{1\leqslant i<j\leqslant n}\phi(X^i-X^j)\geqslant -Bn, \quad \forall X^1,\ldots,X^n; \tag{12}$$

$$\text{(c')}\quad Z_V = 1+\sum_{n=1}\frac{z^n}{n!}\int_{V^n}dX^1\cdots dX^n \times\left[\exp\left(-\beta\sum_{1\leqslant i<j\leqslant n}\phi(X^i-X^j)\right)\right] \tag{13}$$

is convergent; which are Ruelle's criteria of stability. By using linear combinations of Dirac measures with positive (not necessarily integer) coefficients, we arrive at the same results [(11)–(13)].

Proposition 1:
ϕ satisfies inequality (11),

$$\Leftrightarrow \sum_{1\leqslant i\leqslant n}\sum_{1\leqslant j\leqslant n}a_ia_j\phi(X^i-X^j)\geqslant 0,$$
$$\forall a_1,\ldots,a_n\in R^+, \quad X^1,\ldots,X^n\in R^m. \tag{14}$$

Proof: Equation (14) is an obvious consequence of (11) if $a_i\in N$. If $a_i\in Q^+$, and $a_i=p_i/q_i, p_i,q_i\in N^*$, then multiplying (11) by $\Pi_{1<i<n}q_i^2$ leads to an equivalent inequality with integers $a_i'=p_i\,\Pi_{j\neq i}q_j$. The proof is achieved by noticing that any $a_i\in R^+$ is a limit of positive rationals.

Thus we have the following proposition.

Proposition 2: If $\mathscr{G}$ is a set of Radon measures, stability with respect to $\mathscr{G}^1$ implies stability with respect to $\mathscr{G}$.

Proof: Since any positive measure can be approximated by linear combinations of Dirac measures with positive coefficients, the proof can be concluded by applying Proposition 1.

Proposition 3: If $\mathscr{G}$ is a set of Radon measures, and the weak closure of $\mathscr{G}$ contains a Dirac measure $\lambda_0\delta(X^0)$, $\lambda_0>0$, $X^0\in R^m$, then stability with respect to $\mathscr{G}$ and to $\mathscr{G}^1$ are equivalent.

Proof: By conditions (11) and (12), the weak closure of $\mathcal{G}$ contains all linear combinations of Dirac measures with integer coefficients, up to the unimportant factor λ_0.

B. Example 2: "smeared out particles"

We now give an example of a potential unstable in the sense of Ruelle but stable with respect to a certain $\mathcal{G}^2$ of Radon measures.

Let

$$\mathcal{G}^2 = \left\{ \sum_{1 \leqslant i \leqslant n} \mu_0^{X^i}, \quad X^1,\ldots,X^n \in R^m \right\},$$

where μ_0 represents a particle smeared out in a small ball around the origin. For example, in three dimensions,

$$d\mu_0(X) = \begin{cases} dX/4\pi a^3, & \text{if } |X| \leqslant a, \\ 0, & \text{elsewhere.} \end{cases}$$

The potential defined by

$$\phi(X) = \begin{cases} 11, & \text{if } |X| \leqslant R - \epsilon/2, \\ -1, & \text{if } R - \epsilon/2 < |X| < R + \epsilon/2, \\ 0, & \text{elsewhere,} \end{cases}$$

is known to be unstable.

It is not continuous, but we can make it continuous by modifying its value slightly in the vicinity of the spheres $X = R \pm \epsilon/2$. One can easily show the stability of ϕ with respect to $\mathcal{G}^2$ when $a \gg \epsilon$. Physically, the finite and incompressible size of the particles (or the uncertainty in their positions) prevents them from collapsing in a perfect cubic face centered lattice of points with nearest neighbors distance R, a configuration for which the energy would not be linearly bounded from below.

Finally, we remark that it is still an open problem to characterize the potentials which verify condition (11) or condition (6) for some set $\mathcal{G}$. From Proposition 2, we know that any sum of two functions, one being positive and the other being of positive type, is a stable potential for any set $\mathcal{G}$ of Radon measures; but from example 2, we see that for a particular set $\mathcal{G}$, these are in general not the only ones to be stable.

[1] G. V. Abi-Ghanem and V. V. Nguyen, Phys. Rev. B **26**, 4321 (1982).
[2] G. V. Abi-Ghanem and V. V. Nguyen, J. Phys. A: Math. Gen **16**, 3401 (1983).
[3] D. Ruelle, *Statistical Mechanics* (Benjamin, Reading, MA, 1977), p. 219.
[4] N. Bourbaki, *Elements de Mathématique: Intégration* (Hermann, Paris, 1965), Vol. VI, Chaps. 1–4, p. 283.

Effect of the boundary condition at distance of closest approach on Wien dissociation of a weak electrolyte: A singular perturbation solution

Douglas K. McIlroy and David P. Mason
Department of Applied Mathematics, University of Witwatersrand, Johannesburg, South Africa

(Received 6 December 1983; accepted for publication 3 February 1984)

A singular perturbation technique is developed in order to examine the effect of plausible variations in the boundary condition of the ion-pair distribution function f at the distance of closest approach of the oppositely charged ions in a weakly dissociated electrolyte. Such variations from the point-ion model may arise in practice where the ions are weakly solvated or in ion-exchange resins or soluble polyelectrolytes. The relative increase $K(X)/K(0)$ in the dissociation constant of the weak electrolyte to which a uniform electric field $\mathbf{X}$ is applied is calculated to third order in the field-dependent parameter ϵ, assuming that the mean distance of closest approach between the charge centers of oppositely charged ions, a, is nonzero. A novel feature of this problem is that the boundary condition on f at distance a depends on the dissociation constant itself. It is found that the constant of proportionality in Onsager's linear law, derived for the idealized case of a weakly dissociated electrolyte of point ions, remains unchanged when $a \neq 0$ but the second- and third-order terms in the expansion for $K(X)/K(0)$ in powers of ϵ are each increased by an additional term which is proportional to $\exp(-1/a)$.

PACS numbers: 66.10. − x, 82.45. + z

I. INTRODUCTION

The main deduction from the mathematical theory of Wien dissociation of a weak electrolyte to which a uniform electric field $\mathbf{X}$ is applied is Onsager's remarkable formula[1] for the relative increase in the dissociation constant,

$$K(X)/K(0) = J_1[(-8\epsilon)^{1/2}]/(-2\epsilon)^{1/2}, \tag{1.1}$$

$$K(X)/K(0) = 1 + \epsilon + \epsilon^2/3 + \epsilon^3/18 + \cdots, \tag{1.2}$$

where J_1 is the ordinary Bessel function of order 1, $\epsilon = 2\beta q$, and

$$q = \frac{-e_i e_j}{2DkT} > 0, \quad \beta = \frac{X|e_i\omega_i - e_j\omega_j|}{2kT(\omega_i + \omega_j)}. \tag{1.3}$$

Here, q is the Bjerrum association distance, e_i and e_j (of opposite sign) and are the charges on the ions of the weak electrolyte, ω_i and ω_j are their mobility coefficients, D and T are the dielectric constant and absolute temperature of the electrolytic solution, and k is Boltzmann's constant.

An important assumption in the derivation of (1.1) is that a, the distance of closest approach of the i-ions and j-ions, can be taken as zero. Indeed Onsager's solution[1] for the distribution function f describing the pairs of associated, oppositely charged, ions of the weak electrolyte may be considered as a *direct consequence* of this idealization. In this article we will treat the more realistic case $a \neq 0$ and will calculate the resulting modification to (1.2).

There are several reasons why this problem is of interest. For instance Onsager[2] notes that in certain extreme cases, e.g., in applications to colloids or to ionization in gases, a variety of boundary conditions may have to be considered. Or, if the ions of a weak electrolyte should be weakly solvated, the distance of closest approach will be nonzero and the limiting case of zero separation in ionic collisions will be inappropriate. Again, in the case of dissociation of ion-exchange resins and soluble polyelectrolytes, where a charged ion may accept more than one counterion, the collision of point charge ions will be a poor model of the process of ionic association.

The contribution of this article is to give an *analytical* derivation, as distinct from an order of magnitude estimate, of the correction to (1.2) due to $a \neq 0$ and thereby to establish a means of investigating the more complicated, and as yet largely unformulated, problems mentioned above. We find that the correction is in the form of an additional term $O[\exp(-2q/a)]$ in the coefficients of ϵ^2 and ϵ^3. Our results agree with the order of magnitude estimate given by Onsager[1] who, however, estimated only a correction to the equilibrium approximation to f so that exact comparison with our result for $K(X)/K(0)$ is precluded.

Our method of solution[3] leading to (1.2) for $K(X)/K(0)$ is a singular perturbation. A straightforward expansion in powers of ϵ of f, the Legendre transform[4] of which is required in the computation of $K(X)/K(0)$, contains secular terms in r, the separation distance between oppositely charged ions, and hence the expansion is not valid for large r. Furthermore, the boundary conditions are imposed on f at $r = a$ (or zero in the idealized case) and at $r = \infty$, and consequently many of the standard singular perturbation techniques which were developed to deal with secular behavior in initial value problems[5] are not directly applicable.

The perturbation technique which we will employ here can be summarized as follows. By systematically determining the asymptotic behavior of f for both large and small r using a method described by Murray,[6] we first show that f can be written as the product of two functions, one of which is known and contains the asymptotic behavior of f and the other function, h, is unknown but possesses a perturbation expansion which is uniformly valid in r for $a \leqslant r \leqslant \infty$. We show that to calculate $K(X)/K(0)$ it is not necessary to obtain f but only the Legendre transform of h and that by expanding this in powers of ϵ and solving a series of ordinary differential equations for each order in ϵ a uniformly valid solution is

 0022-2488/84/092791-09$02.50

obtained. Variants of this technique have been applied successfully to other problems in this field.[3,7] For the limiting case $a = 0$ (see Ref. 3) the procedure could be simplified because each term in the expansion of h admitted a separable solution and we were able to exploit this to obtain the closed form solution (1.1). (A separable solution for each term in the expansion of h is not, however, possible in the present problem because of the form of the boundary condition at $r = a$.) When considering Wien dissociation of charged macromolecular sites in biomembranes we were able to extend the method employed in this article to a *dual* perturbation,[7] the second perturbation parameter being introduced through the interionic potential.

Section II of this paper is concerned with the mathematical formulation of the problem. The boundary condition on the ion-pair distribution function, which is central to the present work, is closely examined. It is explained how the concentrations of free (dissociated) ions, n_i and n_j, must be functions of the applied electric field intensity $\mathbf{X}$ and we show that this in turn implies that the boundary condition at $r = a$ on that part of the distribution function which describes associated ion pairs (f), depends on the dissociation constant itself. The partial differential equation governing f and an expression for $K(X)/K(0)$ in terms of this distribution function are deduced. In Sec. III the problem is reformulated as a singular perturbation problem and in Sec. IV we derive as our key result the correction to (1.2) arising from $a \neq 0$. Finally, we discuss our results and conclusions in Sec. V.

II. GOVERNING EQUATIONS

A. Boundary conditions on the ion-pair distribution function

The theory of Brownian motion of ions in a weak electrolyte to which a uniform electric field has been applied has been discussed by several authors.[2,8–11] In this section we therefore briefly outline the underlying principles of the equations which we will require in the sequel.

Consider an electrolytic solution in a uniform applied electric field $\mathbf{X}$, and suppose that dV_i and dV_j are a pair of volume elements of this electrolyte, centered at position vectors $\mathbf{r}_i$ and $\mathbf{r}_j$, respectively, with respect to some arbitrary origin; define $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$. Then the distribution function $F_{ji}(\mathbf{r}_j, \mathbf{r}_{ji})$ is the number of pairs of oppositely charged ions which can be formed from the i-ions in dV_i and the j-ions in dV_j, per unit volume of $(\mathbf{r}_i, \mathbf{r}_j)$ space. It will be assumed that there is no velocity gradient applied to the electrolytic solution as a whole so that F_{ji} depends only on the relative position $\mathbf{r}_{ji}$ and not on $\mathbf{r}_i$ and $\mathbf{r}_j$ separately; thus $F_{ji} = F_{ji}(\mathbf{r}_{ji})$. The function F_{ji} is the total ion-pair distribution function in the sense that it describes the state of complete dissociation as well as the undissociated ion pairs. We may also define the reciprocal function $F_{ij} = F_{ij}(\mathbf{r}_{ij})$. Clearly $F_{ji}(\mathbf{r}_{ji}) = F_{ij}(\mathbf{r}_{ij})$.

In order to discuss the boundary conditions on F_{ji} it is first necessary to consider the mean velocity of an i-ion relative to a neighboring j-ion. Since for weak electrolytes the concentrations n_i and n_j of the free ions are very small,

$$q \ll \kappa^{-1} \equiv [DkT/4\pi(n_i e_i^2 + n_j e_j^2)]^{1/2}, \tag{2.1}$$

which ensures that the effects of the ionic atmospheres may be neglected, κ^{-1} being the Debye–Hückel "radius" of the ionic atmospheres. Then it can be shown[12] that, on the assumption that the bulk velocity of the electrolyte is zero, the mean velocity of an i-ion in the neighborhood of a j-ion, $\mathbf{v}_{ji}$, is given by

$$\mathbf{v}_{ji} = \omega_i [e_i \mathbf{X} - \operatorname{grad}_i(e_i e_j/Dr) - kT \operatorname{grad}_i(\log F_{ji})], \tag{2.2}$$

where the subscript i on the gradient operator denotes differentiation with respect to the components of $\mathbf{r}_i$ and $r = |\mathbf{r}_{ji}|$. In (2.2), the first term on the right-hand side is the contribution to $\mathbf{v}_{ji}$ due to the applied electric field $\mathbf{X}$, the second term is that resulting from the (unshielded) interionic Coulombic force, and the third term is the contribution due to the diffusion velocity of the i-ions, the diffusion coefficient of the i-ions being $kT\omega_i$ (see Ref. 13). In the interionic force it is assumed that even though the ions may not be point charges, the force law is still the Coulomb inverse square law where distance is measured between the centers of the ions.

Similarly, if $\mathbf{v}_{ij}$ is the mean velocity of a j-ion in the neighborhood of an i-ion, then

$$\mathbf{v}_{ij} = \omega_j [e_j \mathbf{X} - \operatorname{grad}_j(e_i e_j/Dr) - kT \operatorname{grad}_j(\log F_{ij})]. \tag{2.3}$$

If $\mathbf{v}$ is the mean velocity of an i-ion relative to a neighboring j-ion we have $\mathbf{v} = \mathbf{v}_{ji} - \mathbf{v}_{ij}$, and on subtracting (2.3) from (2.2),

$$\mathbf{v} = kT(\omega_i + \omega_j)[2\beta\mathbf{n} + \operatorname{grad}_i(2q/r) - \operatorname{grad}_i(\log F_{ji})], \tag{2.4}$$

since $\operatorname{grad}_j = -\operatorname{grad}_i$ and $F_{ij}(\mathbf{r}_{ij}) = F_{ji}(\mathbf{r}_{ji})$, where β and q are defined as in (1.3) and

$$\mathbf{n} = (\mathbf{X}/X)\operatorname{sgn}(\omega_i e_i - \omega_j e_j); \tag{2.5}$$

$\operatorname{sgn}(\omega_i e_i - \omega_j e_j) = +1$ if $\omega_i e_i - \omega_j e_j > 0$ and -1 if $\omega_i e_i - \omega_j e_j < 0$. To be specific we will assume that $e_i > 0$ and $e_j < 0$ so that $\operatorname{sgn}(\omega_i e_i - \omega_j e_j) = +1$ and $\mathbf{n}$ is a unit vector parallel to $\mathbf{X}$. We will employ nondimensional variables in the subsequent analysis. A suitable characteristic length is $2q$ and the corresponding characteristic time is $\tau_D = (2q)^2/kT(\omega_i + \omega_j)$; τ_D is the diffusion relaxation time and is measure of the time taken for an i-ion to diffuse, relative to a neighboring j-ion, a distance of $2q$. In nondimensional variables (2.4) is

$$\mathbf{v} = 2\epsilon\mathbf{n} + \operatorname{grad}(1/r) - \operatorname{grad}(\log F), \tag{2.6}$$

where the subscripts on the ion-pair distribution function and on the gradient operator have been suppressed to simplify the notation. The boundary conditions on the distribution function are most conveniently discussed if the interionic force and the force on the ions due to the applied electric field are expressed as the gradient of a potential. To do this we introduce spherical polar coordinates (r, θ, ϕ) based upon the direction of the applied electric field $\mathbf{n}$; θ is the angle which an oppositely charged ion pair makes with the direction of $\mathbf{n}$, and ϕ is the azimuthal angle about $\mathbf{n}$. Then if we define the total nondimensional potential Φ by

$$\Phi(r,\theta) = -1/r - 2\epsilon r\cos\theta, \tag{2.7}$$

(2.6) assumes the form

$$\mathbf{v} = -\operatorname{grad}\Phi - \operatorname{grad}(\log F). \tag{2.8}$$

Now ions are often far from spherical. To obtain an estimate of the effect of the ion's sizes on the kinetics of the weak electrolyte, we therefore suppose that a (nondimensionalized with respect to $2q$) is the *mean* distance of closest

approach of the charge centers of an i-ion and a j-ion, the value of a being determined by the volumes "traversed" by the electrons of the charge centers of the i-ions and j-ions.

Next consider the boundary condition at $r = a$. We will require that the total ionic flux is zero for all θ at $r = a$:

$$F(a,\theta)\mathbf{v}(a,\theta) = \mathbf{0}, \quad \forall\theta, \tag{2.9}$$

which is one of the boundary conditions noted by Onsager[2] as being of importance. This condition in fact corresponds to infinite mobility of the ions in the sphere $r < a$ and is possibly the simplest boundary condition which we can envisage in this problem. Nevertheless the mathematical techniques which we develop here should be applicable in the main to similar problems with more complicated boundary conditions. Since there is obvious symmetry about $\mathbf{n}$, F has no dependence on ϕ. Using (2.8) for $\mathbf{v}$ and expressing (2.9) in component form, we obtain the following two equations valid at $r = a$:

$$\frac{\partial F}{\partial r} + F\frac{\partial \Phi}{\partial r} = 0, \tag{2.10}$$

$$\frac{\partial F}{\partial \theta} + F\frac{\partial \Phi}{\partial \theta} = 0. \tag{2.11}$$

It follows directly from (2.10) and (2.11) that at $r = a$,

$$\frac{dF}{d\Phi} + F = 0, \tag{2.12}$$

whence

$$F(a,\theta) = Ae^{-\Phi(a,\theta)}, \tag{2.13}$$

where A is independent of Φ. Although independent of Φ, A could conceivably depend separately on a, θ, and ϵ (and hence on X); we show that this is, however, not possible. For suppose that A depends on ϵ. Then solving (2.7) for ϵ we have

$$\epsilon = -(1 + a\Phi)/2a^2\cos\theta, \tag{2.14}$$

which implies that A can be expressed as a (different) function of a,θ, and Φ, contradicting the fact that A is independent of Φ. By a similar argument A can be shown to be independent of a and θ. To determine A, we note that when $\mathbf{X} = \mathbf{0}$, the total distribution function is, taking the ground state to be that of the randomly distributed free ions,

$$F = n_i(0)n_j(0)e^{-\Phi}, \tag{2.15}$$

where in this case $\Phi = -1/r$. Thus

$$F(a,\theta) = n_i(0)n_j(0)e^{-\Phi(a,\theta)}, \tag{2.16}$$

and so the total distribution function reduces at $r = a$ to a Maxwell–Boltzmann distribution in the combined Coulomb and externally applied fields.

Now, the total distribution function can be split into two parts: the part $f^{(1)}$ which describes the state of complete dissociation and the part $f^{(2)}$ which describes associated ion pairs of oppositely charged ions. Thus,

$$F = f^{(1)} + f^{(2)} \tag{2.17}$$

and

$$f^{(1)} = n_i(X)n_j(X) = \text{const}, \tag{2.18}$$

where the constant in (2.18) is a function of X because the concentrations of free ions must increase as X is increased. Equation (2.18) arises because $f^{(1)}$ is the number of pairs of oppositely charged ions which can be formed from the free i-ions in the volume element dV_i centered at $\mathbf{r}_i$, and the free j-ions in the volume element dV_j centered at $\mathbf{r}_j$, per unit volume of $(\mathbf{r}_i,\mathbf{r}_j)$ space. Equations (2.16)–(2.18) therefore imply that

$$f^{(2)}(a,\theta) = n_i(0)n_j(0)e^{-\Phi(a,\theta)} - n_i(X)n_j(X). \tag{2.19}$$

Next we normalize $f^{(2)}$ by defining

$$f = f^{(2)}/n_i(0)n_j(0), \tag{2.20}$$

so that (2.19) becomes

$$f(a,\theta) = e^{-\Phi(a,\theta)} - n_i(X)n_j(X)/n_i(0)n_j(0). \tag{2.21}$$

It remains to obtain an expression for $n_i(X)n_j(X)/n_i(0)n_j(0)$. Now for distances which are but slightly greater than a, the great majority of the contribution to F is in the form of associated ion pairs. Thus the total concentration of "bound" ion pairs, ν_{ji}, is essentially given by the volume integral of F from $r = a$ to $r = \frac{1}{2}$, the upper limit being somewhat arbitrary. But in a weak electrolyte the concentration of bound pairs of oppositely charged ions is very large and the concentrations of free ions are very small for all finite $\mathbf{X}$, and consequently the change in ν_{ji} due to the imposition of $\mathbf{X}$ is negligible. Thus we may take

$$\nu_{ji}(X) = \nu_{ji}(0). \tag{2.22}$$

The law of mass action gives

$$\frac{d\nu_{ji}}{dt} = k'n_i(X)n_j(X) - K(X)k'\nu_{ji}(X), \tag{2.23}$$

where k' is the association rate constant which may be shown to be independent of X (see Sec. II C). Since we assume that a steady state has been attained, $d\nu_{ji}/dt = 0$, and hence from (2.23),

$$n_i(X)n_j(X) = K(X)\nu_{ji}(X), \tag{2.24}$$

$$n_i(0)n_j(0) = K(0)\nu_{ji}(0). \tag{2.25}$$

The boundary condition (2.21) may therefore be written as

$$f(a,\theta) = e^{-\Phi(a,\theta)} - K(X)/K(0), \tag{2.26}$$

using (2.22).

Equation (2.26) is the required boundary condition on $f(r,\theta)$ at $r = a$ and shows the close relation between the distribution function and the dissociation constant. The novel feature of this problem is that the boundary condition (2.26) on f at $r = a$ contains the desired function $K(X)/K(0)$. This complication, absent from the previous mathematical treatments of Wien dissociation[1,3,7,10] arises purely as a result of a being nonzero. Unlike the case $a = 0$, the ion-pair distribution function in the present problem cannot be calculated without first obtaining $K(X)/K(0)$ in the process.

Furthermore, as the separation distance r between an i-ion and a j-ion tends to infinity, the ions become completely dissociated. Hence the second boundary condition on f is

$$\lim_{r\to\infty} f(r,\theta) = 0. \tag{2.27}$$

B. The partial differential equation for f

The partial differential equation governing f follows from the continuity equation associated with the motion of

the i-ions and j-ion. Since the number of ions of each species is conserved, the average motion of a pair of oppositely charged ions in six-dimensional space with coordinates $(\mathbf{r}_i,\mathbf{r}_j)$ must satisfy the continuity equation

$$\frac{\partial F}{\partial t} + \mathrm{div}_i(F\mathbf{v}) = 0. \tag{2.28}$$

For a steady state, $\partial F/\partial t = 0$ and (2.28) reduces to

$$\mathrm{div}_i(F\mathbf{v}) = 0. \tag{2.29}$$

Now, if we partition F as in (2.17), it can be easily verified using (2.8) that

$$F\mathbf{v} = f^{(1)}\mathbf{v}^{(1)} + f^{(2)}\mathbf{v}^{(2)}, \tag{2.30}$$

where

$$\mathbf{v}^{(k)} = -\ \mathrm{grad}\ \Phi - \mathrm{grad}(\log f^{(k)}), \quad k = 1,2. \tag{2.31}$$

Since $f^{(1)} = n_i(X)n_j(X) = \mathrm{const}$, it follows that

$$\mathrm{div}(f^{(1)}\mathbf{v}^{(1)}) = f^{(1)}\mathrm{div}\ \mathbf{v}^{(1)} = -f^{(1)}\nabla^2\Phi = 0, \tag{2.32}$$

and hence from (2.29), (2.30), and (2.32)

$$\mathrm{div}(f^{(2)}\mathbf{v}^{(2)}) = 0. \tag{2.33}$$

Thus steady-state continuity equations are satisfied *separately* for each process. Using (2.31) and recalling that $\nabla^2\Phi = 0$, (2.33) can be written as

$$\nabla^2 f^{(2)} + \mathrm{grad}\, f^{(2)} \cdot \mathrm{grad}\ \Phi = 0. \tag{2.34}$$

Since (2.34) is homogeneous in $f^{(2)}$, we can normalize $f^{(2)}$ as in (2.20). On noting that f is independent of ϕ, (2.34) becomes

$$r^2\frac{\partial^2 f}{\partial r^2} + (1 + 2r - 2\epsilon r^2\cos\theta)\frac{\partial f}{\partial r} + \frac{\partial^2 f}{\partial \theta^2} + (\cot\theta + 2\epsilon r\sin\theta)\frac{\partial f}{\partial \theta} = 0, \tag{2.35}$$

which is the required equation for $f(r,\theta)$.

C. Association and dissociation constants

Finally, we consider the association rate constant k' and the dissociation constant $K(X)$ of the weak electrolyte. The time rate of change of concentration of associated pairs of oppositely charged ions is

$$\frac{dv_{ji}}{dt} = -\int_S F\mathbf{v}\cdot d\mathbf{S}, \tag{2.36}$$

where S is an arbitrary surface surrounding the origin. Because of (2.30) this can be rewritten as

$$\frac{dv_{ji}}{dt} = -\int_S f^{(1)}\mathbf{v}^{(1)}\cdot d\mathbf{S} - \int_S f^{(2)}\mathbf{v}^{(2)}\cdot d\mathbf{S}, \tag{2.37}$$

and on comparing (2.37) with (2.23) we find that

$$k' n_i(X)n_j(X) = -\int_S f^{(1)}\mathbf{v}^{(1)}\cdot d\mathbf{S}, \tag{2.38}$$

$$K(X)k' v_{ji} = \int_S f^{(2)}\mathbf{v}^{(2)}\cdot d\mathbf{S}. \tag{2.39}$$

With the aid of (2.18) for $f^{(1)}$ and (2.31) for $\mathbf{v}^{(1)}$, and choosing for S a sphere of arbitrary radius $> a$ and center $r = 0$, it is easily verified[10] that

$$k' = 4\pi. \tag{2.40}$$

Using (2.25) and (2.40) we find that (2.39) may be rewritten as

$$\frac{K(X)}{K(0)} = \frac{1}{4\pi}\int_S f\mathbf{v}^{(2)}\cdot d\mathbf{S}, \tag{2.41}$$

which becomes, using (2.31) for $\mathbf{v}^{(2)}$,

$$\frac{K(X)}{K(0)} = -\frac{1}{2}\int_0^\pi \left[\frac{\partial f}{\partial r} + \left(\frac{1}{r^2} - 2\epsilon\cos\theta\right)f\right] r^2\sin\theta\, d\theta. \tag{2.42}$$

This completes the formulation of our problem which is to calculate $K(X)/K(0)$ defined by (2.42), where f satisfies (2.35), subject to the boundary condition (2.26) at $r = a$ and $f = 0$ at $r = \infty$.

III. PERTURBATION EQUATIONS

A. Asymptotic behavior of f

We determine the asymptotic behavior of $f(r,\theta)$ for large r by using an asymptotic technique described by Murray[6] and since it is an improvement over our previous treatments[3,7] we will outline the analysis.

We first substitute the transformation

$$f(r,\theta) = g(r,\theta)\exp(\epsilon r\cos\theta) \tag{3.1}$$

into (2.35) which becomes

$$r^2\frac{\partial^2 g}{\partial r^2} + (2r + 1)\frac{\partial g}{\partial r} - \epsilon^2 r^2 g = -\left(\frac{\partial^2 g}{\partial \theta^2} + \cot\theta\frac{\partial g}{\partial \theta} + \epsilon\cos\theta g\right). \tag{3.2}$$

Unlike (2.35) for f, this equation admits a separable solution of the form $g(r,\theta) = R(r)Z(\theta)$, where $R(r)$ satisfies

$$\frac{d^2R}{dr^2} + \left(\frac{2}{r} + \frac{1}{r^2}\right)\frac{dR}{dr} - \left(\epsilon^2 + \frac{\alpha}{r^2}\right)R = 0, \tag{3.3}$$

and α is the separation constant. If we let

$$R(r) = (e^{1/2r}/r)S(r), \tag{3.4}$$

then (3.3) can be rewritten in the standard form for asymptotic analysis,[6]

$$\frac{d^2S}{dr^2} - \left(\frac{1}{4r^4} + \frac{\alpha}{r^2} + \epsilon^2\right)S = 0. \tag{3.5}$$

When $\epsilon \neq 0$, the point at infinity is an irregular singular point since the coefficient of S in (3.5) is $O(\epsilon^2)$ and not $O(1/r^2)$ as $r\to\infty$. Hence we seek an asymptotic solution to (3.5) of the form

$$S(r) = e^{\lambda r}r^\sigma\left(A_0 + \frac{A_1}{r} + \frac{A_2}{r^2} + \cdots\right), \quad \text{as } r\to\infty, \tag{3.6}$$

where λ, σ, and $A_n (n \geqslant 0)$ are constants: it is found that for $A_0 \neq 0$, $\lambda = \pm\epsilon$ and $\sigma = 0$. For a solution bounded as $r\to\infty$ we require $\lambda = -\epsilon$ and hence

$$S(r) = A_0 e^{-\epsilon r}[1 + O(1/r)], \quad \text{as } r\to\infty. \tag{3.7}$$

Thus by (3.1), (3.4), and (3.7) we have

$$f(r,\theta) = B_0(\theta)\exp[-\epsilon r(1 - \cos\theta)] \times [1/r + O(1/r^2)], \quad \text{as } r\to\infty, \tag{3.8}$$

where $B_0(\theta) = A_0 Z(\theta)$.

Now, if we perform the straightforward perturbation expansion

$$f(r,\theta)=\sum_{m=0}^{\infty}\epsilon^m f^{(m)}(r,\theta), \tag{3.9}$$

then we see from (3.8) that for $m\geqslant 0$,

$$f^{(m)}(r,\theta)=O\left[r^{m-1}(1-\cos\theta)^m\right], \quad \text{as } r\to\infty. \tag{3.10}$$

If $\cos\theta\neq 1$, $f^{(m)}$ for $m\geqslant 1$ is more singular than $f^{(m-1)}$ as $r\to\infty$ and (3.9) is not valid for large r due to secular behavior in r.

The asymptotic behavior of $f(r,\theta)$ for small r can be obtained in exactly the same way as for large r by making the change of variable $r=1/s$ in (3.3) and considering the asymptotic behavior of $R(s)$ as $s\to\infty$. It is found that

$$f(r,\theta)=C_0(\theta)\exp(1/r+\epsilon r\cos\theta)[1+O(r)], \quad \text{as } r\to 0. \tag{3.11}$$

Thus, even for the idealized case $a=0$, $f^{(1)}$, for instance, will not be more singular than $f^{(0)}$ as $r\to 0$ and we could consider the straightforward expansion (3.9) of f in this region. However, for the case $a=0$ (see Ref. 3), in order to simplify the boundary condition at $r=0$, we removed the factor $e^{1/r}$ from the new independent variable $h(r,\theta)$ and we will do the same here; this allows the two perturbation solutions to be compared more easily. We therefore split $f(r,\theta)$ into two factors:

$$f(r,\theta)=h(r,\theta)\exp[1/r-\epsilon r(1-\cos\theta)]. \tag{3.12}$$

From (3.8), we see that $h(r,\theta)$ will not exhibit secular behavior for large r and we therefore look for a solution in the form of a perturbation expansion of h instead of f.

B. Infinite system of coupled ordinary differential equations

Since (2.35) for f and (3.12) for h are the same as for point ions, we merely outline here the important steps in the analysis.

Because of the form of the potential Φ defined by (2.7), we make the change of variable $u=1/r$, $x=\cos\theta$. Equation (2.35) rewritten in terms of $h(u,x)$ is expanded according to

$$h(u,x)=\sum_{m=0}^{\infty}\epsilon^m h^{(m)}(u,x). \tag{3.13}$$

Now, it is shown in the following section that to calculate $K(X)/K(0)$ to a given order in ϵ, it is sufficient to determine for certain nonnegative integers m and n, the Legendre transforms $H_n^{(m)}(u)$ of $h^{(m)}(u,x)$ defined by[3]

$$H_n^{(m)}(u)=\int_{-1}^{1}P_n(x)h^{(m)}(u,x)dx, \tag{3.14}$$

where $P_n(x)$ denotes the Legendre polynomial of degree n, and that the functions $h^{(m)}(u,x)$ themselves are not required. Expansion (3.13) is therefore substituted in the partial differential equation for $h(u,x)$ and the Legendre transform is taken of the resulting partial differential equation for $h^{(m)}(u,x)$ obtained by equating the coefficients of ϵ^m. With the aid of elementary indentities for Legendre polynomials,[14] we obtain for integers $m\geqslant 0$, $n\geqslant 0$

$$\begin{aligned}\frac{d^2H_n^{(m)}}{du^2}+\frac{dH_n^{(m)}}{du}-\frac{n(n+1)}{u^2}H_n^{(m)}&=\frac{(2-u)}{u^3}H_n^{(m-1)}-\frac{(n+1)}{(2n+1)u^2}H_{n+1}^{(m-1)}\\&\quad-\frac{n}{(2n+1)u^2}H_{n-1}^{(m-1)}-\frac{2}{u^2}\frac{dH_n^{(m-1)}}{du}.\end{aligned} \tag{3.15}$$

Equation (3.15) is valid for $m=0$ by defining $H_n^{(-1)}(u)=0$, $\forall n\geqslant 0$, and for $n=0$ either by defining $H_{-1}^{(m-1)}(u)=0$ or by noting that this transform is multiplied by n in (3.15).

Thus the infinite system of coupled ordinary differential equations (3.15) for integers $m\geqslant 0$, $n\geqslant 0$ replaces the partial differential equation (2.35) in this problem.

C. Dissociation constant

When expressed in terms of x and u, (2.42) is

$$\frac{K(X)}{K(0)}=\frac{1}{2}\int_{-1}^{1}\left(\frac{\partial f}{\partial u}-f+\frac{2\epsilon x}{u^2}f\right)dx. \tag{3.16}$$

Equation (3.16) has the same form as when $a=0$ (see Ref. 3) but, unlike the latter case, $h^{(m)}(u,x)$ is not separable in u and x when $a\neq 0$ [this can be traced to the fact that the boundary condition (3.23) is not separable in a and x] and hence the results derived below in terms of $H_n^{(m)}(u)$ for the expansion of $K(X)/K(0)$ in powers of ϵ are different from those obtained for the case $a=0$.

We have noted that the expansion (3.9) for f is not uniformly valid for $0\leqslant u\leqslant 1/a$. But $K(X)/K(0)$ is independent of u (see Ref. 3). Hence by choosing a value of u, say of order of magnitude unity, for which the expansion (3.9) of f is valid, $K(X)/K(0)$ may be expanded in the form

$$\frac{K(X)}{K(0)}=\sum_{m=0}^{\infty}\epsilon^m K^{(m)}. \tag{3.17}$$

On expanding (3.12) in powers of ϵ, substituting into (3.16), and equating the coefficients of ϵ^m we obtain for $m\geqslant 1$,

$$\begin{aligned}K^{(m)}=\frac{e^u}{2}\int_{-1}^{1}\Bigg(&\sum_{s=0}^{m}\frac{(x-1)^{m-s}}{(m-s)!u^{m-s}}\frac{\partial h^{(s)}}{\partial u}\\&+\sum_{s=0}^{m-1}\frac{(x+1)(x-1)^{m-s-1}}{(m-s-1)!u^{m-s-1}}h^{(s)}\Bigg)dx.\end{aligned} \tag{3.18}$$

Equation (3.18) is also valid for $m=0$ if we omit the second summation.

We list below explicit expressions for $K^{(m)}$ to third order. By expressing all polynomials in x as the sum of Legendre polynomials, a direct calculation yields

$$K^0=\frac{1}{2}e^u\frac{dH_0^{(0)}}{du}, \tag{3.19}$$

$$K^{(1)}=\frac{1}{2}e^u\left(\frac{dH_0^{(1)}}{du}+\frac{1}{u}\frac{dH_1^{(0)}}{du}-\frac{1}{u}\frac{dH_0^{(0)}}{du}+\frac{1}{u^2}H_1^{(0)}+\frac{1}{u^2}H_0^{(0)}\right), \tag{3.20}$$

$$\begin{aligned}K^{(2)}=\frac{1}{2}e^u\Bigg(&\frac{dH_0^{(2)}}{du}+\frac{1}{u}\frac{dH_1^{(1)}}{du}-\frac{1}{u}\frac{dH_0^{(1)}}{du}\\&+\frac{1}{3u^2}\frac{dH_2^{(0)}}{du}-\frac{1}{u^2}\frac{dH_1^{(0)}}{du}\\&+\frac{2}{3u^2}\frac{dH_0^{(0)}}{du}+\frac{1}{u^2}H_1^{(1)}+\frac{1}{u^2}H_0^{(1)}\\&+\frac{2}{3u^3}H_2^{(0)}-\frac{2}{3u^3}H_0^{(0)}\Bigg),\end{aligned} \tag{3.21}$$

$$K^{(3)} = \frac{1}{2}e^{u}\Big(\frac{dH_0^{(3)}}{du} + \frac{1}{u}\frac{dH_1^{(2)}}{du} - \frac{1}{u}\frac{dH_0^{(2)}}{du} + \frac{1}{3u^2}\frac{dH_2^{(1)}}{du} - \frac{1}{u^2}\frac{dH_1^{(1)}}{du} + \frac{2}{3u^2}\frac{dH_0^{(1)}}{du} + \frac{1}{15u^3}\frac{dH_3^{(0)}}{du} - \frac{1}{3u^3}\frac{dH_2^{(0)}}{du} + \frac{3}{5u^3}\frac{dH_1^{(0)}}{du} - \frac{1}{3u^3}\frac{dH_0^{(0)}}{du} + \frac{1}{u^2}H_1^{(2)} + \frac{1}{u^2}H_0^{(2)} + \frac{2}{3u^3}H_2^{(1)} - \frac{2}{3u^3}H_0^{(1)} + \frac{1}{5u^4}H_3^{(0)} - \frac{1}{3u^4}H_2^{(0)} - \frac{1}{5u^4}H_1^{(0)} + \frac{1}{3u^4}H_0^{(0)}\Big). \tag{3.22}$$

As noted previously, $K^{(m)}$ depends on the Legendre transforms of the expansion of h, and not on the terms in the expansion of h explicitly.

D. Boundary conditions on $H_n^{(m)}(u)$

Using (3.12) the boundary condition (2.26) yields the following boundary condition on $h(u,x)$ at $u = 1/a$:

$$h\left(\frac{1}{a},x\right) = \exp[\epsilon a(1+x)] - \frac{K(X)}{K(0)}\Delta\exp[\epsilon a(1-x)], \tag{3.23}$$

where $\Delta = \exp(-1/a)$. If we expand (3.23) in powers of ϵ and equate the coefficients of ϵ^m we find that

$$h^{(m)}\left(\frac{1}{a},x\right) = \frac{a^m(1+x)^m}{m!} - \Delta\sum_{p=0}^{m}K^{(p)}a^{m-p}\times\frac{(1-x)^{m-p}}{(m-p)!}, \quad m\geqslant 0. \tag{3.24}$$

We first note the following general results. The right-hand side of (3.24) is a polynomial in x of degree m and therefore it can be expressed as a sum of Legendre polynomials of degree $\leqslant m$. Hence it follows directly from the orthogonality relation for Legendre polynomials that

$$H_n^{(m)}(1/a) = 0, \quad \forall n \geqslant m+1. \tag{3.25}$$

We see also that $H_n^{(m)}(1/a)$ for $n \leqslant m$ depend only on those $K^{(p)}$ with $0 \leqslant p \leqslant m-n$ and in particular only $H_0^{(m)}(1/a)$ depends on $K^{(m)}$.

Equations (3.19)–(3.22) show which Legendre transforms need to be evaluated in order to calculate $K(X)/K(0)$ to third order in ϵ. The boundary conditions on the required Legendre transforms, and which are not covered by (3.25), are

$$H_0^{(0)}(1/a) = 2 - 2\Delta K^{(0)}, \tag{3.26}$$

$$H_0^{(1)}(1/a) = 2a - 2\Delta(aK^{(0)} + K^{(1)}), \tag{3.27}$$

$$H_1^{(1)}(1/a) = \tfrac{2}{3}a + \tfrac{2}{3}\Delta aK^{(0)}, \tag{3.28}$$

$$H_0^{(2)}(1/a) = \tfrac{4}{3}a^2 - 2\Delta(\tfrac{2}{3}a^2K^{(0)} + aK^{(1)} + K^{(2)}), \tag{3.29}$$

$$H_1^{(2)}(1/a) = \tfrac{2}{3}a^2 + \tfrac{2}{3}\Delta(a^2K^{(0)} + aK^{(1)}), \tag{3.30}$$

$$H_0^{(3)}(1/a) = \tfrac{2}{3}a^3 - \tfrac{2}{3}\Delta(a^3K^{(0)} + 2a^2K^{(1)} + 3aK^{(2)} + 3K^{(3)}). \tag{3.31}$$

The remaining boundary condition on $H_n^{(m)}(u)$ is at $u = 0$. From (3.8) and (3.12) it follows that

$$h(r,\theta) = O(1/r), \quad \text{as } r\to\infty, \tag{3.32}$$

and hence $h(u,x) = 0$ at $u = 0$; thus

$$H_n^{(m)}(0) = 0, \quad \forall n\geqslant 0,\ m\geqslant 0. \tag{3.33}$$

Equations (3.19)–(3.22) for $K^{(m)}$ and the system of coupled ordinary differential equations for $H_n^{(m)}(u)$ given by (3.15) with $m\geqslant 0$, $n\geqslant 0$ form the basis of the perturbation analysis, the problem being to calculate $K^{(m)}$ to third order in ϵ by first solving (3.15) for the required Legendre transforms subject to the boundary conditions (3.25)–(3.31) at $u = 1/a$ and (3.33) at $u = 0$.

IV. PERTURBATION SOLUTION TO THIRD ORDER IN ε

A. A general result

We will prove that $H_n^{(m)}(u) \equiv 0$ for any given $n \geqslant m+1$ provided

$$\sum_{p=0}^{r}\frac{(-1)^p(r+p)!a^p}{(r-p)!p!} \pm \Delta\sum_{p=0}^{r}\frac{(r+p)!a^p}{(r-p)!p!} \neq 0, \tag{4.1}$$

for all integers r such that $n-m \leqslant r \leqslant n+m$, where the $+$ and $-$ signs refer to the cases of r odd and r even, respectively.

Suppose first that $m = 0$ and consider any given $n \geqslant 1$. Since by definition $H_n^{(-1)}(u) \equiv 0$, $\forall n \geqslant 0$, (3.15) with $m = 0$ is

$$\frac{d^2H_n^{(0)}}{du^2} + \frac{dH_n^{(0)}}{du} - \frac{n(n+1)}{u^2}H_n^{(0)} = 0. \tag{4.2}$$

But the general solution of (4.2) is

$$H_n^{(0)}(u) = A_n\sum_{p=0}^{n}\frac{(-1)^p(n+p)!}{(n-p)!p!u^p} + B_n\sum_{p=0}^{n}\frac{(n+p)!e^{-u}}{(n-p)!p!u^p}, \tag{4.3}$$

where A_n and B_n are constants. But to satisfy the boundary condition (3.33) at $u = 0$ we require $A_n = B_n$ if n is odd and $A_n = -B_n$ if n is even. Hence the boundary condition (3.25) at $u = 1/a$, which applies only for $n \geqslant 1$ [although (4.2) is actually valid for $n \geqslant 0$], will be satisfied provided

$$A_n\left(\sum_{p=0}^{n}\frac{(-1)^p(n+p)!a^p}{(n-p)!p!} \pm \Delta\sum_{p=0}^{n}\frac{(n+p)!a^p}{(n-p)!p!}\right) = 0, \tag{4.4}$$

where the $+$ and $-$ signs refer to odd and even n, respectively. Hence if (4.1) is satisfied for $r = n$, then $A_n = B_n = 0$ and $H_n^{(0)}(u) \equiv 0$ for any given $n \geqslant 1$.

Consider now any $m > 0$ and any given $n \geqslant m+1$. From (3.15), the right-hand side of the differential equation for $H_n^{(m)}$ depends on $H_{n-1}^{(m-1)}$, $H_n^{(m-1)}$, and $H_{n+1}^{(m-1)}$. In order to determine these three Legendre transforms we in turn consider the differential equations which they satisfy and note that the right-hand sides of these three differential equations together depend on $H_{n-2}^{(m-2)}$, $H_{n-1}^{(m-2)}$, $H_n^{(m-2)}$, $H_{n-1}^{(m-2)}$, and $H_{n-2}^{(m-2)}$. Continuing in this way we are lead to consider the $2m-1$ differential equations for $H_r^{(1)}$, $n-m+1 \leqslant r \leqslant n+m-1$, the right-hand sides of which together depend on $H_r^{(0)}$, $n-m \leqslant r \leqslant n+m$. But since $n \geqslant m+1$ the Legendre transforms $H_r^{(0)}$, $n-m \leqslant r \leqslant n+m$,

are all zero provided (4.1) is satisfied for $n - m \leqslant r \leqslant n + m$. Thus the $2m - 1$ differential equations for $H_r^{(1)}$, $n - m + 1 \leqslant r \leqslant n + m - 1$, all have the same form as (4.2) for $H_n^{(0)}$ provided (4.1) is satisfied for $n - m \leqslant r \leqslant n + m$, and exactly the same argument as used for $H_n^{(0)}$ shows that $H_r^{(1)}(u) \equiv 0$ for $n - m + 1 \leqslant r \leqslant n + m - 1$. [The boundary condition (3.25) always applies because $n \geqslant m + 1$ and the conditions that (4.1) be satisfied for $n - m + 1 \leqslant r \leqslant n + m - 1$ are a subset of those already imposed.] Repeating this argument we can show that if $s \leqslant m$ then $H_r^{(s)}(u) \equiv 0$ for $n - m + s \leqslant r \leqslant n + m - s$ provided (4.1) is satisfied for $n - m \leqslant r \leqslant n + m$ and in particular for $H_n^{(m)}(u) \equiv 0$.

We see from (3.19)–(3.22) that we will require this general result only when considering the Legendre transforms $H_1^{(0)}(u)$, $H_2^{(0)}(u)$, $H_3^{(0)}(u)$, and $H_2^{(1)}(u)$. These functions will be identically zero if (4.1) is satisfied for $r = 1,2,3$, i.e., if

$$1 - 2a + \Delta(1 + 2a) \neq 0, \tag{4.5}$$

$$1 - 6a + 12a^2 - \Delta(1 + 6a + 12a^2) \neq 0, \tag{4.6}$$

$$1 - 12a + 60a^2 - 120a^3 + \Delta(1 + 12a + 60a^2 + 120a^3) \neq 0. \tag{4.7}$$

In this article we will assume that a is such that conditions (4.5)–(4.7) are satisfied so that

$$H_1^{(0)}(u) = H_2^{(0)}(u) = H_3^{(0)}(u) = H_2^{(1)}(u) = 0. \tag{4.8}$$

B. Zero order in ϵ

It follows from (3.19) that to evaluate $K^{(0)}$ we need to determine only $H_0^{(0)}(u)$. Now, (3.15) with $m = n = 0$ is

$$\frac{d^2H_0^{(0)}}{du^2} + \frac{dH_0^{(0)}}{du} = 0, \tag{4.9}$$

the general solution of which is

$$H_0^{(0)}(u) = A_0^{(0)}e^{-u} + B_0^{(0)}, \tag{4.10}$$

where $A_0^{(0)}$ and $B_0^{(0)}$ are constants. Imposing the boundary conditions (3.26) at $u = 1/a$ and (3.33) at $u = 0$, gives

$$A_0^{(0)} = -B_0^{(0)} = -2 + 2\Delta(K^{(0)} - 1)/(1 - \Delta), \tag{4.11}$$

and on substituting (4.10) into (3.19) we obtain

$$K^{(0)} = -\tfrac{1}{2}A_0^{(0)}. \tag{4.12}$$

Equations (4.11) and (4.12) give an algebraic equation for $K^{(0)}$; we find that $K^{(0)} = 1$, as required.

Finally (4.10) and (4.11) with $K^{(0)} = 1$ imply that

$$H_0^{(0)}(u) = \overline{H}_0^{(0)}(u) = 2(1 - e^{-u}), \tag{4.13}$$

where $\overline{H}_0^{(0)}(u)$ is the solution for $a = 0$.

C. First order in ϵ

We see from (3.20) and (4.8) that to evaluate $K^{(1)}$, $H_0^{(1)}(u)$ remains to be determined. Now (3.15) with $m = 1$, $n = 0$ is

$$\frac{d^2H_0^{(1)}}{du^2} + \frac{dH_0^{(1)}}{du} = \frac{(2 - u)}{u^3}H_0^{(0)} - \frac{2}{u^2}\frac{dH_0^{(0)}}{du}, \tag{4.14}$$

which, using the zero-order solution (4.13), can be rewritten as

$$\frac{d^2H_0^{(1)}}{du^2} + \frac{dH_0^{(1)}}{du} = \frac{2(2 - u)}{u^3} - \frac{2(2 + u)}{u^3}e^{-u}. \tag{4.15}$$

We solve (4.15) by initially treating it as a first-order differential equation in $dH_0^{(1)}/du$; its integrating factor is e^u. On noting that

$$\int \frac{(2 - u)}{u^3}e^u\,du = \frac{-e^u}{u^2} + \text{const}, \tag{4.16}$$

$$\int \frac{(1 + u)}{u^2}e^{-u}\,du = \frac{-e^{-u}}{u} + \text{const}, \tag{4.17}$$

we find that the general solution of (4.15) is

$$H_0^{(1)}(u) = (2/u)(1 - e^{-u}) + A_0^{(1)}e^{-u} + B_0^{(1)}, \tag{4.18}$$

where $A_0^{(1)}$ and $B_0^{(1)}$ are constants. Applying the boundary conditions (3.27) at $u = 1/a$ and (3.33) at $u = 0$, we find that

$$A_0^{(1)} = -2 - \frac{2\Delta(1 - K^{(1)})}{1 - \Delta}, \quad B_0^{(1)} = \frac{2\Delta(1 - K^{(1)})}{1 - \Delta}. \tag{4.19}$$

Using (4.13) and (4.18), (3.20) becomes

$$K^{(1)} = -\tfrac{1}{2}A_0^{(1)}, \tag{4.20}$$

where $A_0^{(1)}$ is the constant of integration in (4.18). We note the similarity in form between (4.20) and (4.12); it appears to be a general result, which we will see is satisfied for second and third orders, that $K^{(m)} = -\frac{1}{2}A_0^{(m)}$, where $A_0^{(m)}$ is the constant of integration multiplying e^{-u} in $H_0^{(m)}(u)$. Solving (4.19) and (4.20) for $K^{(1)}$ gives $K^{(1)} = 1$. Thus correct to $O(\epsilon)$

$$K(X)/K(0) = 1 + \epsilon, \tag{4.21}$$

and we therefore obtain the interesting result that there is no correction to Onsager's linear law[1] due to $a \neq 0$.

Finally using (4.18) and (4.19) with $K^{(1)} = 1$, we have

$$H_0^{(1)}(u) = \overline{H}_0^{(1)}(u) = (2/u)[1 - (1 + u)e^{-u}], \tag{4.22}$$

where $\overline{H}_0^{(1)}(u)$ is the solution for $a = 0$.

D. Second order in ϵ

Since there is no correction to $K(X)/K(0)$ due to $a \neq 0$ in the zero- and first-order terms, it is necessary to go to higher-order terms in ϵ. As well as evaluating the second-order contribution $K^{(2)}$ here, for which we do find a correction, we will also consider the third-order term $K^{(3)}$ in the next subsection. We find a correction to $K^{(3)}$ due to $a \neq 0$ which shows that it is not only the even-order terms which are modified. In evaluating $K^{(0)}$ and $K^{(1)}$ the general procedure of solution has been established above. In obtaining the second- and third-order terms we will therefore merely list relevant intermediate results, emphasizing only new features in the calculation.

It follows from (3.21) and (4.8) that to calculate $K^{(2)}$ it remains to determine only $H_1^{(1)}(u)$ and $H_0^{(2)}(u)$. The form of the differential equation for $H_1^{(1)}(u)$ is slightly different from previous equations. From (3.15) with $m = n = 1$ we have

$$\frac{d^2H_1^{(1)}}{du^2} + \frac{dH_1^{(1)}}{du} - \frac{2}{u^2}H_1^{(1)} = -\frac{1}{3u^2}H_0^{(0)}, \tag{4.23}$$

where $H_0^{(0)}(u)$ is given by (4.13). This can be rewritten as a first-order differential equation in the first derivative by noting that $(1 - 2/u)$ is a particular solution of (4.23) with the

right-hand side set equal to zero and by introducing the new dependent variable $G_1^{(1)}(u)$ defined by

$$H_1^{(1)}(u) = (1 - 2/u)G_1^{(1)}(u); \tag{4.24}$$

Eq. (4.23) becomes

$$\frac{d^2G_1^{(1)}}{du^2} + \frac{(u^2 - 2u + 4)}{u(u-2)}\frac{dG_1^{(1)}}{du} = -\frac{2(1 - e^{-u})}{3u(u-2)}. \tag{4.25}$$

This equation can be solved as previously by treating it as a first-order differential equation in $dG_1^{(1)}/du$; its integrating factor is $[(u-2)^2/u^2]e^u$. On making use of (4.16) and noting that

$$\int \frac{(1-u)}{(u-2)^2}e^{-u}\,du = \frac{e^{-u}}{u-2} + \text{const}, \tag{4.26}$$

$$\int \frac{u^2}{(u-2)^2}e^{-u}\,du = \frac{-(u+2)}{u-2}e^{-u} + \text{const}, \tag{4.27}$$

we find that the solution to (4.23) satisfying the boundary conditions (3.28) and (3.33) is

$$H_1^{(1)}(u) = \bar{H}_1^{(1)}(u) + \frac{\gamma(u+2)}{u}e^{-u} + \frac{\gamma(u-2)}{u}, \tag{4.28}$$

where

$$\gamma = 2(1 + 2a)\Delta\,/3[1 - 2a + (1 + 2a)\Delta\,] \tag{4.29}$$

and

$$\bar{H}_1^{(1)}(u) = (2/3u)[1 - (1 + u)e^{-u}], \tag{4.30}$$

$\bar{H}_1^{(1)}(u)$ being the solution for $a = 0$.

The differential equation for $H_0^{(2)}$ is obtained by putting $m = 2$, $n = 0$ in (3.15):

$$\frac{d^2H_0^{(2)}}{du^2} + \frac{dH_0^{(2)}}{du} = \frac{(2-u)}{u^3}H_0^{(1)} - \frac{1}{u^2}H_1^{(1)} - \frac{2}{u^2}\frac{dH_0^{(1)}}{du}. \tag{4.31}$$

The terms on the right-hand side of (4.31) are known and it can be solved in the same way as (4.15). The integrations required are (4.16), (4.17), and

$$\int \frac{(3-u)}{u^4}e^u\,du = \frac{-e^u}{u^3} + \text{const}, \tag{4.32}$$

$$\int \frac{(u^2 + 2u + 2)}{u^3}e^{-u}\,du = \frac{-(1+u)}{u^2}e^{-u} + \text{const}. \tag{4.33}$$

We find that the solution to (4.31) which satisfies the boundary conditions (3.29) and (3.33) is

$$H_0^{(2)}(u) = (4/3u^2)[1 - (1 + u)e^{-u}] + (\gamma/u)(1 - e^{-u}) + A_0^{(2)}e^{-u} + B_0^{(2)}, \tag{4.34}$$

where

$$A_0^{(2)} = -\frac{2}{3} - \frac{2(1-a)\Delta}{3(1-\Delta)} - \frac{(1 - a + a\Delta)\gamma}{1-\Delta} + \frac{2\Delta}{1-\Delta}K^{(2)}, \tag{4.35}$$

$$B_0^{(2)} = \frac{2(1-a)\Delta}{3(1-\Delta)} + \frac{(\Delta + a\Delta - a)\gamma}{1-\Delta} - \frac{2\Delta}{1-\Delta}K^{(2)}. \tag{4.36}$$

On substituting (4.34), (4.28), (4.22), (4.13), and (4.8) into (3.21) we find that

$$K^{(2)} = -\tfrac{1}{2}A_0^{(2)}, \tag{4.37}$$

and solving the algebraic equations (4.35) and (4.37) for $K^{(2)}$ we obtain

$$K^{(2)} = \tfrac{1}{3} + \Delta\,/3[1 - 2a + (1 + 2a)\Delta\,]. \tag{4.38}$$

A correction to Onsager's result (1.2) due to $a \neq 0$ therefore enters for the first time at the second order in ϵ. Equation (4.38) is well defined because of condition (4.5).

Using (4.38) for $K^{(2)}$, $A_0^{(2)}$ and $B_0^{(2)}$ given by (4.35) and (4.36) can be simplified and (4.34) becomes

$$H_0^{(2)}(u) = \bar{H}_0^{(2)}(u) + \gamma\left(\frac{1}{u} - \frac{2a}{1+2a}\right) - \gamma\left(\frac{1}{u} + \frac{1}{1+2a}\right)e^{-u}, \tag{4.39}$$

where

$$\bar{H}_0^{(2)}(u) = (4/3u^2)[1 - (1 + u + u^2/2)e^{-u}] \tag{4.40}$$

is the solution for $a = 0$.

E. Third order in ϵ

Finally, we outline the derivation of $K^{(3)}$. We see from (3.22) and (4.8) that to evaluate $K^{(3)}$ only $H_1^{(2)}(u)$ and $H_0^{(2)}(u)$ remain to be determined.

Consider first $H_1^{(2)}(u)$. Equation (3.15) with $m = 2$, $n = 1$ is

$$\frac{d^2H_1^{(2)}}{du^2} + \frac{dH_1^{(2)}}{du} - \frac{2}{u^2}H_1^{(2)} = \frac{(2-u)}{u^3}H_1^{(1)} - \frac{1}{3u^2}H_0^{(1)} - \frac{2}{u^2}\frac{dH_1^{(1)}}{du}. \tag{4.41}$$

The terms on the right-hand side of (4.41) are known, and this equation can be solved in exactly the same way as (4.23) for $H_1^{(1)}(u)$. We find that its solution satisfying the boundary conditions (3.30) and (3.33) is

$$H_1^{(2)}(u) = \bar{H}_1^{(2)}(u) + \frac{\gamma(u^2-4)}{2u^2} + \frac{\gamma(u^2 + 4u + 4)}{2u^2}e^{-u}, \tag{4.42}$$

where

$$\bar{H}_1^{(2)}(u) = (2/3u^2)[1 - (1 + u + u^2/2)e^{-u}], \tag{4.43}$$

$\bar{H}_1^{(2)}$ being the solution for $a = 0$.

The differential equation for $H_0^{(3)}(u)$ is obtained by setting $m = 3$, $n = 0$ in (3.15):

$$\frac{d^2H_0^{(3)}}{du^2} + \frac{dH_0^{(3)}}{du} = \frac{(2-u)}{u^3}H_0^{(2)} - \frac{1}{u^2}H_1^{(2)} - \frac{2}{u^2}\frac{dH_0^{(2)}}{du}, \tag{4.44}$$

where again the terms on the right-hand side are known functions. Equation (4.44), subject to the boundary conditions (3.31) and (3.33), is solved in the same way as (4.15). We find that

$$H_0^{(3)}(u) = \frac{2}{3u^3} - \frac{(u^2 + 2u + 2)}{3u^3}e^{-u} + \frac{\gamma(u+2)}{2u^2} - \frac{2a\gamma}{(1+2a)u} - \frac{\gamma(3u+2)}{2u^2}e^{-u} + \frac{2a\gamma}{(1+2a)u}e^{-u} + A_0^{(3)}e^{-u} + B_0^{(3)}, \tag{4.45}$$

where

$$A_0^{(3)} = -\frac{1}{9} - \frac{\Delta}{9(1-\Delta)} - \frac{2(1-a)\Delta}{3(1-\Delta)[1-2a+(1+2a)\Delta]} + [2\Delta/(1-\Delta)]K^{(3)}. \quad (4.46)$$

It is not necessary to determine the other constant $B_0^{(3)}$ in order to calculate $K^{(3)}$ because it is the derivative of $H_0^{(3)}(u)$ and not $H_0^{(3)}(u)$ itself which occurs in (3.22).

Now if we substitute (4.45), (4.42), (4.39), (4.28), (4.22), (4.13), and (4.8) into (3.22) we find that

$$K^{(3)} = -\tfrac{1}{2}A_0^{(3)}, \quad (4.47)$$

and on solving (4.46) and (4.47) for $K^{(3)}$ we obtain

$$K^{(3)} = \tfrac{1}{18} + (1-a)\Delta/3[1-2a+(1+2a)\Delta]. \quad (4.48)$$

The solution for $H_0^{(3)}(u)$ may be completed if desired by determining the constant $B_0^{(3)}$ and simplifying $A_0^{(3)}$ and $B_0^{(3)}$ using (4.48).

F. Higher-order terms

Using the perturbation technique described above, $K(X)/K(0)$ can be calculated in principle to any desired order in ϵ although the calculation of higher-order terms soon becomes complicated. We were able to solve the differential equations (4.23) and (4.41) by using a particular solution of the associated homogeneous equation. This can always be done with (3.15) for general n. For

$$Y_n(u) = \sum_{p=0}^{n} \frac{(-1)^p (n+p)!}{(n-p)!p!u^p} \quad (4.49)$$

is a particular solution of (3.15) with the right-hand side set equal to zero, and by introducing the new unknown function $G_n^{(m)}(u)$ defined by

$$H_n^{(m)}(u) = Y_n(u)G_n^{(m)}(u), \quad (4.50)$$

(3.15) can be reduced to a first-order differential equation in $dG_n^{(m)}/du$; its integrating factor is $Y^2(u)e^u$. We see from (4.3) that another particular solution of the homogeneous equation associated with (3.15) is

$$Z_n(u) = \sum_{p=0}^{n} \frac{(n+p)!e^{-u}}{(n-p)!p!u^p}. \quad (4.51)$$

A generally true result which greatly simplifies our analysis is the fact, noted in Sec. III D, that only $H_0^{(m)}(1/a)$ depends on $K^{(m)}$ in the mth-order calculation. A study of the Legendre transforms which occur on the right-hand sides of, for example, in the case $m = 3$, (4.41) and (4.44), shows that we should expect that in the mth-order calculation, $H_0^{(m)}(u)$ must always be evaluated *last*. Hence in the calculation of $K^{(m)}$ it is only the constants of integration in $H_0^{(m)}(u)$ which depend on $K^{(m)}$. The other Legendre transforms required in the calculation of $K^{(m)}$ depend on $K^{(p)}$, $0 \leqslant p \leqslant m-1$, which are already known.

To third order in ϵ, i.e., for $m = 0,1,2$, and 3 we found that

$$K^{(m)} = -\tfrac{1}{2}A_0^{(m)}, \quad (4.52)$$

where $A_0^{(m)}$ is the constant of integration multiplying e^{-u} in $H_0^{(m)}(u)$. It seems likely that this result will be true in general and if so, it provides a useful check on our analysis.

V. DISCUSSION AND CONCLUSIONS

We have found that when $a \neq 0$,

$$\frac{K(X)}{K(0)} = 1 + \epsilon + \left(\frac{1}{3} + \frac{\Delta}{3[1-2a+(1+2a)\Delta]}\right)\epsilon^2 + \left(\frac{1}{18} + \frac{(1-a)\Delta}{3[1-2a+(1+2a)\Delta]}\right)\epsilon^3 + O(\epsilon^4), \quad (5.1)$$

the power of the method and form of solution lying in the rapid convergence of the infinite series for $K(X)/K(0)$ with $a = 0$.[3] The constant of proportionality in Onsager's linear law remains unchanged when $a \neq 0$ which implies that in the low field limit $K(X)/K(0)$ is insensitive to the value of a. The correction to higher orders in (5.1) is in the form of an additional term proportional to $\Delta = \exp(-1/a)$. For simple ions, a is typically 0.01 to 0.05 and so the correction in (5.1) is negligible, which explains why (1.1) is in such good agreement with experiment in such cases. For the more complicated situations described in Sec. I, a is much larger and the correction will clearly be appreciable; in these cases (5.1) also shows that the effect of the inner boundary condition is really a high field effect. The latter limit has yet to be treated mathematically in the case $a \neq 0$.

The perturbation technique which we established[3] to give an alternative derivation of Onsager's result (1.1) has had to be modified in this paper because the functions $h^{(m)}(u,x)$ are no longer separable in u and x when $a \neq 0$. This can be traced essentially to the fact that the boundary condition (3.23) is not separable in a and x. (The problem treated above is perhaps the simplest problem for which the full power of the present treatment is necessary.) A consequence of this is that the Legendre transforms $H_n^{(m)}(u)$ for the same value of m but different values of n are no longer constant multiples of each other as they are in the problem when $a = 0$: e.g., $H_0^{(2)}(u)$ and $H_1^{(2)}(u)$ given by (4.39) and (4.42), respectively, are not constant multiples of each other but become so in the limit $a \to 0$. This is also the case with $H_0^{(1)}(u)$ and $H_1^{(1)}(u)$ given by (4.22) and (4.28), respectively. We[3] were able to take advantage of this observation when $a = 0$ to calculate the mth-order term in the expansion of $K(X)/K(0)$ but to do that in the present problem would be more difficult.

[1] L. Onsager, J. Chem. Phys. **2**, 599 (1934).
[2] L. Onsager, Ph.D. thesis, Yale University, 1935.
[3] D. P. Mason and D. K. McIlroy, Proc. R. Soc. London Ser. A **359**, 303 (1978).
[4] C. J. Tranter, *Integral Transforms in Mathematical Physics* (Metheun, London, 1966), 3rd ed.
[5] A. H. Nayfeh, *Perturbation Methods* (Wiley, New York, 1973).
[6] J. D. Murray, *Asymptotic Analysis* (Clarendon, Oxford, 1974).
[7] D. P. Mason and D. K. McIlroy, SIAM J. Appl. Math. **40**, 298 (1981).
[8] H. S. Harned and B. B. Owen, *Physical Chemistry of Electrolytic Solutions* (Reinhold, New York, 1958), 3rd ed.
[9] D. P. Mason and D. K. McIlroy, J. Chem Soc. Far. Trans. II **74**, 2019 (1978).
[10] D. K. McIlroy and D. P. Mason, SIAM J. Appl. Math. **40**, 283 (1981).
[11] L. Onsager and R. M. Fuoss, J. Phys. Chem. **36**, 2689 (1932).
[12] D. P. Mason and D. K. McIlroy, J. Math. Phys. **19**, 1340 (1978).
[13] R. C. Tolman, *Statistical Mechanics* (Reinhold, New York, 1927), lst ed.
[14] W. W. Bell, *Special Functions for Scientists and Engineers* (van Nostrand, London, 1968).

Perfect fluid models of Bianchi type-I in modified Brans–Dicke cosmology

T. Singh and Tarkeshwar Singh
Applied Mathematics Section, Institute of Technology, Banaras Hindu University, Varanasi-221005, India

(Received 12 December 1983; accepted for publication 27 January 1984)

The spatially homogeneous and anisotropic Bianchi type-I cosmological solutions of modified Brans–Dicke theory containing barotropic fluid have been obtained by imposing a condition on the cosmological parameter $\Lambda(\phi)$. The particular case when $p=\rho$ has been discussed in detail.

PACS numbers: 98.80.Dr

I. INTRODUCTION

After the cosmological constant was first introduced into general relativity by Einstein, its significance was studied by various cosmologists (for example, see Ref. 1) but no satisfactory result of its meaning has been reported so far. Zel'dovich[2] has tried to visualize the meaning of this term from the theory of elementary particles. Further, Linde[3] has argued that the cosmological term arises from spontaneous symmetry breaking and suggested that the term is not a constant but a function of temperature. Also Dreitlein[4] connects the mass of Higg's scalar boson with both the cosmological term and the gravitational constant. In cosmology, the term may be understood by incorporation with Mach's principle, which suggests the acceptance of the Brans–Dicke Lagrangian as a realistic case.[5] The investigation of particle physics within the context of the Brans–Dicke Lagrangian[6] has stimulated the study of the cosmological term with a modified Brans–Dicke Lagrangian in cosmology and elementary particle physics. Recently considering the proposition of Bergmann[7] and Wagoner[8] that the cosmological term Λ should be a function of a scalar field ϕ, Endo and Fukui[9] obtained modified Brans–Dicke field equations. The variable cosmological term has also been discussed by Endo and Fukui.[10]

In this paper we have considered a cosmological model filled with perfect fluid in spatially homogeneous Bianchi type-I space-time. The particular case when $p=\rho$ has been derived and discussed in detail. A physically reasonable model originating from a singularity is possible only in the presence of a negative energy density scalar field. However, the fundamental properties of the model are similar to the Robertson–Walker model for zero constant curvature as discussed by Endo and Fukui[9] and the Bianchi type-I dust model by Banerjee and Santos[11] (both in the presence of a positive energy density scalar field). Here also the problem of making Λ correspond to a mass term in particle physics remains (see Endo and Fukui,[9] p. 838). Further, negative energy density scalar-tensor solutions may turn out to be useful, e.g., in particle physics (Refs. 12 and 13).

II. FIELD EQUATIONS

The field equations for the modified Brans–Dicke theory with the introduction of $\Lambda(\phi)$ obtained by Endo and Fukui[9] are

$$G_{ij}+g_{ij}\Lambda=(K/\phi)T_{ij}-(\omega/\phi^2)\times(\phi,i\phi,j-\tfrac{1}{2}g_{ij}\phi_{,\alpha}\phi^{,\alpha})-(1/\phi)(\phi;\mathrm{ij}-g_{ij}\Box\phi), \tag{1}$$

$$\Lambda-\phi\frac{\partial\Lambda}{\partial\phi}=\frac{K}{2\phi}T-\frac{2\omega+3}{2\phi}\Box\phi. \tag{2}$$

Here T_{ij} is the energy-momentum tensor for a perfect fluid given by

$$T_{ij}=(p+\rho)U_iU_j-pg_{ij}, \tag{3}$$

with

$$g_{ij}U^iU^j=1, \tag{4}$$

where p and ρ are proper pressure and energy density, respectively, and U^i are the components of the fluid four-velocity. We assume the coordinates to be co-moving so that

$$U^1=U^2=U^3=0 \quad \text{and} \quad U^4=(g_{44})^{-1/2}.$$

It has been assumed further that the matter and scalar field are related through

$$\Box\phi=K\mu T/(2\omega+3), \tag{5}$$

where the constant μ shows how much our theory including $\Lambda(\phi)$ deviates from that of Brans–Dicke and as usual ω is the coupling constant. Substituting relation (5) into (2) we obtain

$$\Lambda-\phi\frac{\partial\Lambda}{\partial\phi}=A\frac{\Box\phi}{\phi}, \tag{6}$$

where A is a constant defined by

$$A=[(2\omega+3)/2](1/\mu-1). \tag{7}$$

As it has been assumed that Λ is a function of ϕ only, from (6) we can easily conclude that $\Box\phi=f(\phi)$. Here we make another assumption that the functional relation $f(\phi)$ is of the form $f(\phi)=m\phi^n$, where m and n are arbitrary constants. We assume that to explain the origin of mass of an elementary particle by symmetry breaking ϕ^n terms are introduced.[14] With this choice of $f(\phi)$ Eq. (6) reduces to

$$\phi\frac{\partial\Lambda}{\partial\phi}-\Lambda+Am\phi^{n-1}=0. \tag{8}$$

From (8) the solution for Λ is given by

$$\Lambda=[Am/(2-n)]\phi^{n-1}+D_1\phi, \quad \text{when } n\neq 2, \tag{9}$$

and

$$\Lambda=-Am\phi\ln\phi+D_2\phi, \quad \text{when } n=2. \tag{10}$$

Here D_1 and D_2 are integration constants.

III. SOLUTIONS FOR BIANCHI TYPE-I

The line element for the spatially homogeneous Bianchi type-I metric can be written as

$$ds^2=dt^2-e^{2\gamma}dx^2-e^{2\theta}dy^2-e^{2\psi}dz^2, \tag{11}$$

 0022-2488/84/092800-03$02.50

where γ, θ, and ψ are functions of t alone. The field equations (1) in terms of the line element (11) can be written as

$$G_0^0 = -\frac{9}{2}\left(\frac{R'}{R}\right)^2 + \frac{1}{2}(\gamma'^2 + \theta'^2 + \psi'^2)$$
$$= -\Lambda - \frac{K\rho}{\phi} - \frac{\omega}{2}\left(\frac{\phi'}{\phi}\right)^2 - \frac{\phi''}{\phi} + \frac{\Box\phi}{\phi}, \quad (12)$$

$$G_1^1 = -\theta'' - \psi''$$
$$- \frac{3R'}{2R}(-\gamma' + \theta' + \psi') - \tfrac{1}{2}(\gamma'^2 + \theta'^2 + \psi'^2)$$
$$= -\Lambda + \frac{Kp}{\phi} + \frac{\omega}{2}\left(\frac{\phi'}{\phi}\right)^2 - \gamma'\frac{\phi'}{\phi} + \frac{\Box\phi}{\phi}, \quad (13)$$

$$G_2^2 = -\gamma'' - \psi'' - \frac{3R'}{2R}(\gamma' - \theta' + \psi')$$
$$- \tfrac{1}{2}(\gamma'^2 + \theta'^2 + \psi'^2)$$
$$= -\Lambda + \frac{Kp}{\phi} + \frac{\omega}{2}\left(\frac{\phi'}{\phi}\right)^2 - \theta'\frac{\phi'}{\phi} + \frac{\Box\phi}{\phi}, \quad (14)$$

$$G_3^3 = -\gamma'' - \theta'' - \frac{3R'}{2R}(\gamma' + \theta' - \psi')$$
$$- \tfrac{1}{2}(\gamma'^2 + \theta'^2 + \psi'^2)$$
$$= -\Lambda + \frac{Kp}{\phi} + \frac{\omega}{2}\left(\frac{\phi'}{\phi}\right)^2 - \psi'\frac{\phi'}{\phi} + \frac{\Box\phi}{\phi}, \quad (15)$$

where the prime means differentiation with respect to time t and

$$R^3 = \exp(\gamma + \theta + \psi). \quad (16)$$

From the field equations (13)–(15), after subtracting one from the other and integrating, we obtain the relations

$$\theta' = \gamma' - C_1/R^3\phi, \quad (17)$$

$$\psi' = \gamma' - C_2/R^3\phi, \quad (18)$$

where C_1 and C_2 are integration constants and they measure the anisotropy of space because when $C_1 = C_2 = 0$, we have $\gamma = \theta = \psi$ (i.e., isotropic line element).

From the conservation equation $T^j_{i;j} = 0$, we obtain

$$\rho' = -(p + \rho)(\gamma' + \theta' + \psi'). \quad (19)$$

We take the equation of state to be that for a barotropic fluid,

$$P = (\lambda - 1)\rho, \quad 1 \leqslant \lambda \leqslant 2. \quad (20)$$

Then Eq. (19) reduces to

$$\rho'/\rho + 3\lambda R'/R = 0. \quad (21)$$

After integration, it gives

$$\rho = CR^{-3\lambda}, \quad (22)$$

where C is a constant of integration. Equations (5), (6), (8), and (22) give

$$R^{-3\lambda} = m\phi^n/d(4 - 3\lambda), \quad (23)$$

where $d = K\mu C/(2\omega + 3)$.

Now with the help of Eq. (22) we can write the density ρ in terms of the scalar field ϕ as

$$\rho = mC\phi^n/d(4 - 3\lambda). \quad (24)$$

From Eqs. (17), (18), and (23) one can easily obtain

$$3\gamma' = -n\phi'/\phi + (C_1 + C_2)/R^3\phi. \quad (25)$$

Now we consider the transformation

$$dt = R^{3(\lambda - 1)}\,d\tau. \quad (26)$$

Putting Eqs. (9), (17), (18), and (22) into (12) we have

$$QR^6\left(\frac{\phi^+}{\phi}\right)^2 - L\frac{R^{6(\lambda-1)}}{\phi^2} - M\frac{R^{6(\lambda-1)}}{\phi^{1-n+2n/\lambda}}$$
$$- S\frac{1}{\phi^{2n-1}} = 0, \quad (27)$$

where a cross (+) overhead denotes differentiation with respect to τ.

Now subtracting (13) from (12) and using (9), (17), and (18) we get

$$QR^6\left(\frac{\phi^+}{\phi}\right)^2 - L\frac{R^{6(\lambda-1)}}{\phi^2} - P\frac{R^{6(\lambda-1)}}{\phi^{1-n+2n/\lambda}} = 0, \quad (28)$$

where

$$Q = \frac{n^2}{\lambda^2} - \frac{3n}{\lambda} - \frac{3}{2}\omega,$$
$$L = C_1^2 + C_2^2 - C_1C_2,$$
$$M = m\left(\frac{d(4 - 3\lambda)}{m}\right)^{2/\lambda}\left(\frac{3A}{2 - n} + \frac{3(2\omega + 3)}{\mu(4 - 3\lambda)}\right),$$
$$S = 3D_1\left(\frac{d(4 - 3\lambda)}{m}\right)^2,$$
$$P = m\left(\frac{d(4 - 3\lambda)}{m}\right)^{2/\lambda}\left[\frac{n}{\lambda} + \frac{3}{2}\left(\frac{\lambda(2w + 3)}{\mu(4 - 3\lambda)} + 1\right)\right].$$

For Eqs. (27) and (28) to be compatible we must have $S = 0$ and $P = M$, which imply that

$$D_1 = 0, \quad (29)$$

and

$$\frac{n}{3\lambda} + \frac{1}{2}\left(\frac{\lambda(2\omega + 3)}{\mu(4 - 3\lambda)} + 1\right) = \frac{A}{2 - n} + \frac{2\omega + 3}{\mu(4 - 3\lambda)}. \quad (30)$$

The solution for ϕ may be obtained now by substituting (23) and (26) into (28). We thus have

$$t + E_1 = \pm\left(\frac{d(4 - 3\lambda)}{m}\right)^{1/\lambda}\sqrt{Q}$$
$$\times\int \frac{\phi^{-n/\lambda}}{(L + P\phi^{n+1-2n/\lambda})^{1/2}}, \quad (31)$$

where E_1 is a constant of integration.

Now we consider the cases of the dust-filled, radiation-dominated, and superdense stages of the universe one by one.

Case I: Dust-filled universe. This case corresponds to incoherent matter for which $p = 0$. Therefore putting $\lambda = 1$ in the Eq. (31) and integrating, one easily gets

$$\phi = \left\{\pm\left[(t + E_1)\frac{m}{d}\frac{(1 - n)}{2}\left(\frac{P}{Q}\right)^{1/2}\right]^2 - \frac{L}{P}\right\}^{1/(1-n)}. \quad (32)$$

This case has been discussed by Banerjee and Santos.[11]

Case II: Radiation-dominated universe. In this case $p = \rho/3$, i.e., $\lambda = \frac{4}{3}$. In the case of radiation-dominated uni-

verse we have $T = 0$, thus from Eq. (5) we get $\Box\phi = 0$, hence $m = 0$. Therefore, from Eq. (24) we have $\rho = 0$ and hence $p = 0$.

Therefore, for $\rho = 3p$ no solution will exist and the model reduces to the vacuum case.

Case III: Superdense universe. Zel'dovich[15] gave the equation of state for stiff matter by choosing $\lambda = 2$; this implies $p = \rho$. Therefore, inserting the value of λ in Eq. (31) we have

$$t + E_1 = \pm\left(-\frac{2d}{m}\right)^{1/2}\sqrt{Q}\times\int\frac{\phi^{-n/2}\,d\phi}{(L+P\phi)^{1/2}}. \tag{33}$$

The integral on the right can be evaluated[16] as

$$t + E_1 = \pm\left(-\frac{2d}{m}\right)^{1/2}\sqrt{Q}\left(\frac{(L+P\phi)^{-n/2}}{1-n} - \frac{n}{2}\frac{L(L+P)^{-n/2-1}}{n+1} - \frac{(n/2)(n/2+1)}{1.2}\frac{L^2(L+P\phi)^{-n/2-2}}{n+3} + \cdots + NL^{-n/2}\right)\left(\frac{2}{P^{1-n/2}(L+P\phi)^{-1/2}}\right), \tag{34}$$

where N is a constant.

Now the cosmological factor given by (9) and (29) is

$$\Lambda = Am/(2-n)\,\phi^{n-1}. \tag{35}$$

The density can now be written as a function of the cosmological factor. Substituting (35) into (24) we have

$$\rho = \frac{mC}{2d}\left(\frac{2-n}{Am}\right)^{n/(n-1)}\Lambda^{n/(n-1)}. \tag{36}$$

Also

$$R^6 = \frac{C}{\rho} = -\frac{2d}{m}\left(\frac{Am}{2-n}\right)^{n/(n-1)}\Lambda^{(n-1)/n}. \tag{37}$$

Final γ, θ, and ψ can be determined in terms of ϕ. Another important relation can be obtained from (5), which reduces to

$$\frac{d}{dt}(\phi' R^3) = -\left(\frac{2K\mu}{2\omega+3}\right)\rho R^3. \tag{38}$$

In view of (22) and (26) the Eq. (38) goes to the form

$$\frac{d\phi}{d\tau} = -\left(\frac{2K\mu C}{2\omega+3}\right)\tau = -2d\tau, \tag{39}$$

where d is a constant already defined. Integrating (39) we get

$$\phi = -d\tau^2. \tag{40}$$

The integration constant is taken as zero in order that we have the model starting from $\tau = 0$ when $\phi = 0$. From (39) and $d > 0$ when $2\omega + 3 > 0$ it follows that $d\phi/d\tau < 0$, which means that ϕ is a decreasing function of τ. This contradicts our choice of initial condition. Hence we take $d < 0$, i.e., $2\omega + 3 < 0$, which is possible only for negative energy scalar fields in scalar-tensor theories. Then $d\phi/d\tau > 0$ which means that ϕ is an increasing function of τ. Now if we consider an expanding universe with $R = 0$ when $\tau = 0$, R^3 is increasing with τ. This in view of (23) fixes the value of n always less than zero. So in this case $n < 0$ and at the epoch $\tau = 0$, $\phi \to 0$ we have $R^3 \to 0$ and $\rho \to \infty$. In the course of time the model expands and has infinite volume $R^3 \to \infty$ and $\rho \to 0$ as $\phi \to \infty$.

The general behavior of such an anisotropic homogeneous model for negative energy scalar fields is analogous to that of the zero curvature RW-model given by Endo and Fukui[9] and the Bianchi type-I dust solution of Banerjee and Santos[12] (both for positive energy scalar fields). Thus for the model exploding from the initial singularity $R^3 \to 0$ as $t \to 0$ automatically excludes the possibility for $n = 2$.

[1]V. Petrosian, Int. Astron. Union Symp. **63**, 31 (1975).
[2]Ya. B. Zel'dovich, Sov. Phys. Usp. **11**, 381 (1968).
[3]A. D. Linde, JETP Lett. **19**, 183 (1974).
[4]J. Dreitlein, Phys. Rev. Lett. **33**, 1243 (1974).
[5]C. Brans and R. H. Dicke, Phys. Rev. **124**, 925 (1961).
[6]Y. Fujii, Phys. Rev. D **9**, 874 (1974).
[7]P. G. Bergmann, Int. J. Theor. Phys. **1**, 25 (1968).
[8]R. V. Wagoner, Phys. Rev. D **1**, 3209 (1970).
[9]M. Endo and T. Fukui, Gen. Relat. Gravit. **8**, 833 (1977).
[10]M. Endo, and T. Fukui, Gen. Relat. Gravit. **14**, 769 (1982).
[11]A. Banerjee and N. O. Santos, Gen. Relat. Gravit. **14**, 559 (1982).
[12]T. Kodama, Phys. Rev. D **18**, 3529 (1978).
[13]G. Clement, Gen. Relat. Gravit. **13**, 747 (1981).
[14]S. Deser, Ann. Phys. (N.Y.) **59**, 248 (1970).
[15]Ya. B. Zel'dovich, Sov. Phys. JETP **14**, 1143 (1962).
[16]G. P. Bois, *Table of Indefinite Integrals* (Dover, New York, 1961), p. 91.

ERRATUM

Erratum: Scattering theory for the dilation group. I. Simple quantum mechanical scattering [J. Math. Phys. 24, 1797 (1983)]

J. V. Corbett
School of Mathematics and Physics, Macquarie University, North Ryde, New South Wales 2113, Australia

(Received 8 May 1984; accepted for publication 18 May 1984)

PACS numbers: 03.65.Nk, 99.10. + g

The proof of the simple relationship between the dilation wave operators, defined by Eq. (11), and the usual wave operators is false. It is based on the erroneous asymptotic relation given in Eqs. (A9), (A11), and (A14). This asymptotic relation does not hold in the L^2-norm, although it is true for the L^p-norm for any $p > 2$. The correct asymptotic relation in the L^2-norm is

$$\lim_{b\to\infty} \|\pi(u(b))f - \pi(v(b^{-1}))\pi(s(b))(\pi(W))^3 f\| = 0,$$

where $v(b)$, $s(b)$, and w are elements of $SL(2,\mathbb{R})$, given in the Appendix and $v(b) = \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}$. The definition of the dilation wave operators, Definition 5, can be changed to accommodate this but the new dilation wave operators then lose their simplicity.

 0022-2488/84/092803-01$02.50