JOURNAL OF
MATHEMATICAL
PHYSICS

VOLUME 5, NUMBER 7 July 1964

Convergence of Virial Expansions*

J. L. LEBOWITZ AND O. PENROSE†
Belfer Graduate School of Science, Yeshiva University, New York, New York
(Received 5 September 1963)

Some bounds are obtained on $\mathfrak{R}(V)$, the radius of convergence of the density expansion for the logarithm of the grand partition function of a system of interacting particles in a finite volume V, and on $\mathfrak{R}$, the radius of convergence of the corresponding infinite-volume expansion (the virial expansion). A common lower bound on $\mathfrak{R}(V)$ and $\mathfrak{R}$ is $0.28952/(u+1)B$, where $u \equiv \exp[-\text{Min}\, s^{-1}\sum_{i<j\leq s} 2\varphi(\mathbf{x}_i - \mathbf{x}_j)]/\kappa T$ [so that $u \geq 1$, with equality for nonnegative $\varphi(\mathbf{r})$], $B \equiv \int|e^{-\varphi(\mathbf{r})/\kappa T} - 1|\, d^3\mathbf{r}$, and $\varphi(\mathbf{r})$ is the binary interaction potential; the irreducible Mayer cluster integrals have the related upper bounds $\beta_k \leq [(u+1)B/0.28952]^k/k$[$u = 1$, when $\varphi(r) \geq 0$]. For potentials with hard cores the maximum density is an upper bound on $\mathfrak{R}(V)$, though possibly not on $\mathfrak{R}$; an example shows how both $\mathfrak{R}(V)$ and $\mathfrak{R}$ can be less than the maximum density, even if there is no phase transition. A theorem is proved, analogous to Yang and Lee's theorem on uniform convergence in the complex z plane, defining a class of domains in the complex ρ plane within which the operations $V \to \infty$ and $d/d\rho$ commute. This theorem is used to show that $\lim_{V\to\infty} \mathfrak{R}(V) \leq \mathfrak{R}$, and that there is no phase transition for $0 \leq \rho < 0.28952/(u+1)B$.

1. INTRODUCTION

RECENTLY several authors[1-3] have obtained upper and lower bounds for the radius of convergence $R(V)$ of the Mayer fugacity expansions.[4]

$$\mathcal{V}^{-1} \log \Xi(z, V) \equiv p(z, V)/\kappa T = \sum_l b_l(V) z^l, \quad (1.1)$$

$$\rho(z, V) = (z/\kappa T)\, dp(z, V)/dz = \sum_l l b_l(V) z^l. \quad (1.2)$$

Here $\Xi(z, V)$ and $\rho(z, V)$ are the grand partition function and the mean number density at fugacity z and temperature T for a system of particles with two-body interactions, confined to a spatial region V whose volume is $\mathcal{V}$. Boltzmann's constant is denoted by κ. The coefficients $b_l(V)$ are the finite-volume Mayer[4] cluster integrals. The s-particle distribution functions $n_s(\mathbf{x}_1 \cdots \mathbf{x}_s \mid z, V)$ can also be expanded as power series in z, with radius of convergence at least[3] $R(V)$.

The thermodynamic pressure and density are given[1-3] for small z by

$$p(z) \equiv \lim_{V\to\infty} p(z, V) = \kappa T \sum_l b_l z^l, \quad (1.3)$$

$$\rho(z) \equiv \lim_{V\to\infty} \rho(z, V) = \sum_l l b_l z^l, \quad (1.4)$$

where

$$b_l \equiv \lim_{V\to\infty} b_l(V) \qquad (l = 1, 2, \cdots). \quad (1.5)$$

Moreover, the common radius of convergence R of these two series satisfies[3]

$$R \geq \liminf_{V\to\infty} R(V), \quad (1.6)$$

since any point $z = a$ with $|a| < \liminf_{V\to\infty} R(V)$ must be a regular point of $p(z)$. This follows from Yang and Lee's theory.[5]

* Supported by the Air Force Office of Scientific Research under Grant 62-64.

† Present address: Imperial College, London, England.

[1] J. Groeneveld, Phys. Letters 3, 50 (1962).

[2] D. Ruelle, *Correlation Functions of Classical Gases* (Institute for Advanced Study, Princeton, 1963); Ann. Phys. (N. Y.) 25, 109 (1963).

[3] O. Penrose, J. Math. Phys. 1312 (1963).

[4] Mayer describes his theory in *Handbuch der Physik* (Springer-Verlag, Berlin, 1958), Vol. 12.

[5] C. N. Yang and T. D. Lee, Phys. Rev. 87, 404 (1952). The theory is generalized to a wider class of potentials by D. Ruelle, Helv. Phys. Acta 36, 183 (1963).

The purpose of this paper is to make a similar study of the radii of convergence $\mathfrak{R}(V)$ and $\mathfrak{R}$ of the finite- and infinite-volume density expansions obtained by eliminating z from (1.1) and (1.2), and from (1.3) and (1.4). These expansions may be written

$$p(z, V) = P(\rho(z, V), V) \equiv \kappa T \rho(z, V) \times \left[1 - \sum_k \frac{k}{k+1} \beta_k(V) \rho(z, V)^k\right], \quad (1.7)$$

$$p(z) = P(\rho(z)) \equiv \kappa T \rho(z) \times \left[1 - \sum_k \frac{k}{k+1} \beta_k \rho(z)^k\right]. \quad (1.8)$$

The $\beta_k(V)$'s can be expressed in terms of the $b\ell(V)$'s by algebraic relations,[4] such as $\beta_1(V) = 2b_2(V)$, $\beta_2(V) = 3b_3(V) - 6b_2(V)^2$, etc., which do not involve V explicitly. It follows by (1.5) that

$$\beta_k = \lim_{V\to\infty} \beta_k(V). \quad (1.9)$$

The β_k's are the irreducible Mayer cluster integrals[4] and (1.8) is the virial expansion. We shall study $\mathfrak{R}(V)$ by a method based on Lagrange's theorem for the expansion of one function of z in powers of another. This method incidentally yields upper bounds on the absolute values of the β_ℓ's and $\beta_\ell(V)$'s. We shall study $\mathfrak{R}$ by means of a generalization to the complex ρ-plane of Yang and Lee's results[5] on uniform convergence in the z plane.

Our lower bounds on $\mathfrak{R}(V)$ and $\mathfrak{R}$ apply to systems of particles whose positions $\mathbf{x}_1, \mathbf{x}_2, \cdots$ are either continuously variable or confined to a lattice. Their interaction stems from a two-body interaction potential $\varphi(r)$ for which there exists a constant Φ such that

$$\sum_{i<j\le s} \varphi(\mathbf{x}_i - \mathbf{x}_j) \ge -s\Phi \quad \text{for all } s, \mathbf{x}_1 \cdots \mathbf{x}_s. \quad (1.10)$$

The circumstances under which (1.10) is satisfied have been discussed by Ruelle[2,6] and Penrose.[3] We shall also make the convergence assumption

$$B \equiv \int_{\text{all space}} |e^{-\varphi(\mathbf{r})/\kappa T} - 1|\, d^\nu \mathbf{r} < \infty, \quad (1.11)$$

where ν is the number of space dimensions (= 1, 2, or 3). In discussing the upper bounds on $\mathfrak{R}(V)$ and $\mathfrak{R}$, we shall further assume that the potential has a hard core, i.e., that a positive constant a exists such that

$$\varphi(\mathbf{r}) = +\infty \quad \text{if} \quad r < a, \quad (1.12)$$

but this assumption is unnecessary in the other parts of the discussion.

[6] D. Ruelle, Ref. 5.

2. LAGRANGE'S THEOREM

Lagrange's theorem,[7] adapted to the expansion of $p(z, V)$ in powers of $\rho(z, V)$ may be stated thus: let the function $z/\rho(z, V)$ be analytic within and on a closed contour C surrounding the origin of the z plane, and let ρ be a complex number satisfying

$$|\rho| < \mu \equiv \operatorname*{Min}_{z \text{ on } C} |\rho(z, V)|. \quad (2.1)$$

Then the equation $\rho(z, V) = \rho$ is satisfied by just one value of z inside C, which we denote by $z(\rho, V)$; further, if the function $p(z, V)$ is analytic within and on C, it has the convergent expansion

$$P(\rho, V) \equiv p(z(\rho, V), V) = \sum_{n=1}^{\infty} c_n \rho^n, \quad (2.2)$$

where

$$c_n \equiv \frac{1}{2\pi i} \oint_{C'} \frac{dp(z, V)}{dz} \frac{dz}{n\{\rho(z, V)\}^n} = \frac{1}{n!} \frac{d^{n-1}}{dz^{n-1}} \left[\frac{dp(z, V)}{dz} \left\{\frac{z}{\rho(z, V)}\right\}^n\right]_{z=0}. \quad (2.3)$$

The path of integration is any contour C' surrounding $z = 0$ such that $|\rho(z, V)| \le \mu$ for all z on C'. The uniqueness of $z(\rho, V)$ follows from Rouché's theorem,[8] which shows that the functions z/ρ and $z/\rho - z/\rho(z, V)$ have the same number (one) of zeros inside C. The formula for c_n is obtained by expanding in powers of ρ on both sides of the following equation derived from Cauchy's residue theorem:

$$P(\rho, V) = \frac{1}{2\pi i} \oint_{C'} p(z, V) \frac{d\rho(z, V)}{dz} \frac{dz}{\rho(z, V) - \rho}, \quad (2.4)$$

and then integrating the resulting formula for c_n by parts. By virtue of the relation (1.2) between $\rho(z, V)$ and $p(z, V)$, and the definition (1.7) of the $\beta_k(V)$, Eq. (2.3) for $n = 2, 3, \cdots$ is equivalent to

$$-k\beta_k(V) = \frac{1}{2\pi i} \oint \frac{dz}{z\{\rho(z, V)\}^k} \quad (2.5)$$

for $k = 1, 2, \cdots$. This formula is used in Sec. 3 to estimate the $\beta_k(V)$'s.

3. LOWER BOUNDS ON $\mathfrak{R}(V)$ AND $\mathfrak{R}$

According to Lagrange's theorem, the series (2.2) converges if $|\rho|$ is less than the lower bound μ of $|\rho(z, V)|$ on the contour C; that is,

$$\mathfrak{R}(V) \ge \mu \equiv \operatorname*{Min}_{z \text{ on } C} |\rho(z, V)|. \quad (3.1)$$

[7] E. T. Whittaker and G. N. Watson, *Modern Analysis* (Cambridge University Press, New York, 1927), Sec. 7.32.

[8] E. T. Copson, *Theory of Functions of a Complex Variable* (Oxford University Press, London, 1935), Sec. 6.21.

A suitable lower bound on $|\rho(z, V)|$ can be found from Penrose's generalization[3] of Groeneveld's estimates[1] of the $b_\ell(V)$'s

$$|\ell b_\ell(V)| \leq u^{\ell-2}[\ell B]^{\ell-1}/\ell! \quad (\ell = 2, 3, \cdots), \tag{3.2}$$

where

$$u \equiv e^{2\Phi/\kappa T} \geq 1, \tag{3.3}$$

and Φ and B are defined in (1.10) and (1.11). These estimates imply[1,3] that $R(V) \geq 1/euB$. In the rest of the section we ensure the convergence of (1.1) and (1.2) by requiring

$$|z| < 1/euB. \tag{3.4}$$

Since $b_1 = 1$ the series (1.1) now gives the inequality

$$\begin{aligned} |\rho(z, V) - z| &= \left|\sum_{\ell=2}^{\infty} \ell b_\ell z^\ell\right| \leq \frac{1}{u^2 B} \sum_{\ell=2}^{\infty} \ell^{\ell-1}(uB\,|z|)^\ell/\ell! \\ &= w/u^2B - |z|/u, \end{aligned} \tag{3.5}$$

where w is defined by

$$we^{-w} = uB\,|z|, \qquad 0 \leq w < 1. \tag{3.6}$$

Since the function we^{-w} increases monotonically from 0 to e^{-1} in the range $0 \leq w < 1$, the condition (3.4) guarantees that w exists and is unique. In deriving the last line of (3.5) we used Euler's expansion[9] for w in powers of we^{-w}:

$$w = \sum \ell^{\ell-1}(we^{-w})^\ell/\ell!. \tag{3.7}$$

From (3.5) and (3.6) we obtain a lower bound on $|\rho(z, V)|$,

$$\begin{aligned} |\rho(z, V)| &\geq (1 + 1/u)\,|z| - w/u^2B \\ &= \{(u + 1)e^{-w} - 1\}w/u^2B. \end{aligned} \tag{3.8}$$

As the contour C in (3.1), we may choose any circle $|z| = \text{const} < 1/euB$. By (3.6) the equation of this circle may be written $w = \text{const}$, and the corresponding value of μ is $\geq \{(u+1)e^{-w}-1\}w/u^2B$. Since (3.1) holds whatever value of w in the range $0 \leq w < 1$ is used to define C, we must have

$$\mathcal{R}(V) \geq \operatorname*{Max}_{0\leq w<1} \{(u + 1)e^{-w} - 1\}w/u^2B. \tag{3.9}$$

To obtain a convenient estimate of $\mathcal{R}(V)$, we use the identity

$$\begin{aligned} \{(u + 1)e^{-w} - 1\}w/u^2B &= \left[v - v^2 g\left(\frac{vu}{1+u}\right)\right]/(u + 1)B, \end{aligned} \tag{3.10}$$

where $v \equiv w(1 + u)/u$ and $g(w) \equiv (1 - e^{-w})/w$. Since $u/(1 + u) \geq \frac{1}{2}$ by (3.3) and $g(w)$ decreases monotonically, the right side of (3.10) is at least $[v - v^2 g(\frac{1}{2}v)]/(u + 1)B$. Hence (3.9) implies

$$\begin{aligned} \mathcal{R}(V) &\geq \operatorname*{Max}_{v} [v - v^2 g(\tfrac{1}{2}v)]/(u + 1)B \\ &= 0.28952/(u + 1)B. \end{aligned} \tag{3.11}$$

The maximum is attained when $v = 0.62984$. If we had not replaced $g(vu/(1 + u))$ by $g(\frac{1}{2}v)$ the numerator in (3.11) would have been replaced by a function of u increasing monotonically from 0.28952 when $u = 1$ (nonnegative potentials) to $e^{-1} = 0.36788$ as $u \to \infty$.

These methods also yield upper bounds on the $\beta_k(V)$'s. Taking the contour in (2.5) to be a circle $|z| = \text{const}$, we obtain the estimate

$$\begin{aligned} k\,|\beta_k(V)| &\leq \frac{1}{2\pi} \oint_C d\,|z| \operatorname*{Max}_{z \text{ on } C} \left|\frac{1}{z\{\rho(z, V)\}^k}\right| \\ &= [\operatorname*{Min}_{z \text{ on } C} |\rho(z, V)|]^{-k}. \end{aligned} \tag{3.12}$$

Choosing the radius of the circle, as before, to maximize the quantity in square brackets, we find that

$$\begin{aligned} k\,|\beta_k(V)| &\leq [\operatorname*{Max}_{w} \{(u + 1)e^{-w} - 1\}w/u^2B]^{-k} \\ &\leq \left[\frac{(u + 1)B}{0.28952}\right]^k \qquad (k = 1, 2, \cdots). \end{aligned} \tag{3.13}$$

Combined with (1.5) this gives upper bounds on the irreducible Mayer cluster integrals

$$k\,|\beta_k| \leq [(u + 1)B/(0.28952)]^k. \tag{3.14}$$

This set of inequalities implies, by (1.8) and Cauchy's kth-root convergence test, that

$$\begin{aligned} \mathcal{R} = \liminf_{k\to\infty} \left|\frac{k}{k+1}\beta_k\right|^{-1/k} \\ \geq 0.28952/(u + 1)B, \end{aligned} \tag{3.15}$$

so that $\mathcal{R}$ and $\mathcal{R}(V)$ have the same lower bound. The result (3.15) can also be obtained by applying to the function $\rho(z)$ the same arguments which when applied to $\rho(z, V)$ led to (3.11); or by using (6.1).

4. UPPER BOUND ON $\mathcal{R}(V)$

One way of finding an upper bound on $\mathcal{R}(V)$ is to locate singularities of the analytic continuation of the function $P(\rho, V)$ defined for small ρ in Sec. 2. This analytic continuation is easiest for the physically possible values of ρ.

The physically possible values of z are the real

[9] G. Pólya and G. Szegö, *Aufgaben und Lehrsätze der Analysis* (Springer-Verlag, Berlin, 1925), Vol. I, Part III, Chap. 5, No. 209.

positive values. The theory of fluctuations shows that $d\rho(z, V)/dz = [\langle N^2\rangle - \langle N\rangle^2]/z\mathcal{V}$ is positive for positive z. Therefore, as z increases from 0 to ∞, $\rho(z, V)$ increases monotonically from 0 to some limiting value $\rho_M(V)$, (which may be $+\infty$) and $p(z, V) = \kappa T \int_0^z \rho(z, V)\, dz/z$ increases monotonically from 0 to ∞. Thus the physically possible values of ρ are $0 < \rho < \rho_M(V)$. For hard-core potentials $\rho_M(V)$ is given by

$$\rho_M(V) = M(V)/\mathcal{V}, \tag{4.1}$$

where $M(V)$ is the largest number of nonintersecting spheres of diameter a whose centers can be fitted into the region V. For potentials without hard cores, $\rho_M(V)$ is $+\infty$.

Since $\rho(z, V)$ increases monotonically for $0<z<\infty$, its inverse function $z(\rho, V)$—although many-valued—has a branch $Z(\rho, V)$, which increases monotonically from 0 to ∞ as ρ increases from 0 to $\rho_M(V)$. For the physically possible values of ρ we may therefore define $P(\rho, V)$ by

$$P(\rho, V) \equiv p(Z(\rho, V), V) \quad (0 < \rho < \rho_M). \tag{4.2}$$

This function increases monotonically from 0 to ∞, and is therefore singular at $\rho = \rho_M(V)$. Moreover, the two definitions (2.2) and (4.2) are equivalent when $0 < \rho < 0.28952/(u + 1)B$. It follows that the series (2.2) must diverge when $\rho = \rho_M(V)$, so that

$$\mathfrak{R}(V) \leq \rho_M(V). \tag{4.3}$$

Unfortunately this upper bound provides information only for hard-core potentials.

Taking the limit $V \to \infty$ we obtain[10]

$$\lim_{V\to\infty} \mathfrak{R}(V) \leq \rho_M \equiv \lim_{V\to\infty} \rho_M(V). \tag{4.4}$$

To obtain an upper bound on $\mathfrak{R}$ we may try using the same argument for $\rho(z)$ and $z(\rho)$ instead of $\rho(z, V)$ and $z(\rho, V)$. Provided that the system has no phase transition [so that $\rho(z)$ is analytic at every point on the positive z axis], and provided that

$$d\rho(z)/dz > 0 \quad \text{for all} \quad z > 0, \tag{4.5}$$

the same argument goes through, giving

$$\mathfrak{R} \leq \rho_M \equiv \lim_{V\to\infty} \rho_M(V) \tag{4.6}$$

if there is no phase transition. However, if there is a phase transition, $\mathfrak{R}$ may perhaps be larger than ρ_M. The approximate equation of state found by Reiss, Frisch, and Lebowitz[11] for the hard-sphere fluid illustrates this possibility since its only singularity is a triple pole at $\rho = 6\rho_M/\pi\sqrt{2} = 1.3505\rho_M$ which suggests that $\mathfrak{R} \cong 1.3505\rho_M > \rho_M$.

[10] If the limits in (4.4) do not exist, the inequality is true for both the largest and the smallest limit points of $\mathfrak{R}(V)$ and $\rho_M(V)$.

5. YANG–LEE THEORY FOR THE ρ PLANE

In this section we generalize Yang and Lee's theory of uniform convergence in the z plane by proving a corresponding theorem for the ρ plane. This theorem indicates, for example, the circumstances under which the operations $\lim_{V\to\infty}$ and $d/d\rho$ are interchangeable.

Theorem. Let $\rho_1 < \rho < \rho_2$ be a segment of the real ρ axis, with $0 \leq \rho_1$ and $\rho_2 < \rho_M$. Let $\mathfrak{D}$ be any bounded simply connected region in the ρ plane, whose intersection with the line segment $0 < \rho < \rho_M$ is the set $\rho_1 < \rho < \rho_2$, and into which analytic continuation of the functions $P(\rho, V)$ defined by (4.2) yields a single-valued regular function for all sufficiently large V. Then the sequence of functions $P(\rho, V)$ converges uniformly on any region bounded by a contour inside $\mathfrak{D}$.

Proof: The proof depends on Vitali's theorem[12] which states that, if $\mathfrak{D}$ is a region and $f(\rho, V)$ is a sequence of analytic functions which are

(i) regular in $\mathfrak{D}$,
(ii) uniformly bounded in $\mathfrak{D}$,
(iii) convergent, as $V \to \infty$, on a set of points having a limit point in $\mathfrak{D}$,

then the sequence $f(\rho, V)$ converges uniformly in any region bounded by a contour inside $\mathfrak{D}$. We shall apply Vitali's theorem to the sequence

$$f(\rho, V) \equiv \rho/Z(\rho, V), \tag{5.1}$$

where $Z(\rho, V)$ is the analytic continuation, into $\mathfrak{D}$, of the function $Z(\rho, V)$ defined for $0 < \rho < \rho_M$ in Sec. 4. We start the sequence (5.1) with V sufficiently large to make the analytic continuation of $P(\rho, V)$ into $\mathfrak{D}$ possible for all larger V. According to the definitions (1.1) and (1.7), the functions $f(\rho, V)$ and $P(\rho, V)$ are related by the differential equation

$$\frac{1}{\kappa T}\frac{dP(\rho, V)}{d\rho} = 1 - \rho\frac{d[\log f(\rho, V)]}{d\rho}. \tag{5.2}$$

[11] H. Reiss, H. L. Frisch and J. L. Lebowitz [J. Chem. Phys. 31, 369 (1959)] find that $\rho/\rho\kappa T \simeq (1 + \alpha + \alpha^2)/(1 - \alpha)^3$ where $\alpha = \sqrt{2}\pi\rho/6\rho_M$. The same equation of state also follows from the Percus–Yevick equation: see M. Wertheim, Phys. Rev. Letters 8, 321 (1963); E. Thiele, J. Chem. Phys. 39, 474 (1963).

[12] E. C. Titchmarsh, *The Theory of Functions* (Oxford University Press, London, 1939), 2nd ed., p. 168.

Since $P(\rho, V)$ is regular and $\mathfrak{D}$ is simply connected, it follows that $\log f(\rho, V)$ is regular and single-valued in $\mathfrak{D}$; therefore, the analytic continuation used in the definition (5.1) leads to no ambiguities, and moreover $f(\rho, V)$ satisfies the condition (i) of Vilati's theorem.

To deal with the condition (ii), consider first the part of $\mathfrak{D}$ where $Z(\rho, V) \geq 1/euB$. Clearly $f(\rho, V)$ is bounded in this part, since the denominator of (5.1) is bounded away from zero and the numerator is bounded because $\mathfrak{D}$ is a bounded region of the ρ plane. For the other part of $\mathfrak{D}$, where $Z(\rho, V) < 1/euB$, we write z for $Z(\rho, V)$ and use (3.5) to show that

$$|\rho(z, V)| \leq w/u^2B + (1 - 1/u)\,|z| \leq w/uB, \tag{5.3}$$

so that

$$|f(\rho, V)| = \left|\frac{\rho(z, V)}{z}\right| \leq w/|z|\,uB = e^w \leq e. \tag{5.4}$$

Thus $f(\rho, V)$ is bounded in both parts of $\mathfrak{D}$, and (ii) is satisfied.

To show that the sequence defined in (5.1) satisfies condition (iii) it is sufficient to show that, as $V \to \infty$, $Z(\rho, V)$ converges to a limit at almost all points on the segment $\rho_1 < \rho < \rho_2$, since then any subsegment $\rho_1' \leq \rho < \rho_2'$, where $\rho_1 < \rho_1' < \rho_2' < \rho_2$, lies within $\mathfrak{D}$ and contains[13] at least one limit point of the points where $Z(\rho, V)$ converges.

We shall begin by proving the corresponding convergence property for the function $\rho(z, V)$ of which $Z(\rho, V)$ is the inverse. The proof depends on the fact, proved in Sec. 4, that $\rho(z, V)$ is an increasing function of z for real positive z; this fact implies that, for any positive z,

$$\rho(z, -h, V) \leq \rho(z, V) \leq \rho(z, h, V), \tag{5.5}$$

where h is a positive number less than z, and

$$\rho(z, \pm h, V) \equiv h^{-1} \int_0^h \rho(z \pm t, V)\, dt/t \tag{5.6}$$

$$= [p(z \pm h, V) - p(z, V)]/(\pm h\kappa T) \tag{5.7}$$

by (1.2). It is known from Yang and Lee's theory[5] that

$$p(z) \equiv \lim_{V\to\infty} p(z, V) \tag{5.8}$$

exists for all positive z; therefore taking the limit $V \to \infty$ in (5.5) gives

$$\rho(z, -h) \leq \liminf_{V\to\infty} \rho(z, V) \leq \limsup_{V\to\infty} (\rho(z, V) \leq \rho(z, +h)), \tag{5.9}$$

where

$$\rho(z, \pm h) \equiv [p(z \pm h) - p(z)]/(\pm h\kappa T). \tag{5.10}$$

Taking the limit $h \to 0$ in (5.9) we find that $\lim_{V\to\infty} \rho(z, V)$ exists, and is equal to $(z/\kappa T)\, dp(z)/dz$, for all positive values of z where $dp(z)/dz$ exists. But $p(z)$, being a nondecreasing function, is[14] differentiable for almost all z; therefore

$$\rho(z) \equiv \lim_{V\to\infty} \rho(z, V) \tag{5.11}$$

exists for almost all positive values of z.

Since the $\rho(z, V)$'s are increasing functions, the limit function $\rho(z)$ is nondecreasing. Its inverse function $z(\rho)$ is therefore uniquely defined[15] for all values of ρ satisfying $0 < \rho < \rho_M$, apart from a set of exceptional values of ρ for which the equation $\rho = \rho(z)$ has more than one solution. Each exceptional value corresponds to a segment of the real z axis on which $\rho(z)$ is constant. Since these segments of the z axis are countable, the exceptional values of ρ form a set of zero measure.

To show that $\lim Z(\rho, V)$ exists, let ρ_0 be any nonexceptional value of ρ, let $z_0 \equiv Z(\rho_0)$, and let ϵ be a small positive number such that $\rho(z_0 - \epsilon)$ and $\rho(z_0 + \epsilon)$ exist. Since $\rho(z)$ is monotonic and ρ_0 is nonexceptional, we have $\rho(z_0 - \epsilon) < \rho_0 < \rho(z_0 + \epsilon)$, and hence by (5.11) the inequality

$$\rho(z_0 - \epsilon, V) < \rho_0 < \rho(z_0 + \epsilon, V) \tag{5.12}$$

holds for all sufficiently large V. Applying the nondecreasing function $Z(\rho, V)$ to (5.12) we find

$$z_0 - \epsilon \leq Z(\rho_0, V) \leq z_0 + \epsilon. \tag{5.13}$$

Since ϵ can be made arbitrarily small, it follows that

$$\lim_{V\to\infty} Z(\rho_0, V) = z_0 = Z(\rho_0) \tag{5.14}$$

for almost all values of ρ_0 in the range $0 < \rho < \rho_M$. Consequently, condition (iii) of Vitali's theorem is satisfied.

Vitali's theorem now tells us that the sequence $f(\rho, V)$ converges uniformly in any region bounded by a contour inside $\mathfrak{D}$; its limiting function $f(\rho)$ is therefore regular inside $\mathfrak{D}$. To prove our theorem that the same is true of the sequence $P(\rho, V)$ we consider two cases separately. Suppose first that $f(\rho)$ has a zero inside $\mathfrak{D}$, say at $\rho = \alpha$. The value of α cannot be zero, since if the point $\rho = 0$ is within $\mathfrak{D}$ then the conditions of the theorem imply

[13] Ref. 7, Sec. 2.21, p. 12.

[14] Ref. 12, Sec. 11.42.

[15] Either as the solution of $\rho = \rho(z)$ or, if this has no solution, by means of a Dedekind section of the real z axis.

that the segment of the nonnegative real axis inside $\mathfrak{D}$ is $0 \leq \rho < \rho_2$, so that by continuity $f(0) = \lim_{\rho \to 0+} \rho/Z(\rho, V) = \lim_{z \to 0} \rho(z, V)/z = 1 \neq 0$. By Hurwitz's theorem,[16] all the $f(\rho, V)$'s for large enough V must also have zeros at points near $\rho = \alpha \neq 0$ inside $\mathfrak{D}$. Hence, by (5.2), all the $P(\rho, V)$'s for large enough V have logarithmic singularities at these points. This contradicts the condition that the $P(\rho, V)$'s must be single-valued within $\mathfrak{D}$ for large enough V, and thus rules out this first case where $f(\rho)$ has a zero inside $\mathfrak{D}$.

In the remaining case the function $f(\rho)$, having no zeros inside $\mathfrak{D}$, must be bounded away from zero inside any contour within $\mathfrak{D}$; consequently all the $f(\rho, V)$'s are also bounded away from zero inside the contour for large enough V. It follows that the sequence $\log f(\rho, V)$ converges uniformly within the contour, and so also do[17] the sequences $d(\log f(\rho, V))/d\rho$ and [by (5.2)] $dP(\rho, V)/d\rho$. Evaluating $P(\rho, V)$ by integration of its derivative along a path inside $\mathfrak{D}$ with one end fixed on the positive real axis, we conclude[17] that the sequence $P(\rho,V)$ does converge uniformly within the contour. Q.E.D.

6. RELATION BETWEEN $\mathfrak{R}$ AND $\lim \mathfrak{R}(V)$

The theorem of Sec. 5 leads at once to a result analogous to (1.6). Let δ be any small positive number. Then the disk $|\rho| < \liminf_{V \to \infty} \mathfrak{R}(V) - \delta$ satisfies the conditions required of the region $\mathfrak{D}$, since the power series (2.2) whose radius of convergence exceeds the radius of $\mathfrak{D}$ for all sufficiently large V, provides the analytic continuation of $P(\rho, V)$ from the real axis into $\mathfrak{D}$. The theorem then implies that $P(\rho)$, being the limit of a uniformly convergent sequence of analytic functions, is itself analytic inside the contour $|\rho| = \liminf_{V \to \infty} \mathfrak{R}(V) - 2\delta$. Therefore the power-series expansion (1.8) for $P(\rho)$ converges if $|\rho| \leq \liminf_{V \to \infty} \mathfrak{R}(V) - 2\delta$. Since δ can be made arbitrarily small, it follows that

$$\liminf_{V \to \infty} \mathfrak{R}(V) \leq \mathfrak{R}. \tag{6.1}$$

7. OTHER DENSITY EXPANSIONS

Besides the pressure, other quantities have useful expansions in powers of ρ. We can relate their radii of convergence to $\mathfrak{R}$ and $\mathfrak{R}(V)$.

Foremost among these expansions is that of the fugacity z. Since the analytic functions $P(\rho, V)$ and $Z(\rho, V)$ are both regular near $\rho = 0$, it follows from the differential equation (5.2) that their singularities in the ρ plane (appropriately cut) coincide, and hence that the series expansion of $Z(\rho, V)$ has radius of convergence $\mathfrak{R}(V)$. Similarly, the series expansion of $Z(\rho) \equiv \lim_{V \to \infty} Z(\rho, V)$ has radius of convergence $\mathfrak{R}$.

The density expansions for the s-particle distribution functions $n_s(\mathbf{x}_1, \cdots \mathbf{x}_s)$ are also important. To study their convergence, consider the disk $|\rho| < \mathfrak{R}(V)$ and its image D in the z plane under the mapping $z = Z(\rho, V)$. Since the function $\rho(z, V)$ is single-valued, it is regular within D; therefore[5] $\Xi(z, V)$ has no zeros in D, so that[3] $n_s(\mathbf{x}_1, \cdots \mathbf{x}_s)$ is a regular function of z within D. It follows that $n_s(\mathbf{x}_1, \cdots \mathbf{x}_s)$ is a regular function of ρ within $|\rho| < \mathfrak{R}(V)$, so that its expansion in powers of ρ has radius of convergence at least $\mathfrak{R}(V)$.

8. DISCUSSION

The information we have obtained about $\mathfrak{R}(V)$ and $\mathfrak{R}$ can be summarized in the formulas

$$0.28952/(u + 1)B \leq \mathfrak{R}(V) \leq \rho_M(V), \tag{8.1}$$

$$\liminf_{V \to \infty} \mathfrak{R}(V) \leq \mathfrak{R}, \tag{8.2}$$

which come from (3.11), (4.3), and (6.1). The quantities u, B, and $\rho_M(V)$ are defined in (3.3), (1.11), and (4.1).

The simplest illustration of these formulas is provided by a system of hard rods in one dimension. Its equation of state is

$$P/kT = \rho/(1 - a\rho) = \rho + a\rho^2 + \cdots, \tag{8.3}$$

where a is the length of each rod. The value of $\mathfrak{R}$ is therefore $1/a$. The value of $\liminf_{V \to \infty} \mathfrak{R}(V)$ is harder to calculate, but (8.1) and (8.2) provide the rather wide bounds

$$0.07238 \leq a \liminf_{V \to \infty} \mathfrak{R}(V) \leq 1 \tag{8.4}$$

since $u = 1$, $B = 2a$, and $\rho_M = 1/a$.

The main physical conclusion to be drawn from our results is that there can be no phase transition for densities less[18] than $0.28952/(u + 1)B$, since the series (1.7) converges for these densities and is equal (by the theorem of Sec. 5) to the thermodynamic pressure, which is therefore an analytic

[16] Ref. 12, Sec. 3.45.
[17] Ref. 8, Secs. 5.13 and 5.12.

[18] D. Ruelle, Ref. 2, shows that for nonnegative potentials there can be no phase transition for densities less than $1/3.8B = 0.26/B$. Using an inequality due to E. Lieb [J. Math. Phys. 4, 671 (1963)], this number can be slightly increased to $1/(1 + e)B = 0.27/B$. For general hard-core potentials the corresponding number is $1/u[(1 + e)B_+ + eB_-]$, where B_+ and B_- are the contributions of the positive and negative parts of $\varphi(r)$ to the integral (1.1) [see O. Penrose, J. Math. Phys. 4, 1488 (1963), Eq. (8.3)]. For more general potentials, however, the bound $0.28952/(u + 1)B$ given in the text is the best available.

function of ρ. Moreover, if lim inf $\mathfrak{R}(V)$ is known, it provides a better lower bound on the density at a phase transition. This follows from the arguments of Secs. 5 and 6.

On the other hand, our results do not prove that $\mathfrak{R}$ is a lower bound on the density at a phase transition. For quantum systems, Fuchs[19] has shown that $\mathfrak{R}$ can actually exceed the value of ρ at a phase transition (for an ideal B–E gas). For classical systems the question remains open, although the example mentioned at the end of Sec. 4 suggests that here too, $\mathfrak{R}$ can exceed the value of ρ at a phase transition.[20]

Although the value of ρ at the first phase transition cannot be less than lim inf $\mathfrak{R}(V)$, it can be greater than both lim inf $\mathfrak{R}(V)$ and $\mathfrak{R}$. This can be shown by considering a one-dimensional system of interacting hard rods with the interaction potential

$$\varphi(r) = \begin{cases} +\infty & (|r| < a), \\ \kappa T \ln 2 & (a \leq |r| < 2a), \\ 0 & (2a \leq |r|). \end{cases} \tag{8.5}$$

For this potential, fugacity and pressure are related by[21]

$$\frac{1}{Z} = \int_0^\infty e^{-[rp+\varphi(r)]/\kappa T}\, dr = \frac{\kappa T}{p} e^{-3ap/2\kappa T} \cosh(ap/2\kappa T), \tag{8.6}$$

so that

$$\frac{1}{\rho \kappa T} = \frac{d(\ln z)}{dp} = \frac{1}{p} + \frac{3a}{2\kappa T} - \frac{a}{2\kappa T} \tanh \frac{ap}{2\kappa T}. \tag{8.7}$$

As p moves in its Argand plane from the origin along the positive imaginary axis, the value of $1/\rho - \frac{3}{2}a$ moves along its imaginary axis from $-i\infty$ to a value $-(1.1322)ia$, achieved when $p = 2ix\kappa T/a$ where $x = 0.7393$ is the real solution of $x = \cos x$, and then retreats again to $-i\infty$. Hence the image point of ρ starts at the origin of its Argand plane, moves along a circular arc whose furthest point from the origin is its other end at $1/(1.5000 - 1.1322i)a$, and returns to the origin. Therefore the function $P(\rho)$ has a branch point at $1/(1.5000 - 1.1322i)a$, and $\mathfrak{R}$ is at most $1/|1.5000 - 1.1322i|\, a = \frac{1}{2}(0.5321)$, which is less than $\rho_M = 1/a$. Thus, for this system, unlike the simple hard-rod system, both $\mathfrak{R}$ and lim inf $\mathfrak{R}(V)$ are less than ρ_M, despite the fact that there is no phase transition for $0 < \rho < \rho_M$. Therefore, the actual values of $\mathfrak{R}$ and lim $\inf_{V\to\infty}$ $\mathfrak{R}(V)$ have, in general, no physical significance since they may be determined by singularities off the real positive ρ axis.[22]

ACKNOWLEDGMENTS

We are indebted to B. Epstein and H. E. Rauch for useful advice.

[19] W. H. J. Fuchs, J. Ratl. Mech. Anal. 4, 647 (1955).

[20] M. Kac, G. E. Uhlenbeck and P. C. Hemmer, J. Math. Phys. 4, 216 (1963), consider a one-dimensional system with

$$\varphi(r) = \begin{cases} \infty & |r| < a \\ -2\alpha\gamma e^{-\gamma r}, & |r| > a, \end{cases}$$

and find that this system has a phase transition in the limit $\gamma \to 0$, which can be obtained from Maxwell's equal area construction applied to

$$p_0(\rho) = \kappa T\left[\frac{\rho}{1-\rho a} - \alpha\rho^2\right] = \kappa T\rho\left[1 - \sum_k \frac{k}{k+1}\beta_k{}^0\rho^k\right],$$

where $\beta_k{}^0 = \lim_{\gamma\to 0}\beta_k(\gamma)$, and $\beta_k(\gamma) = \lim_{V\to\infty}\beta_k(\gamma, V)$. Thus $\mathfrak{R}^0 = a^{-1}$, the radius of convergence of the above series exceeds the value of ρ at the phase transition.

[21] H. Takahasi, Proc. Phys. Soc. Japan 24, 60 (1942); F. Gursey, Proc. Cambridge Phil. Soc. 46, 182 (1950).

[22] The nearest singularity of $P(\rho)$ is off the real positive ρ axis if and only if an infinite number of virial coefficients are negative. This example therefore supplements Wertheim's proof that the virial coefficients need not all be positive even if $\varphi(r)$ is nonnegative. [Wertheim considers the case $\varphi(r) \propto r^{-n}$; forthcoming paper.]

An Algebraic Approach to Quantum Field Theory

RUDOLF HAAG AND DANIEL KASTLER*
Department of Physics, University of Illinois, Urbana, Illinois
(Received 24 December 1963)

It is shown that two quantum theories dealing, respectively, in the Hilbert spaces of state vectors Φ_1 and Φ_2 are physically equivalent whenever we have a faithful representation of the same abstract algebra of observables in both spaces, no matter whether the representations are unitarily equivalent or not. This allows a purely algebraic formulation of the theory. The framework of an algebraic version of quantum field theory is discussed and compared to the customary operator approach. It is pointed out that one reason (and possibly the only one) for the existence of unitarily inequivalent faithful, irreducible representations in quantum field theory is the (physically irrelevant) behavior of the states with respect to observations made infinitely far away. The separation between such "global" features and the local ones is studied. An application of this point of view to superselection rules shows that, for example, in electrodynamics the Hilbert space of states with charge zero carries already all the relevant physical information.

I. INTRODUCTION

THE essential feature which distinguishes quantum field theory within the frame of general quantum physics is the principle of locality. This principle states that it is meaningful to talk of observables which can be measured in a specific space-time region and that observables in causally disjoint regions are always compatible. It is then natural to introduce the following concepts: If $\mathbf{B}$ is a region in Minkowski space, we denote by $\mathfrak{A}(\mathbf{B})$ the algebra generated by the observables in $\mathbf{B}$. A specific field theory will fix the correspondence between regions and algebras

$$\mathbf{B} \rightarrow \mathfrak{A}(\mathbf{B}). \tag{1}$$

In fact, we may consider this correspondence to be the content of the theory. Indeed, once it is known, one can calculate quantities of direct physical interest such as masses of particles and collision cross sections.[1]

This approach has been developed in previous work[2,3] within the customary framework of quantum theory in which observables are considered to be (bounded or unbounded) operators on a Hilbert space. The algebras $\mathfrak{A}(\mathbf{B})$ are then concrete *-algebras of operators and it is mathematically convenient to replace $\mathfrak{A}(\mathbf{B})$ by its associated von Neumann ring $\mathbf{R}(\mathbf{B})$. Properties of this family of von Neumann rings, which follow from general physical principles or are suggested by conventional quantum field theory, have been studied in Ref. 2. Particle aspects and collision theory are treated in Ref. 3.

In the present paper we shall be concerned with another question. Suppose that the algebras $\mathfrak{A}(\mathbf{B})$ are abstractly defined (without reference to operators on a Hilbert space).[4] If we consider a faithful realization of the algebraic elements by operators on a Hilbert space we come back to the previous point of view. However, we expect that there are many unitarily inequivalent irreducible representations. This ambiguity, typical of quantum field theory, has been the subject of some discussion within the past decade.[5] To deal with it, most authors assume that there is one and only one representation space in which the physical vacuum state appears as a vector and that we have to single out this particular representation as the physically

* This work was supported in part by the National Science Foundation.
† On leave from Physique Théorique, Faculté des Sciences, Marseille, France.
[1] At the present stage this claim is an overstatement, but it is a reasonable extrapolation of results described in Ref. 3.
[2] R. Haag, *Colloque Internationale sur les Problèmes Mathématiques de la Théorie Quantique des Champs, Lille, 1957* (Centre National de la Recherche Scientifique, Paris, 1958); R. Haag and B. Schroer, J. Math. Phys. **3**, 248 (1962); H. Araki, "Einführung in die Axiomatische Quantenfeldtheorie," Lecture notes at the Eidgenössischen Technischen Hochschule, Zürich, 1961/62, unpublished.
[3] R. Haag, Phys. Rev. **112**, 669 (1958); D. Ruelle, Helv. Phys. Acta **35**, 147 (1962); H. Araki, see Ref. 2.
[4] In a heuristic manner the commutation relations and field equations of a conventional quantum field theory provide such an abstract characterization.
[5] It was first noticed in the example of various algebras associated with infinitely many creation and destruction operators. See J. von Neumann, Comp. Math. **6**, 1 (1938); K. O. Friedrichs, *Mathematical Aspects of the Quantum Theory of Fields* (Interscience Publishers, Inc., New York, 1953). For further discussions of this phenomenon in its relation to various models in quantum field theory see, for instance, L. Van Hove, Physica **18**, 145 (1952); A. S. Wightman and S. S. Schweber, Phys. Rev. **98**, 812 (1955); R. Haag, Kgl. Danske Videnskab. Selskab Mat.-Fiz. Medd. **29**, No. 12 (1955); I. E. Segal, Trans. Am. Math. Soc. **88**, 12 (1958); J. Lew, Ph.D. thesis, Princeton Univ., 1960, unpublished; and the papers cited in Ref. 6.

relevant one.[6] While this attitude appears to be perfectly consistent it does not go to the heart of the matter. The fact that no actual measurement can be performed with absolute precision implies that the realistic notion of "physical equivalence" is far less stringent than that of unitary equivalence. Our discussion in Sec. II shows that this notion coincides with the mathematical concept of "weak equivalence" as introduced by Fell.[7]

Fell's results then imply that all faithful representations are in fact physically equivalent, thus opening the way to a purely algebraic approach to the theory. The distinction between "physical" and "unitary" equivalence was forced on us by the discussion of examples in a recent paper.[8]

The purely algebraic approach to the theory has been championed for many years by Segal.[9] He pointed out that many questions of physical interest (e.g., the determination of spectral values) can be answered without reference to a Hilbert space if one chooses the algebra of observables to be a C^*-algebra.[10] Applying these ideas to quantum field theory Segal expected to circumvent the difficulties associated with the existence of inequivalent representations.[11] So far this approach has stayed, however, in a somewhat experimental stage, i.e., it has not yet led to a well-defined frame in which a satisfactory physical interpretation is specified. It is the purpose of this paper to establish such a frame making essential use of the principle of locality. This frame is very similar to that in Ref. 2 but differs from it in two respects. First, we consider the algebras $\mathfrak{A}(\mathbf{B})$ as (abstract) C^*-algebras, not as operator algebras on a Hilbert space. Secondly, we exclude from the list of "all" observables those quantities which refer to infinitely extended regions. Thus the total energy, total charge, etc., are considered as unobservable. This is of particular importance in connection with superselection rules (see Sec. III).

We turn now to a precise specification of the frame:

(1) The "regions" $\mathbf{B}$ for which the correspondence (1) is defined shall be the open sets with compact closure[12] in Minkowski space, the algebras $\mathfrak{A}(\mathbf{B})$ shall be (abstract) C^*-algebras.

(2) Isotony: If $\mathbf{B}_1 \supset \mathbf{B}_2$ then $\mathfrak{A}(\mathbf{B}_1) \supset \mathfrak{A}(\mathbf{B}_2)$. We assume in addition that one of the two following situations prevails. Either $\mathfrak{A}(\mathbf{B}_1)$ and $\mathfrak{A}(\mathbf{B}_2)$ have a common unit element, or neither of them has a unit. The first situation can be obtained from the second by formal adjunction of a unit.

(3) Local Commutativity: If $\mathbf{B}_1$ and $\mathbf{B}_2$ are completely spacelike with respect to each other, then $\mathfrak{A}(\mathbf{B}_1)$ and $\mathfrak{A}(\mathbf{B}_2)$ commute.

(4) The set-theoretic union of all $\mathfrak{A}(\mathbf{B})$ is a normed *-algebra.[13] Taking its completion we get a C^*-algebra which we denote by $\mathfrak{A}$ and call the *algebra of quasilocal observables*. We maintain that $\mathfrak{A}$ contains all observables of interest.[14]

(5) Lorentz Covariance: The inhomogeneous Lorentz group is represented by automorphisms $A \in \mathfrak{A} \to A^L \in \mathfrak{A}$ such that

$$\mathfrak{A}(\mathbf{B})^L = \mathfrak{A}(L\mathbf{B}), \tag{2}$$

where $L\mathbf{B}$ is the image of the region $\mathbf{B}$ under the Lorentz transformation L.

(6) $\mathfrak{A}$ is primitive (see Appendix).

Concerning the physical interpretation the essential point is, of course, that the algebra of observables $\mathfrak{A}$ has a texture, namely the family of subalgebras $\mathfrak{A}(\mathbf{B})$, and that the elements of $\mathfrak{A}(\mathbf{B})$ are interpreted as representing physical operations performed in the region $\mathbf{B}$. In Sec. II we discuss to some extent how this information can be exploited and we justify the previously mentioned notion of physical equivalence of representations. Section III deals with the separation of global and local aspects and its application to superselection rules. Section IV gives a brief comparison between the present approach and the operator approach.

II. PHYSICAL INTERPRETATION OF AN ALGEBRAIC SCHEME AND PHYSICAL EQUIVALENCE OF REPRESENTATIONS

We are concerned with two categories of objects: "states" and "operations." The term "state" is

[6] In Wightman's approach the existence of a vacuum state and the relevant properties of this state are postulated on physical grounds. See, e.g., A. S. Wightman, Phys. Rev. **101**, 860 (1956). The following papers discuss the existence and uniqueness of a vacuum state for specific models. H. Araki, J. Math. Phys. **1**, 492 (1960); D. Shale, Ph.D. Thesis, Department of Mathematics, University of Chicago, 1961, unpublished; I. E. Segal, Illinois J. Math. **6**, 500 (1962); H. J. Borchers, R. Haag, and B. Schroer, Nuovo Cimento **29**, 148 (1963).

[7] J. M. G. Fell, Trans. Am. Math. Soc. **94**, 365 (1960).

[8] H. J. Borchers, R. Haag, and B. Schroer, see Ref. 6.

[9] I. E. Segal, Ann. Math. **48**, 930 (1947).

[10] For definitions and relevant theorems see Appendix 1.

[11] I. E. Segal, *Colloque Internationale sur les Problèmes de la Théorie Quantique des Champs, Lille, 1957* (Centre National de la Recherche Scientifique, Paris, 1958).

[12] Physically speaking: 4-dimensional regions with finite extension.

[13] The union of all $\mathfrak{A}(\mathbf{B})$ has an obvious *-algebra structure due to the isotony assumption. Furthermore, the norm of one of its elements is the same in all local algebras, $\mathfrak{A}(\mathbf{B})$ containing it due to the uniqueness of the C^*-norm (see Appendix 1).

[14] $\mathfrak{A}$ is the collection of the uniform limits of all (bounded) observables describing measurements performable in finite regions of space-time. By taking uniform limits we do not essentially change the local character of the observables (hence the name quasilocal).

used for a statistical ensemble of physical systems,[15] the term "operation" for a physical apparatus which may act on the systems of an ensemble during a limited amount of time producing a transformation from an initial state to a final state. We assume that any operation is applicable to any state. This is one of the idealizations inherent in quantum physics. One is frequently interested in operations which transmit only a certain fraction of the systems of the initial ensemble and eliminate the others. This fraction (probability) is a number depending on the initial state and on the operation. It is the one piece of information about the state which is gathered by the experimenter performing the operation.[16] An experiment may always be regarded as the determination of the transmission probabilities for a finite number of operations.

We may say therefore that we have a complete theory if we are able in principle to compute such probabilities for every state and every operation when the state and the operation are defined in terms of laboratory procedures.

It is not the objective of this paper to justify the particular mathematical formalism by means of which "states" and "operations" are represented in quantum theory.[17] We accept here uncritically the following formal structure:[18] One has an algebra $\mathfrak{A}$ (which in our case will be identified with the C^*-algebra described in the introduction). A "state" is mathematically represented by a positive linear form (expectation functional) over $\mathfrak{A}$. Explicitly, if φ denotes a state, then for every $A \in \mathfrak{A}$ we have a complex number $\varphi(A)$, depending linearly on A and such that

$$\varphi(A^*A) \geq 0. \tag{3}$$

The value which φ takes for the unit element I of the algebra defines the normalization of the state. Intuitively speaking, $\varphi(I)$ is proportional to the number of systems of which the ensemble is composed (the proportionality factor being irrelevant). One calls $\varphi(I)$ the norm of the state φ.[19] The collection of all positive linear functionals over $\mathfrak{A}$ is the positive cone of the dual space $\mathfrak{A}^*$ of the algebra and is therefore denoted by $\mathfrak{A}^{*(+)}$. A functional $\varphi \in \mathfrak{A}^{*(+)}$ is called "extremal" if it cannot be decomposed into a positive linear combination of two others, i.e., if $\varphi = \alpha\varphi_1 + \beta\varphi_2$ with $\alpha > 0$, $\beta > 0$, $\varphi_1 \in \mathfrak{A}^{*(+)}$, $\varphi_2 \in \mathfrak{A}^{*(+)}$ is impossible except for the trivial solution $\varphi_1 = \lambda\varphi$. The extremal functionals correspond to pure states.

An "operation" is mathematically represented by a linear transformation of $\mathfrak{A}^*$ which maps $\mathfrak{A}^{*(+)}$ into itself and does not increase the norm. Those special operations which transform pure states into pure states are called "pure operations." It is asserted that the pure operations are in one to one correspondence with the elements contained in the unit sphere of the algebra (elements $A \in \mathfrak{A}$ with $||A|| \leq 1$).[20] The transformation of the (general) state φ by the pure operation A is given by $\varphi \to \varphi_A$ with

$$\varphi_A(B) = \varphi(A^*BA). \tag{4}$$

Therefore one gets for the transmission probability of φ through A the expression

$$P(\varphi, A) = \varphi(A^*A)/\varphi(I). \tag{5}$$

Apart from the emphasis on "operations" instead of "observables" the preceding paragraph was just a description of the standard formal structure of quantum physics. It may be useful to point out the difference between the Hilbert space approach[21] and the purely algebraic approach[22] as far as this general formalism is concerned. In the former case the only states considered are the density matrices in the representation space (positive-definite self-adjoint operators with finite trace). This collection of states is a subset (usually not the whole) of $\mathfrak{A}^{*(+)}$. The purely algebraic approach on the other hand considers all elements of $\mathfrak{A}^{*(+)}$ as possible states. One has to ask therefore whether this richer supply of states makes the physical interpretation more difficult.

We must turn now to the physical interpretation, i.e., to the following question: Suppose a specific

[15] We adopt Segal's terminology in which the word state is used for any statistical ensemble. If the ensemble cannot be decomposed into purer ones it is called a "pure state," otherwise an "impure state" ("mixture" in von Neumann's terminology).

[16] We find it preferable to base our discussion on the notion of "operations" as defined above instead of "observables" as used by Dirac and von Neumann. An "observable" in the technical sense is an idealization, which in general implies suitably defined limits of an infinite number of operations. It is thus a far less simple concept.

[17] We hope to discuss this question in another paper.

[18] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, 1955); I. E. Segal, see Ref. 9.

[19] It is not crucial to assume that the algebra contains the unit element; see Appendix 1. The norm is then defined as $\sup |\Phi(A)|/||A||$.

[20] It has been emphasized by H. Ekstein that, in general, one algebraic element will correspond to many different laboratory procedures which are equivalent insofar as they produce the same transformation of the states. For simplicity we shall, however, always speak of an "operation" instead of an "equivalence class of operations."

[21] J. von Neumann, see Ref. 18.

[22] I. E. Segal, see Ref. 9.

operation (or state) is defined in terms of a laboratory procedure. How do we find the corresponding element in the mathematical description? For the "operations" the question is partially answered by the assertion: An operation in the space–time region $\mathbf{B}$ corresponds to an element from $\mathfrak{A}(\mathbf{B})$. It is very likely that this simple statement provides not only a partial answer but a complete one because ultimately all physical processes are analyzed in terms of geometric relations of (unresolved) phenomena. In any case it is rather evident that one can construct a good mathematical representative of a Geiger counter coincidence arrangement using the subalgebras for finite regions.[23]

The remaining task is to bridge the gap between the physical and the mathematical description of a state. One possible attitude is, of course, to say that the state φ may be described physically (as well as mathematically) by the collection of probabilities $P(\varphi, A)$ for all pure operations A. This would mean that, once an ensemble has been prepared, "somehow" we make all sorts of monitoring experiments to find out which ensemble we have. The other attitude is to assume that the ensemble is prepared by means of a single, specific operation from an initial ensemble which is completely unknown. In practice both methods (preparation of ensemble by a "filtering" operation from an unknown ensemble and determination of ensemble by monitoring experiments) are used to supplement each other. In both respects it is clear that the physical interpretation of *states* is fixed once we know the correspondence between the mathematical and the physical description for the *operations*. It is, however, also clear that no actual experiment will enable us to establish a definite state.

Take, first, the case of monitoring experiments on an ensemble. These will provide us with a finite number of probabilities, measured with a finite accuracy. In the mathematical scheme this information characterizes exactly a neighborhood in the weak topology of $\mathfrak{A}^*$. Namely, if the probabilities are the transmission probabilities for a collection of pure operations A_i $(i = 1, \cdots, N)$ then we know that the state satisfies[24]

$$\varphi(I) = 1, \tag{6}$$

$$|\varphi(A_i^* A_i) - p_i| < \epsilon_i, \tag{7}$$

where p_i are the experimentally determined probabilities and ϵ_i the errors. If, alternatively, one wants to define the ensemble in terms of a single preparatory operation then the discussion is mathematically somewhat more difficult. Let T be the preparing operation and $\mathbf{R}_T$ its range (the image of $\mathfrak{A}^{*(+)}$ under T). Then the only certain knowledge about the prepared state is that it lies in $\mathbf{R}_T$. To obtain definite state we need an operation with a one-dimensional range. There are two reasons why such an ideal operation is impossible. The first has to do with the limited accuracy in the specification of T [the counterpart of the errors ϵ_i in Eq. (7)]. The other comes from the special structure of our algebra $\mathfrak{A}$. Namely, it is evident that no quasilocal operation can have a finite-dimensional range because an operation in a finite region has no effect on the physical situation in a causally disjoint region. While we are unable at the moment to give a precise analysis of the consequences of these two limitations we feel that the first one (limitation in accuracy) will result in the statement that we have in no actual experiment a precisely specified state but rather a weak neighborhood in $\mathfrak{A}^{*(+)}$. This is the conclusion relevant to the remaining discussion in this section. The other limitation, arising from the special nature of the quasilocal algebra probably implies that there exist many unitarily inequivalent irreducible representations of $\mathfrak{A}$. (See Sec. III and Ref. 5).

The foregoing discussion leads us now to the following:

Statement. Let $R^{(1)}$ and $R^{(2)}$ be two representations of $\mathfrak{A}$ and Ω_1, Ω_2 the subsets of states which correspond to density matrices in the two representation spaces. The two representations are *physically equivalent* if every weak neighborhood of any element of Ω_1 contains an element of Ω_2 and vice versa.

The notion of physical equivalence coincides exactly with that of "weak equivalence" as defined by Fell.[25] We can apply *Fell's equivalence theorem*, i.e., two representations are weakly equivalent if and only if they have the same kernel.[26]

The conclusion is thus that all faithful representations of $\mathfrak{A}$ are physically equivalent. The relevant object is the abstract algebra and not the representation. The selection of a particular (faithful) representation is a matter of convenience without physical

[23] This will be discussed in a separate paper. See also Ref. 3.

[24] We chose (arbitrarily) the normalization of the state. For the sake of symmetry with Eq. (7) we could write, instead of (6), equally well,

$$|\Phi(I) - 1| < \epsilon_0.$$

[25] See Ref. 7 and the last paragraph of Appendix I.

[26] The kernel of a representation is the collection of all elements of $\mathfrak{A}$ which are represented by zero.

implications. It may provide a more or less handy analytical apparatus.

It also follows that we should consider only faithful representations because, supposing for a moment that a nonfaithful representation with kernel $\mathbf{K}$ contained all physically relevant information then the only physically equivalent representations would be those with the same kernel. The relevant object is then not the algebra $\mathfrak{A}$ but the quotient $\mathfrak{A}/\mathbf{K}$. According to a well-known theorem this quotient is again a C^*-algebra, and we should have taken this algebra in the first place instead of $\mathfrak{A}$.

As a final remark we might add that it appears natural to assume that $\mathfrak{A}$ is primitive, i.e., that it has at least one representation which is both faithful and irreducible. It would be tempting to assume even that $\mathfrak{A}$ is simple, i.e., that all its representations are faithful.[27]

III. LOCAL AND GLOBAL PROPERTIES. SUPERSELECTION RULES

It was pointed out in the introduction that all actual experiments involve only operations in finite space-time regions. Hence it is natural to introduce the notion of a "partial state with respect to a region."

Definition. A partial state with respect to region $\mathbf{B}$ is a positive linear form over the algebra $\mathfrak{A}(\mathbf{B})$ or, alternatively speaking, an equivalence class of "global states" (positive linear forms over $\mathfrak{A}$) which coincide on $\mathfrak{A}(\mathbf{B})$.

The two alternative definitions are equivalent due to the following theorem which we use again later:

Theorem.[28] If $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are two C^*-algebras and $\mathfrak{A}_1 \subset \mathfrak{A}_2$ then every state (positive linear form) over $\mathfrak{A}_1$ can be extended to at least one state over $\mathfrak{A}_2$. A pure state over $\mathfrak{A}_1$ can be extended to a pure state over $\mathfrak{A}_2$.

It is of interest to understand the coupling between the partial states of different regions which results from algebraic relations between the various subalgebras $\mathfrak{A}(\mathbf{B})$. We shall call the partial states in regions $\mathbf{B}_1$ and $\mathbf{B}_2$ completely uncoupled if, choosing an arbitrary pair of equally normalized partial states $\varphi^{(1)} \in \mathfrak{A}(\mathbf{B}_1)^{*(+)}$ and $\varphi^{(2)} \in \mathfrak{A}(\mathbf{B}_2)^{*(+)}$, one can find a global state φ which is an extension of both $\varphi^{(1)}$ and $\varphi^{(2)}$. The extreme opposite of this situation (i.e., complete coupling) prevails if each partial state in $\mathbf{B}_1$ determines uniquely a partial state in $\mathbf{B}_2$ by the process of extension to $\mathfrak{A}$ and restriction to $\mathfrak{A}(\mathbf{B}_2)$.

On physical grounds we want:

(i) If $\mathbf{B}_2$ is contained in the causal shadow of $\mathbf{B}_1$ then the partial states in $\mathbf{B}_2$ are uniquely determined by those in $\mathbf{B}_1$ (causality).

(ii) If $\mathbf{B}_1$ and $\mathbf{B}_2$ are causally disjoint then the partial states in the two regions are essentially[29] uncoupled (locality).

Property (i) is equivalent to the algebraic requirement

$$\mathfrak{A}(\mathbf{B}_2) \subset \mathfrak{A}(\mathbf{B}_1) \tag{8}$$

for all regions $\mathbf{B}_2$ in the causal shadow of $\mathbf{B}_1$.[30] Property (ii) is related to the local commutativity postulate but we do not know whether this postulate is already enough to guarantee the lack of coupling for partial states in causally disjoint regions or whether some further structure property is needed.

We now give a brief intuitive (nonrigorous) discussion of some phenomena for which the distinction between global and local features plays a role.

A. Existence of Unitarily Inequivalent Irreducible Representations of $\mathfrak{A}$

A pure state over $\mathfrak{A}$ corresponds to a vector in some irreducible representation space of $\mathfrak{A}$. Two pure states belong to the same representation if and only if the one results from the other by transformation with an element of the algebra [in the sense of Eq. (4)].[31] Otherwise they belong to representations which are unitarily inequivalent.

We confine our attention to the states without infinitely extended correlations. These states are characterized by the property that an operation in region $\mathbf{B}$ does not affect the partial state in a far

---

[27] Compare B. Misra, "On the algebra of quasi-local operators of Quantum Field Theory," to be published. See, however, Appendix II for an example of a nonsimple algebra of physical interest.

[28] See, e.g., M. A. Neumark, Ref. 42.

[29] From the physical point of view it would not be necessary that the uncoupling is complete if the separation distance between $\mathbf{B}_1$ and $\mathbf{B}_2$ is finite but only that it becomes complete in the limit of infinite spacelike separation.

[30] Let $\mathfrak{A}_1$ and $\mathfrak{A}_2$ be two subalgebras of $\mathfrak{A}$ and $\mathfrak{A}_1^{\perp}$, $\mathfrak{A}_2^{\perp}$ those subspaces of $\mathfrak{A}^*$ which are composed of the linear forms vanishing respectively on $\mathfrak{A}_1$ and $\mathfrak{A}_2$. If the "partial states" over $\mathfrak{A}_1$ determine those over $\mathfrak{A}_2$ we have

$$\mathfrak{A}_2^{\perp} \supset \mathfrak{A}_1^{\perp}.$$

Thus

$$\mathfrak{A}_2^{\perp\perp} \subset \mathfrak{A}_1^{\perp\perp}.$$

But $\mathfrak{A}_i^{\perp\perp}$, considered as a subset of $\mathfrak{A}$, coincides with the uniform closure of $\mathfrak{A}_i$, i.e., with $\mathfrak{A}_i$ itself.

[31] Compare Appendix I, Kadison's Theorem, Ref. 53.

away region $\mathbf{B}'$ (apart from a change in normalization).[32] In symbols:

$$\varphi(QQ') \approx \varphi(Q)\varphi(Q')/\varphi(I) \tag{9}$$

if Q and Q' belong to the algebras of two far-separated regions.

Consider now an infinite collection of causally disjoint regions $\mathbf{B}_k$. Let Φ be a pure state without infinitely extended correlations; the corresponding partial state in $\mathbf{B}_k$ will be denoted by φ_k. It is clear that a transformation of Φ by an element from $\mathfrak{A}$ will not change the asymptotic tail ($k \to \infty$) of the sequence of partial states φ_k, apart from a common normalization factor, because any element of $\mathfrak{A}$ can be approximated to arbitrary precision by an operation in a finite region. The asymptotic tail of the sequence of partial states φ_k is thus common to all normalized states belonging to the same representation as Φ. It is a "unitary invariant". On the other hand the lack of coupling between partial states in causally disjoint regions [property (ii)] suggests that there is an enormous variety of possible asymptotically different sequences φ_k. This gives us then many unitarily inequivalent representations of $\mathfrak{A}$. They differ in the global aspects of their states but this difference is irrelevant as long as we are interested only in experiments in finite regions ("physical equivalence" of all these representations).

B. Lorentz Transformations

As postulated under item (5) in the introduction we have for every element L of the inhomogeneous Lorentz group an automorphism of the algebra: $A \to A^L$. The automorphism defines a corresponding transformation in the state space, namely

$$\varphi \to \varphi_L \quad \text{with} \quad \varphi_L(A) = \varphi(A^L). \tag{10}$$

This is a linear transformation in the Banach space $\mathfrak{A}^*$ which preserves the norm and transforms the positive cone $\mathfrak{A}^{*(+)}$ into itself. Since

$$(\varphi_{L_1})_{L_2} = \varphi_{L_2 L_1}, \tag{11}$$

we get in this way an "antirepresentation" of the Lorentz group (by isometric operators in a Banach space). In the special case when L is a translation by a 4-vector x we write A^x and φ_x for the transformed quantities.

Since Lorentz transformations are global operations, affecting the far away regions as strongly (or even stronger) than the regions nearby, they do not correspond to elements in the quasilocal algebra $\mathfrak{A}$. Indeed, if we assumed that there is an element $U(L)$ in $\mathfrak{A}$ such that

$$A^L = U(L)AU^{-1}(L) \quad \text{for all} \quad A \in \mathfrak{A}, \tag{12}$$

we would get an immediate contradiction.[33] We could then approximate $U(L)$ by an operation C in a finite region $\mathbf{B}$ such that $||U(L) - C|| < \epsilon$. Taking $A \in \mathfrak{A}(\mathbf{B}')$ with $\mathbf{B}'$ causally disjoint from $\mathbf{B}$, Eq. (12) would imply

$$||A^L - A|| < 2\,||A||\,\epsilon \quad \text{for all} \quad A \in \mathfrak{A}(\mathbf{B}'),$$

which is not true.

We may assume, however, that there are elements in $\mathfrak{A}$ which produce the same effect as a Lorentz transformation within an arbitrarily chosen finite region $\mathbf{B}$. Denoting such an element by $U_{\mathbf{B}}(L)$ we have instead of (12)

$$A^L = U_{\mathbf{B}}(L)AU_{\mathbf{B}}^{-1}(L) \quad \text{for all} \quad A \in \mathfrak{A}(\mathbf{B}). \tag{13}$$

of course the $U_{\mathbf{B}}(L)$ are not uniquely determined by (13).

We may ask next whether the Lorentz transformations can be represented by unitary operators [again denoted by $U(L)$] in an irreducible representation space of the algebra $\mathfrak{A}$. This is the situation assumed in the usual analytic apparatus of quantum field theory. We may call it "Lorentz invariance of the representation" as distinguished from the Lorentz invariance of the algebraic theory which was postulated in the introduction. From the discussion A it follows that this is only possible if all the states φ belonging to the representation have the property[34]

$$||\mathbf{B}_{(\varphi_L)} - \mathbf{B}_\varphi|| \to 0 \tag{14}$$

when the region $\mathbf{B}$ is moved to infinity keeping its shape fixed. In other words, for such a representation all partial states in far away regions must be Lorentz invariant. The requirement that the Lorentz transformations shall be represented by unitary operators in the Hilbert space (Lorentz invariance of the representation) is thus a very powerful restriction eliminating most of the representations discussed under A).

---

[32] "Far away" means the limit of an infinite spacelike separation.

[33] An automorphism of the type (12) is called an inner automorphism of the algebra. Our argument here shows that the Lorentz transformations are outer automorphisms.

[34] The symbol $\mathbf{B}_\varphi$ is used here to denote the partial state in $\mathbf{B}$ resulting from the restriction of φ.

In a Lorentz-invariant irreducible representation of $\mathfrak{A}$, the Lorentz operators $U(L)$ are, of course, obtainable as strong limits of the quasilocal operators since the strong closure of the collection of representatives of $\mathfrak{A}$ is the ring of all bounded operators on the Hilbert space. Consider a family of regions $\mathbf{B}_n$ such that $\mathbf{B}_{n+1} \supset \mathbf{B}_n$ and $\mathbf{B}_\infty = \cup\, \mathbf{B}_n$ is the whole of Minkowski space. The representatives of the corresponding family $U_{\mathbf{B}_n}(L)$ [see Eq. (13)] form a strongly (but not uniformly) converging sequence of operators due to the fact that every state in the representation space has the asymptotic property (14). The limit of the sequence is $U(L)$. In other representations in which the states have different asymptotic properties this sequence does not converge at all. This illustrates how global operations such as Lorentz transformations may be defined in suitable representation spaces as strong limits of quasilocal operations. The strong convergence of such a sequence of operators arises from the (common) asymptotic properties of all the states in the representation space. Strong convergence depends on the representation whereas uniform convergence does not.

C. Superselection Rules

In the usual formalism of Quantum field theory a superselection rule means that there are operators in the Hilbert space which commute with all observables. Typical examples of such "superselecting operators" are the total electric charge or the total baryon number. This customary representation of the algebra of observables is reducible. It can be decomposed into irreducible ones which we shall call "sectors." Each sector corresponds to a definite numerical value of the charge.[35]

We note first that the charge (as well as every other superselecting operator) is a global quantity. The distinction between the different sectors can therefore not be made by means of experiments in finite regions. A simple argument shows then that every sector is physically equivalent to every other sector. Let us demonstrate this for the two sectors corresponding to charge 3 and charge -1, respectively. We have to show that given an arbitrary state φ of charge 3 and an arbitrary finite set of elements $A_i \in \mathfrak{A}$ we can find a substitute state φ' with charge -1 so that the expectation values of the A_i in the two states differ by less than an arbitrarily prescribed tolerance ϵ. The way to construct φ' is physically evident. We change the physical situation described by φ adding 4 elementary particles of negative charge in a remote region of space. The effect of this added charge on the expectation values of the quasilocal quantities A_i tends to zero as the region is moved to infinity.

We conclude then from Fell's equivalence theorem that each single sector is a faithful representation of $\mathfrak{A}$.[36] A single sector contains already all relevant physical information. We note incidentally that we have here an example of a quasilocal algebra which has (at least) a denumerable infinity of unitarily inequivalent, Lorentz-invariant, faithful, irreducible representations (the various sectors).

In the standard treatment of field theory one considers the direct sum of all the sectors. If $\mathfrak{H}_n$ is the representation space of the sector with charge n then one uses the Hilbert space

$$\mathfrak{H} = \sum{}^{\oplus} \mathfrak{H}_n. \tag{15}$$

Let us denote the representation of $\mathfrak{A}$ in $\mathfrak{H}$ by R, the range of this representation (i.e., the set of operators in $\mathfrak{H}$ representing the elements of $\mathfrak{A}$) by $R(\mathfrak{A})$. It is instructive to observe now the difference between weak and uniform closure. Since $\mathfrak{A}$ is a C^*-algebra $R(\mathfrak{A})$ is already uniformly closed. In the decomposition (15) the general element is of the form

$$R(A) = \begin{bmatrix} \ddots & & & \\ & R_{-1}(A) & & 0 \\ & & R_0(A) & \\ & 0 & & R_1(A) \\ & & & & \ddots \end{bmatrix} \tag{16}$$

where R_n is the irreducible representation of sector n. Since each R_n is faithful, these irreducible representations are rigidly coupled together. In other words, a single one of the suboperators $R_n(A)$ uniquely determines A and hence it fixes all the other $R_m(A)$. Thus in particular the projection operator P_n on the subspace $\mathfrak{H}_n$ does not belong to $R(\mathfrak{A})$ because the only element of $R(\mathfrak{A})$ which is zero in some sector is the zero operator on $\mathfrak{H}$. Let us consider now the weak (or strong) closure of $R(\mathfrak{A})$. This is the von Neumann ring generated by $R(\mathfrak{A})$ and can be alternatively obtained as the bicommutant $R(\mathfrak{A})''$. Since the representations R_n are irreducible and unitarily inequivalent Schur's lemma implies that the commutant $R(\mathfrak{A})'$ consists

[35] For the sake of simplicity we pretend that the electric charge is the only superselected quantity and thus use the word "charge" in lieu of "superselected quantities."

[36] The theorem tells us that one sector is equally as faithful as the collection of all sectors taken together.

of all operators of the form $\sum c_n P_n$ where the c_n are an arbitrary bounded sequence of complex numbers and P_n is the projector on $\mathfrak{H}_n$. Taking the commutant again we find that a general element of $R(\mathfrak{A})''$ is of the form

$$K = \begin{pmatrix} \ddots & & & \\ & K_{-1} & & \\ & & K_0 & 0 \\ & 0 & & K_1 \\ & & & & \ddots \end{pmatrix} \tag{17}$$

where the K_n are arbitrary bounded operators on the corresponding sectors (which can be chosen completely independent of each other). Thus the weak (or strong) closure of $R(\mathfrak{A})$ is the (uncoupled) product[37] of all the full matrix rings $\mathfrak{B}(\mathfrak{H}_n)$. It contains in particular all the projectors P_n, all the bounded functions of the charge as well as the Lorentz operators $U(L)$ ("global" quantities).

IV. COMPARISON WITH OPERATOR APPROACH TO QUANTUM FIELD THEORY

The postulates of a purely algebraic theory which have been stated so far [items (1) through (6) in the introduction, and (i) and (ii) in Sec. III] are not as powerful as those in other approaches to quantum field theory (Wightman's axioms or those of Ref. 2). In some respects this is good because a few irrelevant restrictions which are customarily imposed are eliminated. In other respects, however, the scheme as presented here is quite incomplete. It does not yet contain a stability condition and we have not formulated the counterparts of the finer structure properties which can be stated in the operator form of the theory. We shall point out now some of the features which are lacking in the present formulation.

The bridge between the algebraic approach and the customary analytic apparatus is the assumption that there exists a state Φ_0 over the algebra $\mathfrak{A}$ which is called the physical vacuum state and is supposed to have the following properties:

(α) Φ_0 is Lorentz-invariant, i.e., $(\Phi_0)_L = \Phi_0$.

(β) Φ_0 is a vector state of an irreducible, faithful representation of $\mathfrak{A}$ in a separable Hilbert space $\mathfrak{H}$.[38]

(γ) The Hamiltonian[39] is a positive-semidefinite operator in $\mathfrak{H}$.

It appears to us that the existence of a vacuum state with properties (α) and (β) is no *sine qua non* for a physically meaningful theory. In particular, in a theory describing among other things particles with zero rest mass one may have doubts as to whether the assumptions (α) and (β) are physically reasonable.[40] On the other hand, it is clear that some stability condition like (γ) is absolutely essential. The condition (γ) has also been one of the most useful tools in Wightman's approach.

Let us compare now the algebraic approach with that of Ref. 2 which is conceptually almost identical but uses von Neumann rings $\mathbf{R}(\mathbf{B})$ instead of abstract C^*-algebras $\mathfrak{A}(\mathbf{B})$. Given any specific irreducible, faithful representation of $\mathfrak{A}$ (say the representation R_α) we have immediately also a system of von Neumann rings which we denote by $\mathbf{R}_\alpha(\mathbf{B})$ in the representation space $\mathfrak{H}_\alpha$. The ring $\mathbf{R}_\alpha(\mathbf{B})$ is just the weak closure of the concrete C^*-algebra of operators $R_\alpha\{\mathfrak{A}(\mathbf{B})\}$ [the representatives of $\mathfrak{A}(\mathbf{B})$ in the representation R_α]. This ring system will satisfy the conditions of locality and causality; namely $\mathbf{R}_\alpha(\mathbf{B}_1)$ and $\mathbf{R}_\alpha(\mathbf{B}_2)$ commute if $\mathbf{B}_1$ and $\mathbf{B}_2$ are causally disjoint, and $\mathbf{R}_\alpha(\mathbf{B}_2) \subset \mathbf{R}_\alpha(\mathbf{B}_1)$ if $\mathbf{B}_2$ is the causal shadow of $\mathbf{B}_1$. This follows immediately from the corresponding relations for the algebras $\mathfrak{A}(\mathbf{B})$. However, within the set of von Neumann rings one has one important operation which has no direct counterpart in our family of C^*-algebras. This is the passage from a ring $\mathbf{R}$ to its commutant R'. This operation has been extensively used in Ref. 2 to formulate more detailed structure relations of the ring system which are very interesting because they open a way for a discussion of gauge groups and a distinction between theories with interaction from trivial theories in terms of local observables.[41] One typical example of such a relation is the "additivity"

$$\mathbf{R}(\mathbf{B}_1 \cup \mathbf{B}_2) = \{\mathbf{R}(\mathbf{B}_1), \mathbf{R}(\mathbf{B}_2)\}''. \tag{18}$$

Considering only the special case in which $\mathbf{B}_1$ and $\mathbf{B}_2$ are causally disjoint this relation may be assumed

[37] In the sense of Dixmier: *Les Algèbres d'operateurs dans l'espace Hilbertien* (Gauthier-Villars, Paris, 1952), Chap. I, Sec. 2.2.

[38] The representation is determined by Φ_0 via the GNS construction. See Appendix I, Ref. 59. It is irreducible if Φ_0 is pure. The separability would follow from the irreducibility if the algebra were separable in the norm topology.

[39] It follows from (α) that the representation obtained by the GNS construction from Φ_0 is Lorentz-invariant. Hence there exists a 1-parameter group of unitary operators $U(\tau)$ representing the time translations in $\mathfrak{H}$. The Hamiltonian is the infinitesimal generator of this group.

[40] This question was studied in Ref. 8 but the argument given there is inconclusive in some respects.

[41] For some conjectures in this direction see R. Haag, Ann. Physik **11**, 29 (1963) and Proceedings of the Conference on Analysis in Function Spaces, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1963.

to hold for theories without superselection rules but to fail, for instance, in a theory with charged fields.

The question is therefore whether such relations as (18) can be given a meaning in the purely algebraic approach, in other words, whether they are independent of the particular representation α. Since the double commutant is the same as the weak closure this would be the case if all (faithful) representations of $\mathfrak{A}$ were *"locally quasi-equivalent."* The notion of "quasi-equivalence" of two representations was introduced by Mackey and is described in Appendix I. It is more restrictive than weak equivalence and less restrictive than unitary equivalence. By *"local"* quasi-equivalence of two representations we mean that for any (finite) region **B** the two representations of the subalgebra $\mathfrak{A}(\mathbf{B})$ are quasi-equivalent whereas the two representations of the full algebra $\mathfrak{A}$ need not be quasi-equivalent.

As pointed out in the Appendix, one may characterize quasi-equivalence also by the fact that the sets of states which appear as density matrices in the different representations are identical. Thus, the assumption of local quasi-equivalence of all representations means that each finite region **B** has a (universal) set of partial states which corresponds to the collection of density matrices in an arbitrary (faithful) representation of $\mathfrak{A}$ restricted to $\mathfrak{A}(\mathbf{B})$. In more intuitive language this assumption means two things. On the one hand, the results of measurements in a fixed finite region **B** shall be *uniformly* unaffected if one changes the state by "adding particles behind the moon," i.e.,

$$|\Phi(A) - \Phi'(A)| < \epsilon \, ||A|| \quad \text{for all} \quad A \in \mathfrak{A}(\mathbf{B}), \tag{19}$$

if Φ' results from Φ by a unitary operation in a very distant region. Secondly, there shall be no other limiting procedure leading to the construction of inequivalent irreducible representations of $\mathfrak{A}$ besides the one involving large separation in position space and discussed in the last section. In particular one might wonder about the asymptotic limit for high energies. Is it possible to have states which differ in the asymptotic tail of their high-energy behavior (i.e., which give different expectation values for local operations involving "infinitely high" energy transfer)? The answer is probably no. Thus the assumption of local quasi-equivalence seems to us at the moment not unreasonable.

Another even stronger, assumption which is not contradicted by any of our present knowledge is that of *"local unitary equivalence"* of all irreducible, faithful representations of $\mathfrak{A}$. This would mean that if R and S are two such representations of $\mathfrak{A}$ then $R(\mathfrak{A}(\mathbf{B}))$ and $S(\mathfrak{A}(\mathbf{B}))$ are unitarily equivalent for every finite **B**, however, in such a way that no intertwining operator exists which is independent of the region **B**.

ACKNOWLEDGMENTS

We profited considerably from discussions with various members of the "Summer Institute for Theoretical Physics" at Madison, Wisconsin. In particular we wish to thank A. S. Wightman, H. Araki, and H. J. Borchers for discussions, R. Sachs for the organization, and the National Science Foundation for financial support of this stimulating Institute. We are also indebted to Dr. H. Ekstein and Dr. J. Cook of Argonne National Laboratory for constructive criticism.

APPENDIX I. C*-ALGEBRAS[42]

A complex (or real) *algebra* $\mathfrak{A}$ is a complex (or real) linear space such that to each ordered pair A, $B \in \mathfrak{A}$ there corresponds an element $A\,B \in \mathfrak{A}$, called their *product*, which is bilinear and associative (in general not commutative). In the special case where $AB = BA$ for all $A, B \in \mathfrak{A}$, $\mathfrak{A}$ is said to be *Abelian*. $\mathfrak{A}$ is a *-algebra[43] if to each $A \in \mathfrak{A}$ there corresponds a $A^* \in \mathfrak{A}$, called the *adjoint* of A, so that $A \to A^*$ is a conjugate-linear mapping with the properties $A^{**} = A$ and $(AB)^* = B^*A^*$ for any $A, B \in \mathfrak{A}$. $\mathfrak{A}$ is a *normed algebra* if to each $A \in \mathfrak{A}$ there corresponds a nonnegative number $||A||$, called the *norm* of A, in such a way that $||A|| > 0$ whenever $A \neq 0$ and, for any two A, $B \in \mathfrak{A}$ and any number λ, $||A + B|| \leq ||A|| + ||B||$, $||\lambda A|| = |\lambda| \, ||A||$ and $||AB|| \leq ||A|| \cdot ||B||$. Taking $||A - B||$ to be the distance of two elements A, B one defines on $\mathfrak{A}$ the topology of a metric vector space called its *uniform* or *norm topology*. If $\mathfrak{A}$ is both a *-algebra and a normed algebra we call it a **-normed algebra*[44] provided $||A^*|| = ||A||$ for all $A \in \mathfrak{A}$. A very important class of *-normed algebras, that of *Banach *-algebras*[45] is obtained by requiring completeness in the norm topology (i.e., convergence of all Cauchy sequences of elements of $\mathfrak{A}$ with respect to the norm to some element in $\mathfrak{A}$). One easily sees that any norm-closed *-algebra of bounded

[42] For general sources of information on C^*-algebras see M. A. Neumark, *Normierte Algebren* (VEB Deutscher Verlag der Wissenschaften, Berlin, 1959), Chap. V, Sec. 24; and C. E. Rickart, *General Theory of Banach Algebras* (D. Van Nostrand, Inc., New York, 1960), Chap. IV, Sec. 8. In general we use Rickart's terminology.

[43] Called *symmetrische Algebra* by Neumark.

[44] In Rickart's terminology. Neumark's is: *normierte symmetrische Algebra.*

[45] In Rickart's terminology. Neumark's is: *vollständige normierte symmetrische Algebra.*

operators on a Hilbert space is a Banach *-algebra (with respect to the usual addition, scalar multiplication, product, adjoint operation, and norm of bounded operators). However, the converse is not true and it has been shown by Gelfand and Neumark[46] that by requiring $||A^*A|| = ||A||^2$ for all $A \in \mathfrak{A}$ one singles out those particular Banach *-algebras which are isomorphic to (i.e., concretely realizable as) norm-closed *-algebras of bounded operators on some Hilbert space. These algebras are the interesting ones for quantum theory and we will call them *C*-algebras.*[47] A typical example of a C^*-algebra is the algebra $\mathfrak{B}(\mathfrak{H})$ of all bounded operators on some Hilbert space $\mathfrak{H}$.

A linear mapping L of an algebra $\mathfrak{A}$ into an algebra $\mathfrak{A}_1$ which preserves products is called a *homomorphism.* If $\mathfrak{A}$ and $\mathfrak{A}_1$ are *-algebras and adjoints are mapped into adjoints, we speak of a **-homomorphism.* The *kernel* of the homomorphism L is the set of elements of $\mathfrak{A}$ which are mapped into the zero of $\mathfrak{A}_1$. A (*-)homomorphism is one-to-one if and only if its kernel reduces to zero, in which case it is called a **-isomorphism.* A linear subspace J of an algebra $\mathfrak{A}$ is a *left* (resp. *right,* resp. *two-sided*) *ideal* if $A \in \mathfrak{J}$ and $B \in \mathfrak{A}$ imply $BA \in \mathfrak{J}$ (resp. $AB \in \mathfrak{J}$, resp. BA and $AB \in \mathfrak{J}$). If $\mathfrak{A}$ is a *-algebra and $A \in \mathfrak{J}$ implies $A^* \in \mathfrak{J}$, we speak of a **-ideal.* Two-sided ideals (*-ideals) are in one-to-one correspondence with homomorphisms (*-homomorphisms), the first being the kernels of the latter. Given a two-sided ideal (*-ideal) $\mathfrak{J}$ of $\mathfrak{A}$, the corresponding homomorphism (*-homomorphism) is obtained by assigning to each element $A \in \mathfrak{A}$ its class modulo the elements of $\mathfrak{J}$. The algebra (*-algebra) of these equivalence classes is denoted $\mathfrak{A}/\mathfrak{J}$ and called the *quotient of* $\mathfrak{A}$ *by* $\mathfrak{J}$. A *representation* R of the algebra $\mathfrak{A}$ on a linear space $\mathfrak{H}$ is a homomorphism of $\mathfrak{A}$ into the algebra of linear operators on $\mathfrak{H}$: to each $A \in \mathfrak{A}$ R assigns a linear operator $R(A)$ on $\mathfrak{H}$ the correspondence $A \to R(A)$ respecting linear combinations and products. $R(A)$ is called the *representative* of A in R and the set of representatives of all $A \in \mathfrak{A}$ in R is called the *range* of R. R is *faithful* if it is one-to-one and *algebraically irreducible* if the only subspaces of $\mathfrak{H}$ invariant for all $R(A)$ are $\{0\}$ and $\mathfrak{H}$ itself. In the case where $\mathfrak{A}$ is a *-algebra, $\mathfrak{H}$ is a Hilbert space and R is a *-homomorphism into $\mathfrak{B}(\mathfrak{H})$ we speak of a **-representation* and call R *(topologically) irreducible*[48] if $\{0\}$ and $\mathfrak{H}$ are the only *closed* subspaces of $\mathfrak{H}$ invariant for all $R(A)$. Furthermore, we call R *continuous* in case $||R(A)|| \leq C\,||A||$ for all $A \in \mathfrak{A}$ and some positive constant C. One shows that every *-representation of a Banach *-algebra[49] is continuous. An ideal $\mathfrak{J} \subset \mathfrak{A}$ is called *primitive* if it is the kernel of an algebraically irreducible representation. $\mathfrak{A}$ itself is *primitive* if $\{0\}$ is a primitive ideal, i.e., if $\mathfrak{A}$ admits a faithful algebraically irreducible representation. The *radical* of an algebra $\mathfrak{A}$ is the intersection of all its primitive ideals. In the case where $\mathfrak{A}$ is a *-algebra its **-radical* is the intersection of the kernels of all topologically irreducible *-representations. $\mathfrak{A}$ is *simple* if it contains only the trivial ideals $\{0\}$ and $\mathfrak{A}$ itself. It is *semisimple* (**-semisimple*[50]) if its radical (*-radical) reduces to zero. Simplicity obviously implies that all representations are faithful which implies primitivity in the case of a Banach *-algebra. Semisimplicity (*-semisimplicity) means only that the collection of all algebraically irreducible representations (topologically irreducible *-representations) is *faithful* in the sense that no two different elements of $\mathfrak{A}$ can have the same representative in all of them (we also say that the irreducible representations *separate* $\mathfrak{A}$).

The relation between separation properties of representations and the ideal structure of $\mathfrak{A}$ is greatly simplified in the case of C^*-algebras due to several important peculiarities. First, every *-isomorphism of a C^*-algebra into another C^*-algebra is norm-preserving. Second, every closed ideal $\mathfrak{J}$ is a *-ideal and the corresponding quotient algebra $\mathfrak{A}/\mathfrak{J}$ (equipped with its natural norm

$$||A + \mathfrak{J}|| = \inf_{h \in \mathfrak{J}} ||A + h||)$$

is itself a C^*-algebra.[51] Combining these facts, one finds that every *-homomorphism of a C^*-algebra has a uniformly closed range. In particular, the range of a *-representation of a C^*-algebra is itself a C^*-algebra, the representation being norm-pre-

[46] I. M. Gelfand and M. A. Neumark, Mat. Sb. **12**, 197 (1943).

[47] Here we depart from Rickart's terminology who calls *B*-algebra* the abstract C^*-algebra and reserves the term *C*-algebra* for a concrete norm-closed algebra of operators on a Hilbert space (we speak in this case of a *concrete C*-algebra*). Our C^*-algebras (Rickart's B^*-algebras) are called by Neumark *vollreguläre vollständige Algebren.* Note that the condition $||A^*A|| = ||A||^2$ is evidently fulfilled in an operator algebra and that the distinction between abstract and concrete *C*-algebras* is important because different concrete *C*-algebras* can define the same abstract C^*-algebra.

[48] We always use the word irreducible to mean topologically irreducible.

[49] Whether or not $\mathfrak{A}$ contains an identity. See Rickart (Ref. 42), Theorem (4.1.20).

[50] Neumark uses *reduziert* for *-semisimple and *reduzierendes Ideal* for *-radical (in the case of Banach *-algebras).

[51] M. A. Neumark, Ref. 42, Chap. V, Sec. 24, Theorems 6 and 3.

serving if it is faithful. Further, a C^*-algebra is always semisimple and *-semisimple (its radical and *-radical both reducing to zero).[52] Another important result, due to Kadison,[53] is that every topologically irreducible *-representation of a C^*-algebra is algebraically irreducible. Finally, every algebraically irreducible representation of a C^*-algebra in a complex linear space is algebraically equivalent to a *-representation in a Hilbert space. Therefore every primitive ideal is the kernel of an irreducible *-representation. A primitive C^*-algebra can accordingly be defined as a C^*-algebra having at least one faithful irreducible *-representation.

Two representations (*-representations) R and S of $\mathfrak{A}$ on the respective spaces $\mathfrak{H}$ and $\mathfrak{K}$ are called *algebraically (unitarily) equivalent* if there exists a regular linear (unitary) operator U from $\mathfrak{H}$ onto $\mathfrak{K}$ such that

$$UR(A) = S(A)U \quad \text{for all} \quad A \in \mathfrak{A}. \tag{20}$$

Any linear (bounded linear) operator U from $\mathfrak{H}$ into $\mathfrak{K}$ satisfying (20) is called an *intertwining operator* for R and S. The set of such intertwining operators will be denoted by $\mathfrak{R}(R, S)$.[54] In the case of *-representations, one of the two following situations prevails: either $\mathfrak{R}(R, S) = \{0\}$ or there exist invariant closed subspaces $\mathfrak{H}_1$ and $\mathfrak{K}_1$ such that the restrictions of R and S on those subspaces (called subrepresentations of R and S) are unitarily equivalent. In particular, if R and S are irreducible and inequivalent, $\mathfrak{R}(R, S) = \{0\}$ and $\mathfrak{R}(R, R)$ consists of the multiples of unity (those two facts being generalizations of Schur's lemma). Note that $\mathfrak{R}(R, R)$ is a von Neumann ring, namely the commutant of $R(\mathfrak{A})$.

Two *-representations R and S of $\mathfrak{A}$ will be called *disjoint*[55] if $\mathfrak{R}(R, S) = \{0\}$, i.e., if R and S contain no subrepresentations which are unitarily equivalent. We consider now the decomposition of a representation into disjoint parts. Let R be a *-representation on $\mathfrak{H}$ and $\mathfrak{H}_1$ be a (closed) subspace of $\mathfrak{H}$ with projector E. $\mathfrak{H}_1$ is invariant in R if and only if $E \in \mathfrak{R}(R, R)$. In that case its orthogonal complement $\mathfrak{H}_1^\perp$ is also invariant and the restrictions of R on $\mathfrak{H}_1$ and $\mathfrak{H}_1^\perp$ (subrepresentations) are *disjoint* if and only if E is in the center of $\mathfrak{R}(R, R)$ (the center of an algebra being the set of its elements which commute with all others). A *-representation R of $\mathfrak{A}$ is called *primary*[55] if the center of $\mathfrak{R}(R, R)$ contains only the multiples of unity (so if no two subrepresentations of R are disjoint). We have seen that the range $R(\mathfrak{A})$ of R is closed in the norm-topology of operators on $\mathfrak{H}$, but it is in general not closed in the weak topology of operators on $\mathfrak{H}$ (the weak closure of $R(\mathfrak{A})$ is the von Neumann ring $R(\mathfrak{A})''$ which contains in general many more operators than $R(\mathfrak{A})$ itself). R is primary if $R(\mathfrak{A})''$ is a *factor* in the sense of von Neumann.

Two representations R and S of $\mathfrak{A}$ are called *quasi-equivalent*[56] if they have the same kernel (i.e., if their ranges are *-isomorphic) and if this *-isomorphism extends to the weak closures of the ranges in the respective weak topologies of operators on the representation spaces. The *-representations R and S are quasi-equivalent if and only if no subrepresentation of the one is disjoint from the other. If R and S are primary then they are either quasi-equivalent or disjoint.

A role of primary importance in the study of a Banach *-algebra is played by its *positive forms*. A linear form Φ on $\mathfrak{A}$ is called *continuous* if $|\Phi(A)| \leq C\,||A||$ for all $A \in \mathfrak{A}$ and some positive constant C. The smallest such constant is denoted by $||\Phi||$ and called the *norm* of Φ. Under this norm the set of all continuous forms on $\mathfrak{A}$ is a Banach space called the *dual space of* $\mathfrak{A}$ and denoted by $\mathfrak{A}^*$. (The Banach space topology of $\mathfrak{A}^*$ is called its *uniform* or *norm topology*.) Another topology of interest on $\mathfrak{A}^*$ is its *weak topology* (with respect to $\mathfrak{A}$) characterized by the pseudo-norms $N_A(\Phi) = |\Phi(A)|$ where A runs through $\mathfrak{A}$ (or by the complete set of neighborhoods of zero $\mathbf{V}_{\{A_i,\epsilon\}}$, $A_1, A_2, \cdots, A_n \in \mathfrak{A}$, $\epsilon > 0$, where $\mathbf{V}_{\{A_i,\epsilon\}}$ consists of all $\Phi \in \mathfrak{A}^*$ such that $|\Phi(A_i)| < \epsilon$, $i = 1, 2, \cdots, n$). A linear form Φ is positive if $\Phi(A^*A) \geq 0$ for all $A \in \mathfrak{A}$. If $\mathfrak{A}$ has a unit I the continuity of Φ follows from the positivity and one has $||\Phi|| = \Phi(I)$. Positive forms are also ipso facto continuous for C^*-algebras with or without unit.[57] The set of all positive continuous forms on $\mathfrak{A}$ (or *states* on $\mathfrak{A}$) is called the *positive cone* of $\mathfrak{A}^*$ and denoted by $\mathfrak{A}^{*(+)}$. We will denote by $\mathbf{\Sigma}$ and $\dot{\mathbf{\Sigma}}$ the subsets consisting of all $\Phi \in \mathfrak{A}^*$ such that $||\Phi|| \leq 1$ and $||\Phi|| = 1$, respectively ($\mathbf{\Sigma}$ is the *unit ball* of $\mathfrak{A}^*$). By Tychonov's theorem $\mathfrak{A}^{*(+)} \cap \mathbf{\Sigma}$ (and $\mathfrak{A}^{*(+)} \cap \dot{\mathbf{\Sigma}}$ if $\mathfrak{A}$ has a unit) are compact subsets of $\mathfrak{A}^*$ in its weak topology. As a compact convex set $\mathfrak{A}^{*(+)} \cap \dot{\mathbf{\Sigma}}$ has extremal elements and is equal

[52] M. A. Neumark, Ref. 42, Chap. V, Sec. 24, Theorem 4.
[53] R. V. Kadison, Proc. Natl. Acad. Sci. U. S. 43, 273 (1957).
[54] See G. W. Mackey, "The Theory of Group Representations," University of Chicago, mimeographed lecture notes.
[55] See Mackey, Ref. 54.
[56] For the notion of quasi-equivalence, see Mackey, Ref. 54, Chap. I.
[57] See C. E. Rickart, Ref. 42, Theorems (4.5.14), (4.5.11), and (4.8.14).

to the weak closure of their convex hull (theorem of Krein-Milman). The state $\Phi \in \mathfrak{A}^{*(+)} \cap \dot{\Sigma}$ is called *extremal* or *pure* if it cannot be written as $\lambda_1\Phi_1 + \lambda_2\Phi_2$ with $\Phi_1, \Phi_2 \in \mathfrak{A}^{*(+)} \cap \dot{\Sigma}$, $\Phi_1 \neq \Phi_2$, $0 < \lambda_1 < 1$, $\lambda_1 + \lambda_2 = 1$).

The importance of positive linear forms for Banach *-algebras lies in their connection with *-representations. Each *-representation of a *-algebra is (by transfinite induction) the direct sum of *-representations all of which (except the null representation) are cyclic.[58] Now for a Banach *-algebra with approximate unit, in particular for a C^*-algebra, giving a positive (*ipso facto* continuous) linear form Φ amounts to the same thing as specifying a cyclic representation R and a cyclic vector ξ. This representation is irreducible if and only if Φ is pure. Given R and ξ we have $\Phi(A) = (\xi \,|R(A)|\, \xi)$. Conversely, given Φ, its *null space* $\mathfrak{N}$ (i.e., the set of elements $A \in \mathfrak{A}$ for which $\Phi(A^*A) = 0$ or, equivalently, for which $\Phi(A^*B) = 0$ for all $B \in \mathfrak{A}$) is a left ideal in $\mathfrak{A}$ and one recovers consistently the vectors $R(A)\xi$, their scalar products $(R(A)\xi \mid R(B)\xi)$, and the operator $R(C)$ acting on $R(A)\xi$ by the following identification:

$$\begin{aligned} &\text{vector } R(A)\xi \leftrightarrow \text{class (modulo } \mathfrak{N}\text{) of the algebraic elements } A + \mathfrak{N},\\ &\text{scalar product} \quad (R(B)\xi \mid R(A)\xi) = \Phi(B^*A),\\ &\text{action of operator} \quad R(C)R(A)\xi \leftrightarrow CA + \mathfrak{N}. \end{aligned}$$

$\mathfrak{H}$ is then constructed by completion and the operator $R(C)$ in the complete $\mathfrak{H}$ by continuous extension. The cyclic vector ξ corresponds to $I + \mathfrak{N}$ if I exists and is otherwise obtained with the help of an approximate unity.[59]

It is useful to characterize the relations between different *-representations of a C^*-algebra in terms of certain subsets of $\mathfrak{A}^*$ determined by their representation spaces. Let R be a *-representation of $\mathfrak{A}$ on the Hilbert space $\mathfrak{H}_R$. Given $\Psi \in \mathfrak{H}_R$ and $A \in \mathfrak{A}$ we denote by $\omega_\Psi(A)$ the expectation value of $R(A)$ in the vector Ψ:

$$\omega_\Psi(A) = (\Psi|\, R(A)\, |\Psi) = \operatorname{Tr} R(A)\, |\Psi)(\Psi|.$$

We have thus defined a positive form ω_Ψ on $\mathfrak{A}$ which we call the *vector state determined by* $\Psi \in \mathfrak{H}_R$. When Ψ runs through $\mathfrak{H}_R$, ω_Ψ runs through a subset $\omega(R)$ of $\mathfrak{A}^{*(+)}$ which can be shown to be uniformly closed in $\mathfrak{A}^*$. For cyclic representations $\omega(R)$ determines R up to unitary equivalence; for two cyclic representations R and R' of $\mathfrak{A}$ to be unitarily equivalent, it is necessary and sufficient that $\omega(R) = \omega(R')$.[60] We now pass from $\omega(R)$ to its convex hull conv $\{\omega(R)\}$ (i.e., we consider all finite linear combinations of its elements with positive coefficients). If we close this convex hull respectively in the uniform and in the weak topology of $\mathfrak{A}^*$, we get two sets of states $\overline{\overline{\text{conv}}}\ \{\omega(R)\}$ and $\overline{\text{conv}}\ \{\omega(R)\}$ which respectively determine R up to quasi-equivalence and weak equivalence.[61] The elements of the uniform closure $\overline{\overline{\text{conv}}}\ \{\omega(R)\}$ can be characterized as the set of states Φ of the form

$$\Phi(A) = \operatorname{Tr}(\Phi_{op} \cdot A) \tag{21}$$

where Φ_{op} is any positive linear operator on $\mathfrak{H}_R$ with finite trace (the norm $||\Phi||$ being equal to the trace of Φ_{op}. These states will be referred to as the *density matrices* in the representation R. The elements of conv $\{\omega(R)\}$ are correspondingly the *density matrices of finite rank* in R. The linear spans of $\overline{\overline{\text{conv}}}\ \{\omega(R)\}$ and conv $\{\omega(R)\}$ are obtained by taking Φ_{op} in (21) to be respectively the operators of the trace class and of finite rank on $\mathfrak{H}_R$. We will denote them accordingly by $\mathfrak{T}(\mathfrak{A}, R)$ and $\mathfrak{F}(\mathfrak{A}, R)$. One has $\mathfrak{T}^+(\mathfrak{A}, R) = \overline{\overline{\text{conv}}}\ \{\omega(R)\}$ and $\mathfrak{F}^+(\mathfrak{A}, R) = \text{conv}\ \{\omega(R)\}$ where $^+$ indicates the restriction to positive elements. Let $t^S(R)$, $t^s(R)$, $t^W(R)$, $t^w(R)$ denote the topologies respectively defined on $\mathfrak{A}$ by the strongest, the strong, the σ-weak and the weak topologies of operators on $\mathfrak{H}_R$ (those topologies are not separating if R is not faithful). $\mathfrak{T}(\mathfrak{A}, R)$, resp. $\mathfrak{F}(\mathfrak{A}, R)$ [$\mathfrak{T}^+(\mathfrak{A}, R)$, resp. $\mathfrak{F}^+(\mathfrak{A}, R)$] can be characterized as the set of linear forms on $\mathfrak{A}$ (of positive linear forms on $\mathfrak{A}$) continuous with respect to either $t^S(R)$ or $t^s(R)$, resp. either $t^W(R)$ or $t^w(R)$. So quasi-equivalence of R and R' means that $t^S(R) = t^S(R')$, or equivalently $t^s(R) = t^s(R')$.

Let S and T be two *-representations of the C^*-algebra $\mathfrak{A}$, and let Ker (S) and Ker (T) be their kernels. Fell's equivalence theorem states that $\underline{\text{Ker}}\ (S) \supseteq \text{Ker}\ (T)$ is equivalent to $\omega(S) \subseteq \underline{\text{conv}}\ \{\omega(T)\}$, or alternatively to $\omega(S) \cap \Sigma \subseteq \underline{\text{conv}}\ \{\omega(T) \cap \Sigma\}$ or again to $\omega(S) \cap \dot{\Sigma} \subseteq \text{conv}\ \{\omega(T) \cap \Sigma\}$. If this is the case Fell calls S *weakly contained* in T. If S and T are weakly con-

[58] A representation R in the space $\mathfrak{H}$ is called *cyclic* with *cyclic vector* $\xi \in \mathfrak{H}$ if the set of vectors $R(\mathfrak{A})\ \xi$ is dense in $\mathfrak{H}$.

[59] This construction is due to I. M. Gelfand and M. A. Neumark, Isvertija Ser. Mat. **12**, 445 (1948); and I. E. Segal, Bull. Am. Math. Soc. **53**, 73 (1947). We call it the GNS construction. See M. A. Neumark, Ref. 42, Chap. IV, Sec. 17.3 or, for the case of an algebra without unit, C. E. Rickart, Ref. 42, Chap. IV, Sec. 5.

[60] These results can be inferred from R. V. Kadison, Trans. Am. Math. Soc. **103**, 304 (1962). $\omega(R) \subseteq \omega(R')$ means that R is unitarily equivalent to some subrepresentation of R'.

[61] For this characterization of quasi-equivalence, see Z. Takeda, Tôhoku Mat. J. **6**, 212 (1954).

tained in each other they are *weakly equivalent* (for us *physically equivalent*). If S is cyclic with cyclic vector Ψ it is sufficient for S to be weakly contained in T that $\omega_\Psi \in \overline{\text{conv}}\,\{\omega(T)\}$. If S and T are irreducible and S is weakly contained in T then $\omega(S)$ is already contained in the weak closure of $\omega(T)$.

APPENDIX II. NONSIMPLICITY OF THE FERMION CURRENT ALGEBRA[62]

A. Finite-Dimensional Case

Let $\mathfrak{E}$ be an n-dimensional metric vector space over the complex numbers. Corresponding to each vector x from $\mathfrak{E}$ we consider two algebraic elements $a^*(x)$ and $a(x)$ (adjoints of each other). They shall satisfy the commutation relations of creation and destruction operators in Fermi statistics,

$$\begin{aligned} a^*(x)a(y) + a(y)a^*(x) &= (x, y), \\ a^*(x)a^*(y) + a^*(y)a^*(x) &= 0, \end{aligned} \tag{22}$$

and the "creators" $a^*(x)$ shall depend linearly on x. The *-algebra generated by the a^* and a will be denoted by $\mathfrak{C}$.[63] A representation of $\mathfrak{C}$ is obtained (the standard representation in physical applications) by stipulating that the representation space $\mathfrak{H}(\mathfrak{E})$ shall contain a vector Φ_0 satisfying

$$a(x)\Phi_0 = 0 \quad \text{for all} \quad x \in \mathfrak{E} \tag{23}$$

and that the other vectors of $\mathfrak{H}(\mathfrak{E})$ are obtained from Φ_0 by application of polynomials of the a^*. From the commutation relations one infers immediately that the space $\mathfrak{H}$ has 2^n dimensions, that $\mathfrak{C}$ has 4^n linearly independent elements and is isomorphic to the full matrix ring over $\mathfrak{H}$. Thus $\mathfrak{C}$ is simple.

If $\Pi \in \mathfrak{C}$ is a product of p creators and q annihilators (in any order) we define the *grade of* Π to be the difference $p - q$. $\mathfrak{C}$ thus becomes a graded algebra whose zero-grade part will be called $\mathfrak{C}_0$. $\mathfrak{C}_0$ is represented in $\mathfrak{H}$ by operators which leave the homogeneous subspaces $\mathfrak{H}_p$ invariant. ($\mathfrak{H}_p$ is that subspace which is generated from Φ_0 by homogeneous polynomials of the a^* of order p; p runs from 0 to n.) Calling $R_p(\mathfrak{C}_0)$ the restriction of the representation of $\mathfrak{C}_0$ to $\mathfrak{H}_p$ one sees by counting dimensions that the R_p are a separating family of irreducible inequivalent representations of $\mathfrak{C}_0$. Therefore $\mathfrak{C}_0$ is a semisimple (but not simple) finite dimensional algebra. Note that $\mathfrak{C}_0$ can be regarded as the "*algebra of currents*" where the "*current*" $j(K)$ corresponding to the linear operator K on $\mathfrak{E}$ is defined by

$$j(K) = \sum_{i,k} \langle x_i \,|K|\, x_k\rangle a^*(x_i)a(x_k) \tag{24}$$

(x_i being a complete orthonormal basis of $\mathfrak{E}$).

B. Infinite-Dimensional Case

Instead of the finite-dimensional $\mathfrak{E}$ we take now an infinite-dimensional Hilbert space $\mathfrak{H}$. For any finite dimensional subspace $\mathfrak{E} \subset \mathfrak{H}$ we can consider the algebra $\mathfrak{C}(\mathfrak{E})$ has previously defined and one sees easily that for $\mathfrak{E}_2 \supset \mathfrak{E}_1$ the algebra $\mathfrak{C}(\mathfrak{E}_1)$ is is canonically embedded in $\mathfrak{C}(\mathfrak{E}_2)$ (as a normed *-algebra). Thus we can define $\mathfrak{C}(\mathfrak{H})$ as the completion of the union of all the $\mathfrak{C}(\mathfrak{E})$ for the finite-dimensional subspaces $\mathfrak{E} \subset \mathfrak{H}$. The fact that each $\mathfrak{C}(\mathfrak{E})$ is simple implies that all *-representations of $\mathfrak{C}(\mathfrak{H})$ are faithful and isometric, i.e., that also $\mathfrak{C}(\mathfrak{H})$ is simple.[64]

In the case of a free Dirac field $\mathfrak{H}$ is the direct sum of two spaces $\mathfrak{H}_1$ and $\mathfrak{H}_2$ which correspond, respectively, to the states of a single electron and to those of a single positron. We are interested now in the zero grade part of $\mathfrak{C}(\mathfrak{H})$. This algebra $\mathfrak{C}_0$ may be regarded as the algebra of currents in the theory of a free Dirac field. We consider the two familiar irreducible representations of $\mathfrak{C}$:

(1) the old-fashioned one which results if we assume the existence of a state Φ_0 satisfying

$$a(x)\Phi_0 = 0 \quad \text{for all} \quad x \in \mathfrak{H}; \tag{25}$$

(2) the "charge symmetric" one in which one assumes a state Ψ_0 satisfying

---

[62] We are indebted to Professor H. Araki for pointing out to us the main facts described in this appendix.

[63] $\mathfrak{C}$ is the Clifford algebra over the space $\mathfrak{E} \oplus \bar{\mathfrak{E}}$ with respect to the bilinear scalar product

$$g(x \oplus \bar{y}, x' \oplus \bar{y}') = \tfrac{1}{2}\{(x, y') + (y, x)\} \quad x \in \mathfrak{E}, y \in \mathfrak{E}.$$

Here $\bar{\mathfrak{E}}$ is the "complex conjugate" of $\mathfrak{E}$; i.e., it is isomorphic to $\mathfrak{E}$ as an additive group:

$$x \in \mathfrak{E} \leftrightarrow \bar{x} \in \bar{\mathfrak{E}},$$

but its scalar multiplication reverses the sign of i:

$$\overline{((\alpha + i\beta)x)} = (\alpha - i\beta)\bar{x}.$$

Since (x, y) is Hermitian symmetric the form g is bilinear. $\mathfrak{C}$ can be defined as the quotient of the tensor algebra over $\mathfrak{E} \oplus \bar{\mathfrak{E}}$ by an ideal $\mathfrak{T}$ which is generated by the tensors $\xi \times \xi - g(\xi, \xi)$ with $\xi \in \mathfrak{E} \oplus \bar{\mathfrak{E}}$. One has $a^*(x) = x \oplus 0 \bmod \mathfrak{T}$ and $a(y) = 0 \oplus \bar{y} \bmod \mathfrak{T}$. By extension of the adjoint operation

$$x \oplus \bar{y} \leftrightarrow y \oplus \bar{x}$$

one defines on $\mathfrak{C}$ the structure of a *-algebra. Equation (23) defines a faithful realization of $\mathfrak{C}$ by the linear operators on the space $\mathfrak{H}(\mathfrak{E})$ which latter coincides with the Grassmann algebra over $\mathfrak{E}$.

[64] Since it is known to have many inequivalent irreducible representations it is an NGCR-algebra.

$$a(x)\Psi_0 = 0 \quad \text{for all} \quad x \in \mathfrak{H}_1, \qquad a^*(x)\Psi = 0 \quad \text{for all} \quad x \in \mathfrak{H}_2. \tag{26}$$

Both representations, restricted to $\mathfrak{C}_0$, split up into irreducible parts corresponding to the different values of the charge. Now in Case 1 none of those subrepresentations of $\mathfrak{C}_0$ is faithful. In the subspace corresponding to charge n all operators having more than n annihilators on the right have zero representatives. Thus $\mathfrak{C}_0$ has nontrivial ideals and is accordingly not simple. On the other hand, in the charge-symmetric representation of $\mathfrak{C}$ (Case 2) all the subrepresentations of $\mathfrak{C}_0$ corresponding to a fixed value of the charge are faithful. This is an immediate consequence of the semisimplicity of the $\mathfrak{C}(\mathfrak{E})$ for finite-dimensional $\mathfrak{E}$.

Note on Wigner's Theorem on Symmetry Operations

V. BARGMANN
Palmer Physical Laboratory, Princeton University, Princeton, New Jersey
(Received 28 January 1964)

Wigner's theorem states that a symmetry operation of a quantum system is induced by a unitary or an anti-unitary transformation. This note presents a detailed proof which closely follows Wigner's original exposition.

INTRODUCTION

THE states of a quantum system S are described by unit vectors f (i.e., vectors of norm 1)[1] in some Hilbert space $\mathfrak{H}$. Assume that, conversely, to every unit vector f in $\mathfrak{H}$ corresponds a state of S.[2] This correspondence is, of course, not one-to-one since f and $f\tau$ describe the same state if τ is a scalar factor of modulus 1. The states of S are therefore in a one-one correspondence with *unit rays* $\mathbf{f}$, a unit ray being defined as the set of all vectors of the form $f_0\tau$, where f_0 is a fixed unit vector in $\mathfrak{H}$ and τ any scalar of modulus 1. Any significant statement in quantum theory is therefore a statement about unit rays.

Every vector f contained in the ray f ($f \in \mathbf{f}$) will be called a *representative* of $\mathbf{f}$. The transition probability from a state $\mathbf{f}$ to a state $\mathbf{g}$ equals $|(f, g)|^2$ where f, g are representatives of the rays $\mathbf{f}$, $\mathbf{g}$, respectively. This suggests the introduction of the inner product of two rays by the definition

$$\mathbf{f}\cdot\mathbf{g} = |(f, g)| \qquad (f \in \mathbf{f},\ g \in \mathbf{g}),$$

which is evidently independent of the choice of the representatives f, g.

A symmetry operation $\mathbf{T}$ of the system S maps the states in $\mathfrak{H}$ either onto themselves or onto the states in some other Hilbert space $\mathfrak{H}'$, with preservation of transition probabilities. (The second alternative corresponds to the mapping of one coherent subspace onto another. See footnote 2.) In terms of rays, $\mathbf{T}$ defines a mapping, $\mathbf{f}' = \mathbf{T}\mathbf{f}$, of unit rays onto unit rays such that $\mathbf{f}_1'\cdot\mathbf{f}_2' = \mathbf{f}_1\cdot\mathbf{f}_2$ if $\mathbf{f}_i' = \mathbf{T}\mathbf{f}_i$.

It has been shown by Wigner[3] that every such ray mapping $\mathbf{T}$ may be replaced by a vector mapping U of $\mathfrak{H}$ onto $\mathfrak{H}'$ which is either unitary or anti-unitary.[4] (For a precise formulation see Sec. 1.3 below.) For a long time this theorem has played a fundamental role in the analysis of symmetry properties of quantum systems.

The reason for returning to this question is the following. In Wigner's book the theorem is not proved in full detail. The construction of the mapping U, however, is clearly indicated, so that it is not difficult to close the gaps in the proof. In recent years several papers have appeared in which a proof of Wigner's theorem is presented.[5,6] To this writer most of these proofs seem unsatisfactory in one significant aspect: They obscure the quite elementary nature of Wigner's theorem.[7]

In addition, some authors state—or imply—the view that it is desirable, if not necessary, to depart from Wigner's construction in order to arrive at a simple or rigorous demonstration of his theorem. This writer, on the contrary, has always felt that Wigner's construction provides an excellent basis for an elementary and straightforward proof.

The present note is expository and contains no new results. It gives a complete proof of Wigner's theorem, by a method which closely adheres to his original construction. The only change of any consequence is the following. While Wigner relates U to an orthonormal set defined once for all, the proof below uses orthonormal sets adjusted to the vectors under consideration. As a result, it suffices to employ sets of at most two or three vectors.

Remarks on the notation. Re λ and Im λ denote, respectively, the real and the imaginary part of the complex number λ, and λ^* its complex conjugate.

[1] Here f corresponds of course to a wavefunction ψ. In this note vectors will be denoted by italics and scalars by lower-case Greek letters. The product of a vector f by the scalar λ will be written $f\lambda$.

[2] If superselection rules hold for S, $\mathfrak{H}$ will be considered a coherent subspace of the Hilbert space of all states. See the discussion in Wightman [A. S. Wightman, Nuovo Cimento Suppl. 14 (1959), p. 81].

[3] E. P. Wigner, *Gruppentheorie* (Frederick Vieweg und Sohn, Braunschweig, Germany, 1931), pp. 251–254; *Group Theory* (Academic Press Inc., New York, 1959), pp. 233–236.

[4] Although Wigner did not explicitly formulate his theorem in terms of rays, it is essentially equivalent to the one stated here and certainly follows from the theorem proved below.

[5] For a bibliography and a critical analysis of the proofs, see Uhlhorn, Ref. 6. The recent paper by Lomont and Mendelson [J. S. Lomont and P. Mendelson, Ann. Math. 78, 548 (1963).] should be added to his list.

[6] U. Uhlhorn, Arkiv Fysik 23, 307 (1963).

[7] This criticism does not apply at all to the very interesting papers by Emch and Piron [G. Emch and C. Piron, J. Math. Phys. 4, 469 (1963)] and by Uhlhorn[6], who start from more general premises and, consequently, obtain more comprehensive results.

1. STATEMENT OF THE THEOREM

1.1. *Preliminary remarks on rays.* Let $\mathfrak{H}$ be a complex Hilbert space—which may be finite dimensional—with vectors $f, g, \cdots$. The inner product (f, g) of two vectors f, g has Hermitian symmetry, i.e., $(g, f) = (f, g)^*$, and for any complex scalar λ

$$(f, g\lambda) = (f, g)\lambda. \tag{1}$$

$||f|| = (f, f)^{\frac{1}{2}}$ is the norm of f.

A ray $\mathbf{f}$ in $\mathfrak{H}$ is the set of all vectors $f_0\tau$, where f_0 is a fixed vector in $\mathfrak{H}$ and τ any scalar of modulus 1.[8] Every vector $f \in \mathbf{f}$ is an element or a "representative" of $\mathbf{f}$. Two vectors f', f'' are *equivalent* if they belong to the same ray, which is the case if and only if $f'' = f'\omega$ ($|\omega| = 1$). It is clear that a ray $\mathbf{f}$ is uniquely determined by any one of its representatives f, and we write

$$\mathbf{f} = \{f\}. \tag{1a}$$

$\mathbf{0}$ is the ray consisting of the vector 0.

The inner product of two rays $\mathbf{f}, \mathbf{g}$ is defined by

$$\mathbf{f}\cdot\mathbf{g} = |(f, g)| \qquad (f \in \mathbf{f}, g \in \mathbf{g}) \tag{1b}$$

and the *norm* of the ray $\mathbf{f}$ by

$$|\mathbf{f}| = (\mathbf{f}\cdot\mathbf{f})^{\frac{1}{2}} = ||f|| \qquad (f \in \mathbf{f}). \tag{1c}$$

A *unit ray* is a ray of norm 1.

For nonnegative real scalars ρ we define

$$\mathbf{f}\rho = \{f\rho\} \qquad (f \in \mathbf{f}), \tag{2}$$

i.e., if $f_0 \in \mathbf{f}$, the elements g of $\mathbf{f}\rho$ are given by $g = f_0\tau$ ($|\tau| = \rho$). Clearly,

$$(\mathbf{f}\rho)\sigma = \mathbf{f}(\rho\sigma), \qquad |\mathbf{f}\rho| = |\mathbf{f}|\,\rho, \tag{2a}$$

$$\mathbf{f}\rho\cdot\mathbf{g}\sigma = (\mathbf{f}\cdot\mathbf{g})\rho\sigma. \tag{2b}$$

Every ray $\mathbf{a}$ may be expressed in the form

$$\mathbf{a} = \mathbf{e}\rho \qquad (|\mathbf{e}| = 1, \rho \geq 0). \tag{2c}$$

In all cases $\rho = |\mathbf{a}|$. If $\mathbf{a} = \mathbf{0}$, then $\rho = 0$, and the unit ray $\mathbf{e}$ may be chosen arbitrarily. If $\mathbf{a} \neq \mathbf{0}$, $\mathbf{e}$ is uniquely determined as $\mathbf{a}\rho^{-1}$.

1.2. It is reasonable to impose the following conditions on a symmetry operation $\mathbf{T}$.

(a) $\mathbf{T}$ is defined for every unit ray $\mathbf{e}$ in $\mathfrak{H}$, and $\mathbf{e}' = \mathbf{Te}$ is a unit ray in $\mathfrak{H}'$.

(b) $\mathbf{Te}_1\cdot\mathbf{Te}_2 = \mathbf{e}_1\cdot\mathbf{e}_2$ (preservation of transition probabilities).

(c) If $\mathbf{Te}_1 = \mathbf{Te}_2$, then $\mathbf{e}_1 = \mathbf{e}_2$ (the mapping is one-to-one).

(d) Every unit ray $\mathbf{e}'$ in $\mathfrak{H}'$ is the image of some $\mathbf{e}$ in $\mathfrak{H}$ (the mapping is *onto* the unit rays in $\mathfrak{H}'$).

It is easily seen that (c) is superfluous, because it is an immediate consequence of (b). By Schwarz's inequality, two unit rays $\mathbf{f}, \mathbf{g}$ coincide if and only if $\mathbf{f}\cdot\mathbf{g} = 1$. Hence if $\mathbf{Te}_1 = \mathbf{Te}_2$, then $\mathbf{e}_1 = \mathbf{e}_2$ by (b).

In order to make the structure of the theorem more transparent we also drop the condition (d) and reinstate it in a corollary.

1.3. Thus our aim is the proof of the following

Main Theorem. Let $\mathbf{e}' = \mathbf{Te}$ *be a mapping of the unit rays* $\mathbf{e}$ *of a Hilbert space* $\mathfrak{H}$ *into the unit rays* $\mathbf{e}'$ *of a Hilbert space* $\mathfrak{H}'$ *which preserves inner products, i.e., such that*

$$\mathbf{Te}_1\cdot\mathbf{Te}_2 = \mathbf{e}_1\cdot\mathbf{e}_2. \tag{3}$$

Then there exists a mapping $a' = Ua$ *of all vectors* a *in* $\mathfrak{H}$ *into the vectors* a' *in* $\mathfrak{H}'$ *such that*

$$Ua \in \mathbf{Ta} \quad \text{if} \quad a \in \mathbf{a} \tag{4}$$

(if $\mathbf{Ta}$ *is defined) and, in addition*

$$\left.\begin{aligned} &\text{(a)} \quad U(a + b) = Ua + Ub, \\ &\text{(b)} \quad U(a\lambda) = (Ua)\chi(\lambda), \\ &\text{(c)} \quad (Ua, Ub) = \chi((a, b)), \end{aligned}\right\} \tag{5}$$

where either $\chi(\lambda) = \lambda$ *for all* λ *or* $\chi(\lambda) = \lambda^*$ *for all* λ.

A vector mapping U which satisfies (4) is called *compatible* with $\mathbf{T}$.

U is isometric since $||Ua|| = ||a||$ by (5). It is a *linear* or an *antilinear* isometry according as $\chi(\lambda) = \lambda$ or $\chi(\lambda) = \lambda^*$.

Corollary. If the $\mathbf{T}$ *of the theorem is a mapping onto all unit rays in* $\mathfrak{H}'$*—in which case we call it a "ray correspondence"—*U *is a mapping onto* $\mathfrak{H}'$. *It is unitary if* $\chi(\lambda) = \lambda$ *and anti-unitary if* $\chi(\lambda) = \lambda^*$.[9]

The corollary is an immediate consequence of the theorem.

1.4. *The one-dimensional case* is of course trivial. $\mathfrak{H}$ contains only one unit ray $\mathbf{e}$, and $\mathbf{T}$ is completely determined by $\mathbf{Te} = \mathbf{e}'$. Let $e \in \mathbf{e}$, and $e' \in \mathbf{e}'$. The two vector mappings $U_1(e\alpha) = e'\alpha$ and $U_2(e\alpha) = e'\alpha^*$ are compatible with $\mathbf{T}$. The first is linear, the second antilinear.

Hereafter we assume $\mathfrak{H}$ to be at least *two dimensional.*

1.5. It is worth mentioning that, if $\dim \mathfrak{H} \geq 2$,[10] the linear or antilinear character of U may be

[8] Many authors define a ray differently, by including *all* multiples $f_0\lambda$ of a fixed vector f_0 (irrespective of $|\lambda|$) in one ray—provided $f_0 \neq 0$ and $\lambda \neq 0$—as is suggested by projective geometry. It seems to the writer that in the present context the definition of the text is more convenient.

[9] By definition, a unitary (anti-unitary) mapping is a linear (antilinear) isometry which has an inverse.

[10] $\dim \mathfrak{H}$ denotes the dimension of $\mathfrak{H}$.

expressed in terms of $\mathbf{T}$. It describes, therefore, an *intrinsic* property of the mapping $\mathbf{T}$ and is independent of the choice of U.

Consider three rays $\mathbf{a}_j$, and let $a_j \in \mathbf{a}_j$. The expression

$$\Delta(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3) = (a_1, a_2)(a_2, a_3)(a_3, a_1)$$

is *independent* of the choice of the representatives a_j and is therefore a function of the rays $\mathbf{a}_j$. In fact, if a_j are replaced by $a_j' = a_j\tau_j$ ($|\tau_j| = 1$) the factors τ_j cancel in Δ. It follows now from (5c) that

$$\Delta(\mathbf{Te}_1, \mathbf{Te}_2, \mathbf{Te}_3) = \chi(\Delta(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)).$$

As it should be, this criterion is vacuous if $\dim \mathfrak{H} = 1$ because then $\mathbf{e}_j = \mathbf{e}$, and $\Delta = 1$. But if $\dim \mathfrak{H} \geq 2$, Δ is not always real and may serve to distinguish linear from antilinear mappings. [Let e and f be two orthogonal unit vectors in $\mathfrak{H}$, and set $e_1 = e$, $e_2 = 2^{-\frac{1}{2}}(e - f)$, $e_3 = 3^{-\frac{1}{2}}(e + f(1 - i))$. Then $||e_j|| = 1$, $\Delta = i/6$.]

1.6. A vector mapping U which transforms equivalent vectors into equivalent vectors *induces* (i.e., is compatible with) a uniquely defined *ray mapping* $\mathbf{T}$ by the equation

$$\mathbf{T}\{a\} = \{Ua\}$$

[see (1a)]. It is clear that every (linear or antilinear) isometry—in view of (5b) and (5c)—induces a ray mapping $\mathbf{T}$ which preserves inner products. Wigner's theorem asserts that no other ray mappings of this kind exist.

2. EXTENSION OF THE MAPPING T

Before constructing U we extend, following Wigner, $\mathbf{T}$ from a mapping of unit rays to a mapping of *all* rays $\mathbf{a}$ in $\mathfrak{H}$ into the rays $\mathbf{a}'$ in $\mathfrak{H}'$ by defining

$$\mathbf{T}(\mathbf{e}\rho) = (\mathbf{Te})\rho \qquad (\rho \geq 0,\ |\mathbf{e}| = 1). \tag{6}$$

Note that (6) defines $\mathbf{Ta}$ unambiguously for every ray $\mathbf{a} = \mathbf{e}\rho$. If $\mathbf{a} = \mathbf{0}$, then $\rho = 0$, hence $\mathbf{T0} = \mathbf{0}$. If $\mathbf{a} \neq \mathbf{0}$, both ρ and $\mathbf{e}$ are uniquely determined [see (2c)].

For the extended mapping we have

$$\left.\begin{aligned} &\text{(a)}\quad \mathbf{T}(\mathbf{a}\sigma) = (\mathbf{Ta})\sigma \qquad (\sigma \geq 0), \\ &\text{(b)}\quad \mathbf{Ta}_1 \cdot \mathbf{Ta}_2 = \mathbf{a}_1 \cdot \mathbf{a}_2, \qquad \text{(c)}\quad |\mathbf{Ta}| = |\mathbf{a}|. \end{aligned}\right\} \tag{7}$$

[(a) If $\mathbf{a} = \mathbf{e}\rho$, both sides of (7a) equal $(\mathbf{Te})\rho\sigma$. (b) Set $\mathbf{a}_j = \mathbf{e}_j\rho_j$. The assertion follows from (2b) and (3), and if $\mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}$, we obtain (c).]

In the sequel we deal throughout with the *extended* mapping $\mathbf{T}$—which is assumed to be given—and construct U so as to be compatible with it. Equation (4) is the *only* condition imposed on U. [(5) results from the construction.]

For later use we note the following. If, in the course of the construction, U has been defined [in accordance with (4)] for all multiples $a\lambda$ of a vector $a \neq 0$, then by (7)

$$||Ua|| = ||a||, \qquad U(a\lambda) = (Ua)\chi_a(\lambda), \tag{8}$$

$$\chi_a(1) = 1, \qquad |\chi_a(\lambda)| = |\lambda|, \tag{8a}$$

where $\chi_a(\lambda)$ is a uniquely defined function of a and λ. If U has been defined for a and b,

$$|(Ua, Ub)| = |(a, b)|. \tag{8b}$$

In Sec. 3 the mapping U is constructed for a subclass of all vectors. The partially defined U is analyzed in Sec. 4, and in Sec. 5 the construction of U—and the proof of the main theorem—is completed.

3. PARTIAL CONSTRUCTION OF U

3.1. *Preliminary remarks.* Let $\mathbf{f}_\rho$ ($\rho = 1, \cdots, m$) be m *orthonormal* rays (m finite!) so that $\mathbf{f}_\rho \cdot \mathbf{f}_\sigma = \delta_{\rho\sigma}$, and set $\mathbf{f}_\rho' = \mathbf{Tf}_\rho$. If $f_\rho \in \mathbf{f}_\rho$ and $f_\rho' \in \mathbf{f}_\rho'$, then $(f_\rho, f_\sigma) = (f_\rho', f_\sigma') = \delta_{\rho\sigma}$. Let $a = \sum_\rho f_\rho\alpha_\rho$, and $\mathbf{a} = \{a\}$. For any $a' \in \mathbf{Ta}$,

$$a' = \sum_\rho f_\rho'\alpha_\rho' \qquad |\alpha_\rho'| = |\alpha_\rho|. \tag{9}$$

Proof: Note that $||a'|| = ||a||$, $|(f_\rho', a')| = |(f_\rho, a)|$, and $(f_\rho, a) = \alpha_\rho$. Therefore

$$||a' - \sum_\rho f_\rho'(f_\rho', a')||^2 = ||a'||^2 - \sum_\rho |(f_\rho', a')|^2$$

$$= ||a||^2 - \sum_\rho |(f_\rho, a)|^2 = ||a - \sum_\rho f_\rho(f_\rho, a)||^2 = 0,$$

hence $a' = \Sigma_\rho f_\rho'(f_\rho', a')$, and the assertion follows, with $\alpha_\rho' = (f_\rho', a')$.

3.2. Fix a unit ray $\mathbf{e}$ in $\mathfrak{H}$, and let $\mathbf{e}' = \mathbf{Te}$, so that $|\mathbf{e}'| = 1$. Select $e \in \mathbf{e}$, and $e' \in \mathbf{e}'$. We define

$$\text{A.} \qquad Ue = e'$$

in accordance with (4).[11]

Denote by $\mathfrak{P}$ the set of vectors in $\mathfrak{H}$ orthogonal to e, by $\mathfrak{P}'$ the set of vectors in $\mathfrak{H}'$ orthogonal to e'.

Every vector a in $\mathfrak{H}$ has a unique decomposition

$$a = e\alpha + z \qquad (z \in \mathfrak{P}); \tag{10}$$

viz., $\alpha = (e, a)$, $z = a - e(e, a)$. In this section we construct U for those a for which $\alpha = 0$ or 1.

Let $a = e + z$ ($z \in \mathfrak{P}$, $z \neq 0$), and set $f = z/||z||$.

[11] The selection of e' constitutes the only arbitrary choice in the construction of U. All other definitions are uniquely determined by e'.

Let **a**, **f**, **z** be the corresponding rays, so that

$$\mathbf{z} = \mathbf{f}\,||z||, \qquad |\mathbf{f}| = 1.$$

If $a' \in \mathbf{Ta}$, and $f' \in \mathbf{f}' = \mathbf{Tf}$, then, by (9), $a' = e'\alpha_0' + f'\alpha_1'$, $|\alpha_0'| = 1$, $|\alpha_1'| = ||z||$.[12] Hence **Ta** contains a *uniquely determined* vector a'' $(= a'\alpha_0'^{-1})$ of the form $e' + f'\beta'$ $(|\beta'| = ||z||)$. Setting $f'\beta' = Vz$ we define[13]

B. $\quad U(e + z) = e' + Vz \qquad (z \in \mathcal{P},\ Vz \in \mathcal{P}')$.

Clearly, $Vz = f'\beta' \in (\mathbf{Tf})||z|| = \mathbf{Tz}$, and we are allowed to set

C. $\quad Uz = Vz (= U(e + z) - Ue) \qquad (z \in \mathcal{P})$.

If $z = 0$, we set $Vz = 0$, so that B reduces to A.

In the next section we analyze the mapping V of $\mathcal{P}$ into $\mathcal{P}'$ in greater detail.

4. ANALYSIS OF THE MAPPING V

4.1. *The real part of* (Vw, Vx). Let w, x be in $\mathcal{P}$. By C, B, and (8b),

$$|(Vw, Vx)|^2 = |(w, x)|^2, \tag{11}$$

and $|(e' + Vw, e' + Vx)|^2 = |(e + w, e + x)|^2$, or $|1 + (Vw, Vx)|^2 = |1 + (w, x)|^2$. Since for every complex number ζ, $|1 + \zeta|^2 = 1 + |\zeta|^2 + 2\,\mathrm{Re}\,\zeta$, it follows from (11) that

$$\mathrm{Re}\,(Vw, Vx) = \mathrm{Re}\,(w, x), \tag{12}$$

$$(Vw, Vx) = (w, x) \quad \text{if} \quad (w, x) \quad \text{is real}. \tag{12a}$$

4.2. It will now be shown—in the remainder of this section—that, for any two nonvanishing y, z in $\mathcal{P}$, (a) $V(y + z) = Vy + Vz$, (b) $\chi_w(\lambda) = \chi_z(\lambda)$ [see (8)], (c) $(Vy, Vz) = \chi_w((y, z))$.

4.3 Set $f_1 = y/||y||$. If $\dim \mathfrak{C} = 2$ all vectors in $\mathcal{P}$ are multiples of f_1, hence

$$y = f_1\rho, \qquad z = f_1\sigma. \tag{13a}$$

If $\dim \mathfrak{C} \geq 3$ choose a second unit vector f_2 in $\mathcal{P}$ *orthogonal* to f_1 (whether or not y and z are independent) such that

$$y = f_1\rho, \qquad z = f_1\sigma + f_2\tau. \tag{13b}$$

In both cases let $\mathfrak{L}$ be the set of linear combinations of the m orthonormal vectors f_ρ $(m = 1 \text{ or } m = 2)$.

4.4. *The functions* $\chi_\rho(\alpha)$. Set $\mathbf{f}_\rho = \{f_\rho\}$. Then $f_\rho' = Vf_\rho \in \mathbf{Tf}_\rho$, and the vectors f_ρ' are *orthonormal*. By (8),

$$V(f_\rho\alpha) = f_\rho'\chi_\rho(\alpha), \qquad |\chi_\rho(\alpha)| = |\alpha|. \tag{14}$$

Applying (12) to $f_\rho\alpha$ and $f_\rho\beta$ we obtain

$$\mathrm{Re}\,(\chi_\rho(\alpha)^*\chi_\rho(\beta)) = \mathrm{Re}\,(\alpha^*\beta). \tag{15}$$

Set $\alpha = 1$. Since $\chi_\rho(1) = 1$ we conclude from (12) and (12a) that

$$\mathrm{Re}\,\chi_\rho(\beta) = \mathrm{Re}\,\beta, \tag{15a}$$

$$\chi_\rho(\beta) = \beta \quad \text{for} \quad \textit{real}\ \beta. \tag{15b}$$

4.5.[14] Let $x = \sum_\rho f_\rho\alpha_\rho$. By (9), $Vx = \sum_\rho f_\rho'\alpha_\rho'$, $|\alpha_\rho'| = |\alpha_\rho|$. We prove first that $\alpha_\rho' = \chi_\rho(\alpha_\rho)$. This is trivial if $\alpha_\rho = 0$. If $\alpha_\rho \neq 0$, set $\gamma_\rho = \alpha_\rho^{*-1}$. Then $(f_\rho\gamma_\rho,\ f_\rho\alpha_\rho) = (f_\rho\gamma_\rho,\ x) = 1$. Hence, by (12a), $\chi_\rho(\gamma_\rho)^*\chi_\rho(\alpha_\rho) = \chi_\rho(\gamma_\rho)^*\alpha_\rho' = 1$, i.e., $\alpha_\rho' = \chi_\rho(\alpha_\rho)$.

We show next that $\chi_2(\alpha) = \chi_1(\alpha)$ if $m = 2$. Let $w = \sum_\rho f_\rho$. Then $Vw = \sum_\rho f_\rho'$, and $V(w\alpha) = \sum_\rho f_\rho'\chi_\rho(\alpha) = (Vw)\chi_w(\alpha)$, by (8). Thus $\chi_1(\alpha) = \chi_2(\alpha) = \chi_w(\alpha)$. As a result,

$$V(\sum_\rho f_\rho\alpha_\rho) = \sum_\rho f_\rho'\chi_1(\alpha_\rho). \tag{16}$$

4.6. *Determination of* $\chi_1(\beta)$. (1) Set $\beta = i$. Then $|\chi_1(i)| = 1$, $\mathrm{Re}\,\chi_1(i) = 0$; thus $\chi_1(i) = \eta i$, $\eta = 1$ *or* $\eta = -1$. (2) For any complex ζ, $\mathrm{Im}\,\zeta = \mathrm{Re}\,(i^*\zeta)$. Hence, from (15), $\mathrm{Im}\,\chi_1(\beta) = \mathrm{Re}\,(i^*\chi_1(\beta)) = \eta\,\mathrm{Re}\,(\chi_1(i)^*\chi_1(\beta)) = \eta\,\mathrm{Re}\,(i^*\beta) = \eta\,\mathrm{Im}\,\beta$. Combining this with (15a),

$$\chi_1(\beta) = \beta \text{ if } \eta = 1, \quad \chi_1(\beta) = \beta^* \text{ if } \eta = -1. \tag{17}$$

Note the obvious relations:

(a) $\chi_1(\alpha + \beta) = \chi_1(\alpha) + \chi_1(\beta)$,

(b) $\chi_1(\alpha\beta) = \chi_1(\alpha)\chi_1(\beta)$,

(c) $\chi_1(\alpha)^* = \chi_1(\alpha^*)$.

4.7. *The structure of* V. Let $w = \sum_\rho f_\rho\alpha_\rho$ and $x = \sum_\rho f_\rho\beta_\rho$ be two vectors in $\mathfrak{L}$. From (16) and from the properties of χ_1 just stated we draw the following conclusions. By (a), $V(w + x) = Vw + Vx$, by (b), $V(x\lambda) = (Vx)\chi_1(\lambda)$. Since both y and z belong to $\mathfrak{L}$ this proves the assertions (a) and (b) of Sec. 4.2, with $\chi_w(\lambda) = \chi_z(\lambda) = \chi_1(\lambda)$. To establish (c) in (4.2) we note that $(y, z) = \rho^*\sigma$ [by (13)], and $(Vy, Vz) = \chi_1(\rho)^*\chi_1(\sigma) = \chi_1(\rho^*)\chi_1(\sigma) = \chi_1(\rho^*\sigma)$, Q.E.D.

By (b) in Sec. 4.2, $\chi_z(\lambda)$ is the *same* function, say, $\chi(\lambda)$, for every nonvanishing vector z in $\mathcal{P}$. To sum up, the mapping V has the following properties:

$$\begin{aligned} &\text{(a)} \quad V(y + z) = Vy + Vz, \\ &\text{(b)} \quad V(z\lambda) = (Vz)\chi(\lambda), \\ &\text{(c)} \quad (Vy, Vz) = \chi((y, z)), \end{aligned} \tag{18}$$

[12] e and f are orthonormal, so are e' and f'.

[13] The definitions A and B are the crucial steps in Wigner's construction.

[14] If $m = 1$, Sec. 4.5 may be omitted since Eq. (16) reduces then to Eq. (14).

where χ is one of the functions in (17). [The equations (18) have been explicitly proved for nonvanishing y and z. But they hold trivially if $y = 0$ or $z = 0$.]

5. THE CONSTRUCTION OF U COMPLETED

It remains to define U for vectors $a = e\alpha + z$ ($z \in \mathfrak{P}$) for which $\alpha \neq 0, 1$ [see (10)]. Set $b = e + z\alpha^{-1}$, so that $a = b\alpha$, and $\mathbf{Ta} = (\mathbf{Tb})\,|\alpha|$ if $\mathbf{a}, \mathbf{b}$ are the corresponding rays. Ub ($\in \mathbf{Tb}$) is defined by B in Sec. 3. Hence $(Ub)\chi(\alpha) \in \mathbf{Ta}$, and we may therefore define $Ua = (e' + V(z\alpha^{-1}))\chi(\alpha)$, or, by (18),

D. $\qquad U(e\alpha + z) = e'\chi(\alpha) + Vz \qquad (z \in \mathfrak{P})$.

If $\alpha = 1$ or 0, D coincides with A, B, or C of Sec. 3. Thus it defines the mapping U for all vectors a in $\mathfrak{H}$.

By virtue of (18) it is an immediate consequence of D that U satisfies *all conditions* (5) of the theorem.

[For an example, let $a_i = e\alpha_i + z_i$ ($j = 1, 2$). Then $(a_1, a_2) = \alpha_1^*\alpha_2 + (z_1, z_2)$, and $(Ua_1, Ua_2) = \chi(\alpha_1)^*\chi(\alpha_2) + Vz_1, Vz_2) = \chi(\alpha_1^*\alpha_2) + \chi((z_1, z_2)) = \chi(\alpha_1^*\alpha_2 + (z_1, z_2)) = \chi((a_1, a_2))$.]

This concludes the proof of the main theorem.

6. UNIQUENESS OF U

It is of course important to know to what extent U is determined by a ray mapping $\mathbf{T}$ which preserves inner products. Without using the main theorem in this section the following can be asserted. ($\mathbf{T}$ stands for the *extended* mapping)

(a) Let U be a vector mapping compatible with $\mathbf{T}$. If a_1, a_2 are (linearly) independent, so are Ua_1, Ua_2. *Proof:* Two vectors a_i are independent if and only if $G(a_1, a_2) = (a_1, a_1)(a_2, a_2) - |(a_1, a_2)|^2 > 0$. Since, by (8b), G is not changed by the mapping U, the assertion follows.

(b) If U_2 and U_1 are compatible with the same $\mathbf{T}$, then $U_2 0 = U_1 0 = 0$, and for every $a \neq 0$

$$U_2 a = (U_1 a)\tau(a), \qquad |\tau(a)| = 1.$$

If $\tau(a) = \theta$ (independent of a), we write $U_2 = U_1\theta$.

A mapping U is *additive* if $U(a + b) = Ua + Ub$.

Theorem 2. If two additive vector mappings U_2 and U_1 are compatible with the same $\mathbf{T}$, *and* dim $\mathfrak{H} \geq 2$, *then $U_2 = U_1\theta$.*[15] ($|\theta| = 1$.)

Proof: We proceed in two steps. (1) *If a, b are independent, $\tau(a) = \tau(b)$.* Set $c = a + b$. Then $U_2 c = U_2 a + U_2 b$, $U_1 c = U_1 a + U_1 b$, and $U_2 c = (U_1 c)\tau(c)$. Therefore

$$(U_1 a)\tau(a) + (U_1 b)\tau(b) = (U_1 a + U_1 b)\tau(c).$$

Since $U_1 a$ and $U_1 b$ are independent, $\tau(c) = \tau(a) = \tau(b)$.

(2) Fix a vector $a_0 \neq 0$ in $\mathfrak{H}$, and set $\tau(a_0) = \theta$. *For every vector $a \neq 0$, $\tau(a) = \theta$.* If a and a_0 are independent, this follows from (1). If $a = a_0\mu$ ($\mu \neq 0$), choose b independent of a_0 (and hence of a). Then, by (1), $\tau(b) = \tau(a)$, $\tau(b) = \theta$, Q.E.D.

Let $U_2 = U_1\theta$. *If* $U_1(a\lambda) = (U_1 a)\chi(\lambda)$, *then also* $U_2(a\lambda) = (U_2 a)\chi(\lambda)$. In fact,

$$U_2(a\lambda) = (U_1 a)\chi(\lambda)\theta = (U_1 a)\theta\chi(\lambda) = (U_2 a)\chi(\lambda) \qquad (19)$$

in accordance with our result in Sec. 1.5.

APPENDIX. WIGNER'S THEOREM IN QUATERNION QUANTUM THEORY

In recent years there has been some interest in a modification of the quantum theoretical formalism which consists in replacing the complex Hilbert space of quantum states by a quaternion Hilbert space.[16] We wish to indicate the changes in the theorem and in its proof that must be made. The above exposition is so arranged that these changes are concentrated in a few places.

1. *Preliminary remarks on quaternions.*[17] Let Q be the set of all quaternions. We write a quaternion λ in the form $\lambda = \sum_{\nu=0}^{3} \lambda_\nu i_\nu$. Here, λ_ν are real numbers, $i_0 = 1$ (i.e., $i_0\lambda = \lambda i_0 = \lambda$ for every $\lambda \in Q$), while i_r ($r = 1, 2, 3$) are the imaginary units, with the multiplication rules

$$i_r^2 = -1, \qquad i_r i_s = -i_s i_r = i_t \qquad (A1)$$

where r, s, t is an *even* permutation of 1, 2, 3.

The conjugate λ^* of λ is defined by

$$\lambda^* = \lambda_0 i_0 - \sum_{r=1}^{3} \lambda_r i_r$$

so that $(\lambda^*)^* = \lambda$. A quaternion is *real* if $\lambda^* = \lambda$, i.e., $\lambda = \lambda_0 \cdot 1 = \lambda_0$. (The real quaternions, and only they, *commute* with all of Q.) In general we denote the real part of λ by

$$\mathrm{Re}\ \lambda = \tfrac{1}{2}(\lambda + \lambda^*) = \lambda_0$$

Note that $\mathrm{Re}\ \lambda^* = \mathrm{Re}\ \lambda$. Let $\kappa = \sum_\nu \kappa_\nu i_\nu$. Since $(i_\mu i_\nu)^* = i_\nu^* i_\mu^*$ (for all μ, ν), $(\kappa\lambda)^* = \lambda^*\kappa^*$.

For all μ, ν

$$\mathrm{Re}\ (i_\mu^* i_\nu) = \mathrm{Re}\ (i_\mu i_\nu^*) = \delta_{\mu\nu}. \qquad (A2)$$

It follows that

$$\mathrm{Re}\ (\kappa^*\lambda) = \mathrm{Re}\ (\lambda^*\kappa) = \sum_\nu \kappa_\nu\lambda_\nu, \qquad (A2a)$$

$$\mathrm{Re}\ (\kappa^*\lambda) = \mathrm{Re}\ (\kappa\lambda^*). \qquad (A2b)$$

[15] In terms of the construction in Sec. 3.2 this merely means that A is replaced by $U_2 e = e'\theta$ (see footnote 11). It follows from Sec. 1.4 that dim $\mathfrak{H} \geq 2$ is a necessary assumption.

[16] D. Finkelstein, J. M. Jauch, S. Schiminovich, and D. Speiser, J. Math. Phys. 3, 207 (1961).

[17] The reader is assumed to be familiar with quaternions. These introductory remarks fix the notation and review several relations to be used later on.

In particular, $\lambda\lambda^*$ and $\lambda^*\lambda$ being real,

$$\lambda\lambda^* = \lambda^*\lambda = \sum_\nu \lambda_\nu^2 = |\lambda|^2,$$

where $|\lambda|$ is the modulus of λ. We have $|\kappa\lambda| = |\kappa|\,|\lambda|$ since $|\kappa\lambda|^2 = \kappa(\lambda\lambda^*)\kappa^* = (\kappa\kappa^*)(\lambda\lambda^*) = |\kappa|^2\,|\lambda|^2$. If $\lambda \neq 0$, $\lambda^{-1} = |\lambda|^{-2}\lambda^*$, $\lambda\lambda^{-1} = \lambda^{-1}\lambda = 1$.

Q may be considered a four-dimensional *real vector space*. (A2a) defines then an inner product in Q, and by (A2) the units i_ν form an *orthonormal basis*. Hence

$$\lambda_\nu = \mathrm{Re}\,(i_\nu^*\lambda). \tag{A3}$$

More generally, let j_ν ($\nu = 0, 1, 2, 3$) be four orthonormal quaternions, so that $\mathrm{Re}\,(j_\mu^* j_\nu) = \delta_{\mu\nu}$. Then every κ may be written as

$$\kappa = \sum_\nu \mathrm{Re}\,(j_\nu^*\kappa) j_\nu. \tag{A3a}$$

For a later application we add a few words about the automorphism

$$\lambda' = \sigma_\gamma(\lambda) = \gamma\lambda\gamma^{-1} = \gamma\lambda\gamma^* \tag{A4}$$

where γ is a fixed quaternion of modulus 1. Clearly

(a) $\sigma_\gamma(\kappa + \lambda) = \sigma_\gamma(\kappa) + \sigma_\gamma(\lambda)$,

(b) $\sigma_\gamma(\kappa\lambda) = \sigma_\gamma(\kappa)\sigma_\gamma(\lambda)$,

(c) $\sigma_\gamma(\lambda^*) = \sigma_\gamma(\lambda)^*$.

In addition,

(d) $\mathrm{Re}\,\sigma_\gamma(\lambda) = \mathrm{Re}\,\lambda$,

(e) $\mathrm{Re}\,(\sigma_\gamma(\kappa)^*\sigma_\gamma(\lambda)) = \mathrm{Re}\,(\kappa^*\lambda)$.

[(d) follows from $\mathrm{Re}\,((\gamma\lambda)\gamma^*) = \mathrm{Re}\,(\gamma^*\gamma\lambda) = \mathrm{Re}\,\lambda$, and this in turn implies (e) because $\sigma_\gamma(\kappa)^*\sigma_\gamma(\lambda) = \sigma_\gamma(\kappa^*\lambda)$.] By (e), σ_γ is an orthogonal transformation on Q. Conjugation is also an orthogonal mapping, by (A2a). Combining it with σ_γ one obtains a second type of orthogonal transformation,

$$\lambda' = \sigma_\gamma(\lambda^*). \tag{A4a}$$

2. *Wigner's theorem.* The states of the system S are again put in a one-one correspondence with the unit rays in the Hilbert space $\mathfrak{H}$. The discussion in Sec. 1.1 remains unchanged except that the scalars λ and the values of the inner products (f, g)—in $\mathfrak{H}$ and $\mathfrak{H}'$—are now quaternions. More generally, the whole content of Secs. 1–6 remains applicable with the exception of those instances where (a) specific properties of complex numbers are used or (b) the commutative law of multiplication is applied.

The only instance of type (a) is the determination of χ_1 in Sec. 4.6 and, consequently, the characterization of $\chi(\lambda)$ in the statement of the main theorem. The only instance of Type (b) is Eq. (19) in Sec. 6.[18]

These two points are now re-examined.

3. *The two-dimensional case.* In the quaternion case Wigner's theorem *no longer holds* if $\dim \mathfrak{H} = 2$.[19]

Every vector z in $\mathcal{P}$ has now the form $f_1\alpha$, and $V(f_1\alpha) = f_1'\chi_1(\alpha)$ [see (14)] where χ_1 satisfies the three equations (15), (15a), and (15b). From (A4) and (A4a) we obtain two types of solutions, viz.,

$$(1)\quad \chi_1(\alpha) = \sigma_\gamma(\alpha), \qquad (2)\quad \chi_1(\alpha) = \sigma_\gamma(\alpha^*). \tag{A5}$$

(It is not difficult to show that no other solutions exist.)

Now the proof of the main theorem in Sec. 5 is based, in part, on Eq. (18b), whose derivation in turn depends on the relation (b) at the end of Sec. 4.6, namely, $\chi_1(\alpha\beta) = \chi_1(\alpha)\chi_1(\beta)$. While the first solution in (A5) satisfies this relation we find for the second $\chi_1(\alpha\beta) = \sigma_\gamma((\alpha\beta)^*) = \sigma_\gamma(\beta^*\alpha^*) = \sigma_\gamma(\beta^*)\sigma_\gamma(\alpha^*) = \chi_1(\beta)\chi_1(\alpha)$. The order of the factors is reversed: χ_1 is an *antiautomorphism*.

Choose, for simplicity, $\gamma = 1$, so that $V(f_1\beta) = f_1'\beta^*$. As the arguments of Sec. 5 show we may set $U(e\alpha + f_1\beta) = (e' + f_1'(\beta\alpha^{-1})^*)\alpha$ if $\alpha \neq 0$. For convenience we define a new mapping U_0 compatible with $\mathbf{T}$ by

$$U_0(e\alpha + f_1\beta) = (e' + f_1'(\beta\alpha^{-1})^*)\alpha \quad (\alpha \neq 0),$$

$$U_0(f_1\beta) = f_1'\beta \quad (\alpha = 0).$$

(U_0 differs from U only if $\alpha = 0$.)

To disprove Wigner's theorem in this case it remains to show (1) that U_0 actually induces a ray mapping $\mathbf{T}$ with the required properties [i.e., that no conditions have been overlooked that might rule out the second solution in (A5)]; (2) that no additive vector mapping is compatible with $\mathbf{T}$.

(1) By straightforward computation one verifies that, for every vector $a = e\alpha + f_1\beta$, $U_0(a\lambda) = (U_0 a)\lambda$, and that $|(U_0 a_1, U_0 a_2)| = |(a_1, a_2)|$. Thus U_0 induces indeed a ray mapping $\mathbf{T}$ which preserves inner products.

(2) will be proved by contradiction. Let W be an additive vector mapping compatible with $\mathbf{T}$. Then

$$W(e\alpha + f_1\beta) = (U_0(e\alpha + f_1\beta))\Phi(\alpha, \beta) \quad |\Phi(\alpha, \beta)| = 1$$

if $(\alpha, \beta) \neq (0, 0)$. In particular, $W(e\alpha) = e'\alpha\Phi(\alpha, 0)$, $W(f_1\beta) = f_1'\beta\Phi(0, \beta)$. Setting $\eta(\alpha) = \alpha\Phi(\alpha, 0)$ and

[18] It should be added that the remarks in Sec. 1.5 do not apply to the quaternion case. The proof that Δ does not depend on the choice of the representatives a_i uses commutativity of multiplication.

[19] Uhlhorn's contrary assertion (Ref. 6, pp. 335, 336) is incorrect. He overlooked the second solution in (A5).

$\zeta(\beta) = \beta\Phi(0, \beta)$ we conclude from the additivity of W that

$$e'\eta(\alpha) + f_1'\zeta(\beta) = (e' + f_1'\alpha^{*-1}\beta^*)\alpha\Phi(\alpha, \beta).$$

Assume $\alpha \neq 0$, $\beta \neq 0$. Then η, ζ and $\Phi \neq 0$, and

$$\alpha^*\zeta(\beta) = \beta^*\eta(\alpha). \tag{A6}$$

Setting, in succession, $\alpha = \beta = 1$, $\beta = 1$, and $\alpha = 1$, one finds $\zeta(1) = \eta(1)$, $\eta(\alpha) = \alpha^*\eta(1)$, $\zeta(\beta) = \beta^*\eta(1)$. Multiplying (A6) with $\eta(1)^{-1}$ from the right, we finally obtain $\alpha^*\beta^* = \beta^*\alpha^*$ or $\beta\alpha = \alpha\beta$, which is absurd.

4. *Determination of* χ_1 *if dim* $\mathfrak{H} \geq 3$. Here $m = 2$, and we first derive a further condition on χ_1 to supplement Eqs. (15), (15a), (15b). Let $w = f_1 + f_2\alpha$. Then $Vw = f_1' + f_2'\chi_1(\alpha)$, and $V(w\beta) = f_1'\chi_1(\beta) + f_2'\chi_1(\alpha\beta) = (Vw)\chi_w(\beta)$, so that $\chi_1(\beta) = \chi_w(\beta)$, $\chi_1(\alpha\beta) = \chi_1(\alpha)\chi_w(\beta)$. Thus

$$\chi_1(\alpha\beta) = \chi_1(\alpha)\chi_1(\beta). \tag{A7}$$

Set now $j_\nu = \chi_1(i_\nu)$. By (15b), $j_0 = 1$, and by (15) and (A2)

$$\mathrm{Re}\,(j_\mu^* j_\nu) = \mathrm{Re}\,(i_\mu^* i_\nu) = \delta_{\mu\nu}.$$

Let $\beta = \sum_{\nu=0}^{3} \beta_\nu i_\nu$. By (15), $\mathrm{Re}\,(j_\nu^*\chi_1(\beta)) = \mathrm{Re}\,(i_\nu^*\beta) = \beta_\nu$, [see (A3)], so that, by (A3a)

$$\chi_1(\beta) = \sum_\nu \beta_\nu j_\nu. \tag{A8}$$

By (15b), $\chi_1(-1) = -1$; hence by (A7), $\chi_1(-\beta) = -\chi_1(\beta)$. Applying (A7) to j_r $(r > 0)$ we find

$$\begin{aligned} j_r^2 &= \chi_1(i_r^2) = -1, \\ j_r j_s &= \chi_1(i_r i_s) = \chi_1(i_t) = j_t, \\ j_s j_r &= \chi_1(-i_t) = -j_t, \end{aligned}$$

if (r, s, t) is an even permutation of $(1, 2, 3)$. Together with $j_0 = 1$, this shows that the j_ν satisfy the multiplication rules of the units i_ν.

It is well known—and easily proved—that then $j_\nu = \gamma i_\nu \gamma^{-1} = \sigma_\gamma(i_\nu)$ for some fixed γ of modulus 1. Inserting this in (A8) we finally obtain

$$\chi_1(\beta) = \sigma_\gamma(\beta). \tag{A9}$$

This solution satisfies all three relations listed at the end of Sec. 4.6, and the arguments in Sec. 4.7 and Sec. 5 apply without change. Thus the main theorem is valid, but $\chi(\lambda)$ is an automorphism $\sigma_\gamma(\lambda)$, and U a *semilinear* isometry.

5. *Theorem 2 of Sec. 6 holds.* The transition from U_1 to $U_2 = U_1\theta$, however, has now more radical consequences. Instead of Eq. (19) we have

$$U_2(a\lambda) = U_1(a)\chi(\lambda)\theta = (U_1 a)\theta(\theta^{-1}\chi(\lambda)\theta).$$

Assuming $\chi(\lambda) = \sigma_\gamma(\lambda)$,

$$\begin{aligned} U_2(a\lambda) &= (U_2 a)\chi'(\lambda), \\ \chi'(\lambda) &= \theta^{-1}\sigma_\gamma(\lambda)\theta = \sigma_{\theta^{-1}\gamma}(\lambda). \end{aligned} \tag{A10}$$

$\chi'(\lambda) = \chi(\lambda)$ *if and only if* $U_2 = \pm U_1$. In fact, θ must commute with all $\sigma_\gamma(\lambda)$ and hence must be *real*. Since $|\theta| = 1$, $\theta = \pm 1$.

In particular, if $\theta = \gamma$, then $\chi'(\lambda) = \lambda$, so that U_2 is *linear*.

To sum up: *In the quaternion case, if dim* $\mathfrak{H} \geq 3$, *every ray mapping* **T** *which preserves inner products is induced by a linear mapping* U, *and* **T** *determines* U *up to a sign.*

6. *Remarks on Uhlhorn's theorem.* The following remarks apply to the complex as well as the quaternion case. Uhlhorn has obtained the very interesting result that Wigner's theorem holds under considerably weaker assumptions. In terms of the conditions listed in Sec. 1.2 it suffices to maintain (a) and (d) while (b) is replaced by the condition b': $\mathbf{Te}_1 \cdot \mathbf{Te}_2 = 0$ if and only if $\mathbf{e}_1 \cdot \mathbf{e}_2 = 0$ (preservation only of the transition probability zero!) On the other hand it is necessary to assume dim $\mathfrak{H} \geq 3$.

Since, however, the condition (d)—or possibly some weaker substitute—is actually needed for the proof of Uhlhorn's result it seems to the writer that the main theorem proved in the present note retains an independent mathematical interest.

In conclusion it may be mentioned that a minor modification of Wigner's construction also yields a simple proof of Uhlhorn's theorem.

Zero-Mass Representations of the Proper Inhomogeneous Lorentz Group*

DAVID KORFF
Department of Physics and Astronomy, University of Maryland, College Park, Maryland
(Received 14 January 1963)

The representations of the proper inhomogeneous Lorentz group are investigated as a function of both real and imaginary mass, in the limit as the mass approaches zero. One obtains only the physical mass-zero representations in either limit, the infinite spin representations being unique to zero mass. It is found that there exists a superselection rule which prohibits the position and spin operators from being physically observable in the mass-zero limit.

1. INTRODUCTION

THE classification of the unitary irreducible representations of the proper inhomogeneous Lorentz group usually differentiates between three distinct classes.[1-5] These classes, corresponding to a particle with a squared mass greater than, less than, or equal to zero, are qualitatively very different. However, one might expect on physical grounds that a particle of vanishingly small mass, or vanishingly small imaginary mass, should behave like, *and be indistinguishable from*, a zero-mass particle. The "physical" mass-zero representations (i.e., non-infinite spin) have been obtained by Coester[6] as a limit of those representations characterized by a finite (real) mass. The purposes of this paper are twofold: First, to expand on Coester's results for the zero-mass limit, with particular attention to the role of the position and spin operators. Second, to obtain the "physical" zero-mass representations as a limit of those characterized by a finite imaginary mass (i.e., negative squared mass). We set $\hbar = c = 1$ throughout.

2. THE REPRESENTATIONS OF THE PROPER INHOMOGENEOUS LORENTZ GROUP[7]

As is well known, the generators of the proper inhomogeneous Lorentz group are ten in number corresponding to the linear four-momentum (translations) and the four-angular momentum (rotations and pure Lorentz transformations). These satisfy the commutation rules

$$[p_\mu, p_\nu] = 0, \quad [M_{\mu\nu}, p_\sigma] = i[g_{\nu\sigma}p_\mu - g_{\mu\sigma}p_\nu],$$
$$[M_{\mu\nu}, M_{\rho\sigma}] = -i(g_{\mu\rho}M_{\nu\sigma} - g_{\nu\rho}M_{\mu\sigma} + g_{\mu\sigma}M_{\rho\nu} - g_{\nu\sigma}M_{\rho\mu}), \tag{2.1}$$
$$g_{00} = -g_{11} = -g_{22} = -g_{33} = 1.$$

It is convenient to define the pseudovector $w_\mu = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}M^{\nu\rho}p^\sigma$ which is orthogonal to p_μ, $p_\mu w^\mu = 0$, and which has the following commutation rules with the generators:

$$[M_{\mu\nu}, w_\rho] = i(g_{\nu\rho}w_\mu - g_{\mu\rho}w_\nu); \quad [w_\mu, p_\nu] = 0;$$
$$[w_\mu, w_\nu] = i\epsilon_{\mu\nu\rho\sigma}w^\rho p^\sigma. \tag{2.2}$$

It then follows that $p_\mu p^\mu \equiv P$ and $w_\mu w^\mu \equiv -W$ commute with the ten generators and hence, via Schur's Lemma, are multiples of the identity matrix in any irreducible representation. The eigenvalue of P is the mass squared and the eigenvalue of $P^{-1}W$ is related to the spin, s, by $P^{-1}W = s\,(s+1)$. In addition, there may be other invariants for particular classes of representations. The irreducible unitary representations of the proper inhomogeneous Lorentz group may be classified as follows:

(A) $P > 0$. The relation $p_\mu w^\mu = 0$ requires that $W > 0$. It follows from the commutation rules that the spin is an integer (half-integer) for single (double)-valued representations. In addition, $p_0/|p_0|$, the sign of the energy, commutes with the ten generators and hence there are two irreducible representations for each value of P and W—one for each sign of the energy. A basis which spans the representation space may be taken to be the eigenstates of the four-momentum and helicity $(w_0/|\mathbf{p}|)$. The spectrum is given by $(2s+1)$ eigenstates of helicity for each value of the four-momentum, going

* This research was supported by the National Science Foundation under Grant GP 1193.
[1] E. P. Wigner, Ann. Math. 40, 149 (1939).
[2] V. Bargmann and E. P. Wigner, Proc. Natl. Acad. Sci. U. S. 34, 211 (1948).
[3] Yu. M. Shirokhov, Zh. Eksperim. i Teor. Fiz. 33, 861 (1957) [English transl.: Soviet Phys.—JETP 6, 664 (1958)].
[4] Yu. M. Shirokhov, Zh. Eksperim. i Teor. Fiz. 33, 1196 (1957) [English transl.: Soviet Phys.—JETP 6, 919 (1958)].
[5] Yu. M. Shirokhov, Zh. Eksperim. i Teor. Fiz. 33, 1208 (1957) [English transl.: Soviet Phys.—JETP 6, 929 (1958)].
[6] F. Coester, Phys. Rev. 129, 2816 (1963).
[7] Sections 2 and 3 are a summary of well known results, which we have included for purposes of self-containment. Section 2 leans heavily on Ref. 4. The results of Sec. 3 have previously been derived by Coester (Ref. 6). Our transformation matrix (3.1) differs from his by a similarity transformation.

from $-s$ to $+s$ in integer steps. An explicit representation of the generators for, e.g., $p_0 > 0$, in which the z component of spin is diagonal, rather than the helicity, has been given by Shirokhov[4]:

$$(\mathbf{p})_{op} = \mathbf{p}, \qquad (p_0)_{op} = (\mathbf{p}^2 + m^2)^{\frac{1}{2}}, \tag{2.3}$$

$$\mathbf{M} = -i(\mathbf{p} \times \nabla_p) + \mathbf{S},$$

$$\mathbf{N} = ip_0\nabla_p - \mathbf{S} \times \mathbf{p}/(p_0 + m),$$

$$\mathbf{M} \equiv (M_{23}, M_{31}, M_{12}), \quad \mathbf{N} \equiv (M_{01}, M_{02}, M_{03}).$$

$S = (S_1, S_2, S_3)$ denote a set of constant (i.e., momentum-independent), $(2s + 1) \times (2s + 1)$-dimensional matrices, satisfying $[S_i, S_j] = \epsilon_{ijk}S_k$. The usual convention, which we shall adopt, is to take S_3 diagonal. The unitary irreducible representations of Class A are denoted in the literature by $P^s_{\pm m}$, the sign referring to the sign of the energy. We shall have occasion to deal only with unitary representations in this paper. The wavefunctions which constitute a representation space for $P^s_{\pm m}$ have $(2s + 1)$ components.

(B) $P = 0$. The relation $p_\mu w^\mu = 0$ requires that either $W > 0$ or else $w_\mu = cp_\mu$, where c is a constant. The spin is undefined in both cases. Both $p_0/|p_0|$ and $w_0/|\mathbf{p}|$ commute with the ten generators if $W=0$. If $W > 0$, we have the "infinite spin" representations considered by Bargmann and Wigner.[2] We are not interested in these representations for reasons which are made clearer below (Sec. 5). For $W = 0$, the irreducible representations contain the helicity as a diagonal operator and the wavefunctions are all one component. The helicity operator has eigenvalues $\Sigma = w_0/|\mathbf{p}| = \pm w_0/p_0$, where Σ is an integer (half-integer) for single (double)-valued representations. An explicit representation of the generators for, e.g., $p_0 > 0$ and $w_0/|\mathbf{p}| = \Sigma$ has been given by Shirokhov[5,8,9]:

$$(\mathbf{p})_{op} = \mathbf{p}, \qquad (p_0)_{op} = |\mathbf{p}|, \tag{2.4}$$

$$M_1 = -i(\mathbf{p} \times \nabla_p)_1 + [\Sigma p_1 |\mathbf{p}|/(p_1^2 + p_2^2)](1 - p_3/|\mathbf{p}|),$$

$$M^2 = -i(\mathbf{p} \times \nabla_p)_2 + [\Sigma p_2 |\mathbf{p}|/(p_1^2 + p_2^2)](1 - p_3/|\mathbf{p}|),$$

$$M_3 = -i(\mathbf{p} \times \nabla_p)_3 + \Sigma,$$

$$N_1 = ip_0\partial/\partial p_1 + [p_0p_2/(p_1^2 + p_2^2)](1 - p_3/|\mathbf{p}|)\Sigma,$$

$$N_2 = ip_0\partial/\partial p_2 - [p_0p_1/(p_1^2 + p_2^2)](1 - p_3/|\mathbf{p}|)\Sigma,$$

$$N_3 = ip_0\partial/\partial p_3.$$

[8] Strictly speaking, the representation (2.4) differs from Shirokhov's by a unitary equivalence. In addition, some errors of sign in Ref. 5 have been corrected.

[9] Yu. M. Shirokhov, Nucl. Phys. 15, 1 (1960).

The Class B representations are denoted in the literature by $P^{\Sigma}_{\pm 0}$. Once again, the sign refers to the sign of the energy.

(C) $P < 0$. The relation $p_\mu w^\mu = 0$ imposes no restriction on W. If $P^{-1}W \geq 0(-\frac{1}{4})$ for single (double)-valued representations, then $w_0/|w_0|$, the sign of the helicity, commutes with the ten generators and is a multiple of the identity in an irreducible representation. In this case $P^{-1}W = s(s + 1)$ where s is an integer (half-integer) greater than, or equal to, zero (minus one-half) for single (double)-valued representations. Depending on the sign of the helicity in the irreducible representation, the helicity runs from $\pm (s + 1)$ to $\pm\infty$ in integer steps. We may therefore say, after a fashion, that the magnitude of the helicity *exceeds* that of the spin for Class C representations. "Spin" is perhaps a misleading term, since $P^{-1}W$ does not have the physical significance in this case of being the angular momentum in the rest frame of the particle. Indeed, a Class C "particle" has no rest frame.[10,11] If $P^{-1}W < 0(-\frac{1}{4})$ there are no supplementary invariants. The spectrum of $P^{-1}W$ is continuous in this case, and the spectrum of helicity, for any value of the four-momentum, encompasses all integers (half-integers) for single (double)-valued representations. An explicit representation of the generators for Class C representations has been given by Shirokhov.[4] Inasmuch as the sign of the energy is no longer an invariant, the energy is a two-valued function of the three-momentum within the representation. To get around this one resorts to four-dimensional angle variables:

$$p_0 = \Pi \sinh\chi, \quad p_1 = \Pi \cosh\chi \sin\theta \cos\varphi,$$

$$p_2 = \Pi \cosh\chi \sin\theta \sin\varphi, \quad p_3 = \Pi \cosh\chi \cos\theta, \tag{2.5}$$

$$-\infty < \chi < \infty, \quad 0 \leq \theta \leq \pi, \quad -\pi \leq \varphi \leq \pi.$$

$$\Pi^2 = -p_\mu p^\mu \equiv -m^2 > 0.$$

In terms of these variables the representation is given by

$$(\mathbf{M})_1 = M_{23} = i \sin\varphi \frac{\partial}{\partial\theta} + i \cot\theta \cos\varphi \frac{\partial}{\partial\varphi} + \frac{T_1 \sinh\chi + T_0 \cosh\chi \sin\theta \cos\varphi}{1 + \cosh\chi \cos\theta}, \tag{2.6}$$

$$M_{31} = -i \cos\varphi \frac{\partial}{\partial\theta} + i \cot\theta \sin\varphi \frac{\partial}{\partial\varphi}$$

[10] D. Korff, Ph.D. Thesis, Brandeis University (1962) (to be published).

[11] O. M. P. Bilaniuk, V. K. Deshpande, and E. C. G. Sudarshan, Am. J. Phys. 30, 718 (1962).

$$+\frac{T_2\sinh\chi+T_0\cosh\chi\sin\theta\sin\varphi}{1+\cosh\chi\cos\theta},$$

$$M_{12}=-i\partial/\partial\varphi+T_0,$$

$$(\mathbf{N})_1=M_{01}=i\sin\theta\cos\varphi\frac{\partial}{\partial\chi}+i\tanh\chi\times\cos\theta\frac{\partial}{\partial\theta}-i\tanh\chi\frac{\sin\varphi}{\sin\theta}\frac{\partial}{\partial\varphi}-T_2,$$

$$M_{02}=i\sin\theta\sin\varphi\frac{\partial}{\partial\chi}+i\tanh\chi\cos\theta\times\sin\varphi\frac{\partial}{\partial\theta}+i\tanh\chi\frac{\cos\varphi}{\sin\theta}\frac{\partial}{\partial\varphi}+T_1,$$

$$M_{03}=i\cos\theta\frac{\partial}{\partial\chi}-i\tanh\chi\sin\theta\frac{\partial}{\partial\theta}+\frac{\cosh\chi\sin\theta(T_2\cos\varphi-T_1\sin\varphi)}{1+\cosh\chi\cos\theta}.$$

T_0, T_1, and T_2 constitute a representation of the following commutation rules:

$$[T_0,T_1]=iT_2,\quad [T_2,T_0]=iT_1,\quad [T_1,T_2]=-iT_0. \tag{2.7}$$

It has been shown by Bargmann that there are no finite-dimensional unitary representations of these commutation rules.[12]

An explicit representation of (2.7) with T_0 diagonal has the following form for those representations corresponding to $P^{-1}W<0(-\frac{1}{4})$:

$$\begin{aligned}(T_0)_{mn}&=n\delta_{mn},\\ (T_1+iT_2)_{mn}\equiv(T_+)_{mn}&=a_n\delta_{m,n+1};\\ (T_1-iT_2)_{mn}\equiv(T_-)_{mn}&=b_n\delta_{m+1,n},\\ a_nb_n&=-(P^{-1}W)+n(n+1);\\ &-\infty<n,\quad m<\infty.\end{aligned} \tag{2.8}$$

For $P^{-1}W=s(s+1)\geq 0(-\frac{1}{4})$, $|n|\geq s+1$ is a requirement that the representation be unitary so that $b_s=a_{-s}=0$ and n is restricted by the inequalities $-\infty<n\leq-(s+1)$ and $\infty>n\geq(s+1)$. In this latter case, as we have remarked, the sign of the helicity, which does not selectively favor the z direction in any way, is an invariant. Strictly speaking, then, the above representation is not irreducible but can be reduced to the direct sum of irreducible representations. To obtain the irreducible representations, one must transform to a *helicity* diagonal representation. We return to this point below (Sec. 5).

[12] V. Bargmann, Ann. Math. 48, 568 (1947).

The representations of Class C corresponding to $P^{-1}W<0(-\frac{1}{4})$ are denoted by P^{α}_{II}, α referring to the eigenvalue of $P^{-1}W$. For $P^{-1}W\geq 0(-\frac{1}{4})$, the conventional notation is $P^{\pm l}_{\text{II}}$, l denoting the spin, and the sign referring to the sign of the helicity. The representations given by (2.6) are now reducible. We show that there exists a unitary transformation which transforms (2.6) into the direct sum $P^{+l}_{\text{II}}\oplus P^{-l}_{\text{II}}$.

Finally, there is a scalar representation, P^{0}_{II}, corresponding to $T_0=T_1=T_2=0$.

3. THE ZERO-MASS LIMIT FROM ABOVE

Since the zero-mass irreducible representations are of necessity diagonal in the helicity, it is convenient to investigate the limit by transforming the finite-mass representations to a helicity diagonal form. The transforming operator is that of a rotation in spin space from the z axis to the $\mathbf{p}$ axis. The angles involved are momentum-dependent and hence radically affect the explicit form of the generators. For reasons of convenience we choose

$$U=\mathfrak{D}^l(-\varphi,\theta,\varphi)=e^{iS_3\varphi}e^{-iS_1\theta}e^{-iS_3\varphi}, \tag{3.1}$$
$$\theta=\tan^{-1}(p_-p_+)^{\frac{1}{2}}/p_3,\quad \varphi=\tan^{-1}(p_1/p_2).$$

In (3.1), $p_\pm\equiv p_1\pm ip_2$, and S_1, S_3 are $(2l+1)$-dimensional spin matrices. With this transformation our new operators $O'=UOU^{-1}$, are

$$\begin{aligned}M_1'&=-i(\mathbf{p}\times\nabla_p)_1+S_3(p_1|\mathbf{p}|/p_-p_+)(1-p_3/|\mathbf{p}|),\\ M_2'&=-i(\mathbf{p}\times\nabla_p)_1+S_3(p_2|\mathbf{p}|/p_-p_+)(1-p_3/|\mathbf{p}|),\\ M_3'&=-i(\mathbf{p}\times\nabla_p)_3+S_3,\\ N_1'&=ip_0\frac{\partial}{\partial p_1}-\frac{mp_1p_3}{|p|^2(p_-p_+)^{\frac{1}{2}}}S_1+\frac{mp_2}{|p|(p_-p_+)^{\frac{1}{2}}}S_2+\frac{p_0p_2}{p_-p_+}\left(1-\frac{p_3}{|p|}\right)S_3,\\ N_2'&=ip_0\frac{\partial}{\partial p_2}-\frac{mp_2p_3}{|p|^2(p_-p_+)^{\frac{1}{2}}}S_1-\frac{mp_1}{|p|(p_-p_+)^{\frac{1}{2}}}S_2-\frac{p_0p_1}{p_-p_+}\left(1-\frac{p_3}{|\mathbf{p}|}\right)S_3,\\ N_3'&=ip_0\frac{\partial}{\partial p_3}+\frac{m(p_-p_+)^{\frac{1}{2}}}{|\mathbf{p}|^2}S_1.\end{aligned} \tag{3.2}$$

A few pedagogical remarks might be in order here. There are singularities in the operators $\mathbf{M}$ and $\mathbf{N}$ for $p_1=p_2=0$. These are not physical, however, as can be seen by transforming to four-dimensional angle variables:

$$\begin{aligned}p_0&=m\cosh\chi,\quad p_1=m\sinh\chi\sin\theta\cos\varphi,\\ p_2&=m\sinh\chi\sin\theta\sin\varphi,\quad p_3=m\sinh\chi\cos\theta.\end{aligned} \tag{3.3}$$

The singularities are all of the form $(\sin\theta)^{-1}$, and are exactly canceled in all matrix elements via the factor $d^3p = p^2 \sin\theta\, d\theta\, d\varphi = m^2 \sinh^2\chi \sin\theta\, d\theta\, d\varphi$. Secondly, the form of the operators favors the direction of the z axis. This is, however, compensated by a corresponding imbalance in the wavefunction. As an illustration, consider an eigenstate of helicity and spin with eigenvalues $\frac{1}{2}, \frac{1}{2}$.

$$\left[\left(\frac{1+\hat{p}_3}{2}\right)^{\frac{1}{2}}, \quad \frac{\hat{p}_+}{[2(1+\hat{p}_3)]^{\frac{1}{2}}}\right]\psi(|\mathbf{p}|^2) = \Theta(\mathbf{p}), \tag{3.4}$$

$$\hat{p}_\pm \equiv \frac{p_\pm}{|\mathbf{p}|}, \qquad \hat{p}_3 \equiv \frac{p_3}{|\mathbf{p}|}.$$

If we define θ and φ by Eq. (3.1), then we have $\mathfrak{D}^{\frac{1}{2}}(-\phi, \theta, \phi)\Theta(\mathbf{p}) = [1\ 0]\psi(|\mathbf{p}|^2)$, which is a spherically symmetric wavefunction. However, $\Theta(\mathbf{p})$ is certainly not spherically symmetric and ∇_p operates on the functions of $\mathbf{p}$ inside the brackets as well as on $\psi(|\mathbf{p}|^2)$. To compensate, in the helicity diagonal representation, where these functions are $\mathbf{p}$-independent (namely, the constants one and zero), the operators must be asymmetric in form.

By a comparison of (2.4) and (3.2), it is clear that

$$\lim_{m\to 0} P^s_{\pm m} = P^{-s}_{\pm 0} \oplus P^{-(s-1)}_{\pm 0} \oplus \cdots \oplus P^s_{\pm 0}. \tag{3.5}$$

Two results may be immediately drawn:

(a) Particles with "small" mass behave like particles with zero mass. By "small" mass we mean that the momentum-space wavefunction of the particles differs appreciably from zero only for $m/|\mathbf{p}| \ll 1$.

(b) Particles with the same helicity and different spin behave alike in the limit of "small" mass.

4. THE SPIN AND POSITION OPERATORS FOR A MASS-ZERO PARTICLE

There are other results which follow from (3.2). Consider two particles with identical four-momenta and helicity but different spins. These two states will be distinguishable in the case of nonzero mass. In the limit, however, one has

$$\lim_{m\to 0} [\langle a|\, O\, |a\rangle - \langle b|\, O\, |b\rangle] = 0, \tag{4.1}$$

where $|a\rangle$ and $|b\rangle$ are the states referred to, and O is any of the ten generators of the Lorentz group. Equation (4.1) holds for any operator made up of sums of products of the generators with coefficients which remain finite in the zero-mass limit (i.e., which are not proportional to inverse powers of the mass). Let us call such operators "proper" operators. All other operators we refer to as "improper." An interesting example of an improper operator is the spin operator $P^{-1}W = -(m)^{-2}w_\mu w^\mu$. If proper operators are the only physically measurable operators in the limit, then one may say that $|a\rangle$ and $|b\rangle$ become indistinguishable. This follows from the fact that all proper operators may be written in the form $f_1(p)I + f_2(p)S_3$ where I is the identity matrix. Information concerning the spin resides in those terms involving S_+ and S_-. These, however, vanish from proper operators in the limit. Thus, for example, a photon need not be considered as the unique limit of a neutral vector meson but may, with equal validity, be looked on as the mass-zero limit of a spin-*two*, helicity-one state—or, for that matter, spin-n, helicity-one. Strictly speaking, the terms dependent on spin (i.e., proportional to S_+ or S_-) go to zero as $m/|\mathbf{p}|$ rather than m, so that we may roughly say that if $m < 1/\mathfrak{a}$, where $\mathfrak{a}$ is the order of the dimensions of the measuring apparatus, then those momentum states for which the states $|a\rangle$ and $|b\rangle$ are resolvable, are undetectable. Hence we may write

$$[\langle a|\, O\, |a\rangle - \langle b|\, O\, |b\rangle] = 0(\mathfrak{a}m), \quad m \ll 1/\mathfrak{a}, \tag{4.1a}$$

where O is a proper operator. We show below that although *improper* operators may be well defined in the mass-zero limit, they are nonphysical. By this we mean that their measurement cannot be experimentally performed. We see that the mathematical form of this statement is the existence, in the limit, of a superselection rule which prohibits improper operators from being physical observables. As a convenient example of an improper operator, we first consider the spin.

It is commonly assumed that a photon is a spin-one particle. That is, the mass-zero limit of a neutral vector meson with a supplementary condition attached ruling out the longitudinal, helicity-zero eigenstate. Suppose, for the sake of argument, one proposed the "ridiculous" assertion that the photon had spin two. If the photon had a finite mass, then a measurement of spin could be made by viewing the particle in different Lorentz frames and counting the number of helicity states. Let us assign the photon a finite, though vanishingly small, mass and investigate this experiment in the mass zero limit. For purposes of calculational ease, consider the decay $\pi^0 \to 2\gamma$. We denote the photon's mass by m_γ. Then if $m_\gamma \neq 0$, the statement that the photon has helicity one is no longer a Lorentz-invariant one. One may, however, make this statement in a particular preferred Lorentz frame, which we take as the rest frame of the decaying π^0. Taking

the momentum direction of the emergent photons as our z axis, we ask the following question: How fast must an observer be moving in the direction of the x axis in order to detect a component δ of helicity two (i.e., relative probability δ^2 that the photon be in a helicity-two state)? δ is presumed to be an experimental parameter determined by the resolvability of the detecting apparatus.

This calculation may be performed (Appendix A) to yield the result

$$v = \tanh(m_\pi \delta/2m_\gamma), \tag{4.2}$$

where m_π is the mass of the π^0 meson.

Clearly, if $m_\gamma \ll m_\pi \delta$, then $v \approx 1$. Equivalently, we may interpret (4.2) as the velocity the π^0 must have relative to the laboratory frame before a detectable helicity-two component is observed. If $E_{\max}$ is the greatest energy which a π^0 beam can be experimentally given, then if

$$v_{\max} = (E_{\max}^2 - m_\pi^2)^{\frac{1}{2}}/E_{\max} < v, \tag{4.3}$$

no helicity-two component will ever be observed. Note that in order to get a detectable matrix element one needs $u = \sinh^{-1}[v/(1-v^2)^{\frac{1}{2}}] \sim p_\gamma \delta/m_\gamma$. This is what one might expect inasmuch as the raising and lowering parts of $\mathbf{N}$ [i.e., those terms proportional to $S_+ (\equiv S_1 + iS_2)$ and $S_- (\equiv S_1 - iS_2)$] go to zero as $m/|\mathbf{p}|$.

As a possible experimental consequence of the above discussion we note that if one extends a massless, helicity q, particle off the light cone by giving it a small mass, then this extension is not unique, since the spin is unspecified and need not equal q. In particular, if one wishes to speculate about the physical consequences of assigning a finite mass to the muon's neutrino, then one may take into account the possibility of a spin different than one-half.

As we have noted, any proper operator will, in the limit of mass zero, have no matrix elements connecting states of different helicity and identical spin. Due to finite resolvability parameters of any measuring apparatus, this condition obtains at a finite, though infinitesimal, value of the mass, which is determined by the experimental parameters.

If proper operators are the only physically meaningful operators in the limit, then each eigenstate of helicity, together with the four-momentum continuum, spans a Hilbert space. That improper operators have no physical significance in the mass-zero limit can be seen from the following argument: The four components of the four-momentum, together with w_0 and $w_\mu w^\mu$, constitute a complete commuting set of observables. Any two states which have the same eigenvalues of all the above operators must be physically identical, else we would not have a complete set. In the limit, $w_\mu w^\mu$, being essentially the spin *times the mass*, approaches zero, and information concerning the spin becomes increasingly difficult to obtain in the sense described above. Any operator which differentiates between two states with identical eigenvalues of a complete set must be nonphysical. *Improper operators are exactly such operators.* Another way of looking at this is the following: P and W represent only two invariants of the Lorentz group. Allowing the spin a physical interpretation yields *three* invariants; namely P (equal to zero), W (equal to zero), and $P^{-1}W$ [equal to $s(s+1)$].

Alternatively, the above discussion can be given the form of a superselection rule. Consider two states of a spin-s particle with finite mass

$$|1\rangle = |a\rangle + |b\rangle, \qquad |2\rangle = |a\rangle + e^{i\delta}|b\rangle, \tag{4.4}$$

where a and b are states of different helicity. States 1 and 2 are resolvable since the representations (3.2) contain raising and lowering operators which will take into account the value of the phase factor. In the limit, however, these terms vanish [cf. (3.2)], and no measurement may distinguish the two states. In this case we say that the Hilbert spaces consisting of different helicity states (plus the four-momentum manifold) are incoherent. Any operator which has matrix elements between these incoherent Hilbert spaces is unphysical. One says that a superselection rule operates against it. The existence of this superselection rule corresponds to an operator, the helicity, which commutes with all the elements of the Lorentz group in the mass-zero limit.

An additional example of an improper operator is the *position* operator, which can be written[13]

$$\mathbf{x} = \mathbf{N}p_0^{-1} + i\mathbf{p}/2p_0^2 - (\mathbf{p}\times\mathbf{w})/mp_0(m+p_0), \qquad \mathbf{w} \equiv (w_1, w_2, w_3). \tag{4.5}$$

In the mass-zero limit we have, using the helicity diagonal form,

$$\begin{aligned} x_1' &= Ux_1U^{-1} = i\partial/\partial p_1 + S_1[p_1p_3/|\mathbf{p}|^2\,(p_-p_+)^{\frac{1}{2}}] \\ &\quad + S_2[p_2/|\mathbf{p}|\,(p_-p_+)^{\frac{1}{2}}] - S_3[p_3p_2/|\mathbf{p}|\,(p_-p_+)], \\ x_2' &= i\partial/\partial p_2 - S_1[p_2p_3/|\mathbf{p}|^2\,(p_-p_+)^{\frac{1}{2}}] \\ &\quad - S_2[p_1/|\mathbf{p}|\,(p_-p_+)^{\frac{1}{2}}] + S_3[p_3p_1/|\mathbf{p}|\,(p_-p_+)], \\ x_3' &= i\partial/\partial p_3 + S_1[(p_-p_+)^{\frac{1}{2}}/|\mathbf{p}|^2]. \end{aligned} \tag{4.6}$$

[13] C. Kuang-Chao and M. I. Shirokhov, Zh. Eksperim. i Teor. Fiz. **34**, 1230 (1958) [English transl.: Soviet Phys.—JETP **7**, 851 (1958)].

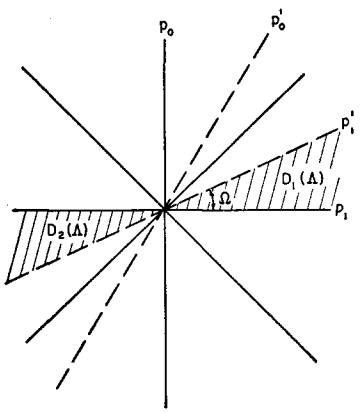


FIG. 1. The domains D_1 and D_2 for a pure Lorentz transformation along the x axis.

Improper operators are thus to be considered perfectly *definable* even in the mass-zero limit. However, as the limit is approached they take on an increasingly nonphysical character, due to the fact that they resolve states which are *physically* unresolvable.

Let us briefly summarize the above discussion. From (3.2) we obtain the result that states with similar wavefunctions in the helicity diagonal representations are certainly distinguishable. By *similar* wavefunctions we mean the following: Two wavefunctions are said to be *similar* if all nonzero components of the same helicity differ at most by a constant phase factor. Thus

$$\Theta_1(\mathbf{p}) = (0\ 0\ \psi(\mathbf{p})\ 0\ 0) \tag{4.7a}$$

and

$$\Theta_2(\mathbf{p}) = e^{i\delta}\psi(\mathbf{p}) \tag{4.7b}$$

are similar wavefunctions. That states with similar wavefunctions but different spin are resolvable is due to the occurrence of S_+ and S_- operators in the expressions (3.2). The coefficients, in the generators, of these operators *vanish* in the mass-zero limit. In this limit no proper operator will resolve states with different spin but similar wavefunctions. Improper operators, which *will* do this, become nonphysical in this limit in a continuous manner.

5. THE LIMIT FROM BELOW

The zero-mass limit of imaginary mass "particles" has many features in common with that of real mass. There is one new feature, however. The manifold $p_\mu p^\mu = m^2 < 0$ is a connected one-sheet hyperboloid. For the zero-mass limit the irreducible representations are *direct sums* of irreducible representations characterized by a definite sign of the energy. The representations (2.6) do not lend themselves to this limit. We proceed as follows. Restricting ourselves for the moment to the scalar case, consider the following substitutions of wavefunctions

$$\psi'(\theta, \varphi, \chi) \equiv \begin{bmatrix} \theta(\chi)\psi(\theta, \varphi, \chi) \\ \theta(-\chi)\psi(\theta, \varphi, \chi) \end{bmatrix} \tag{5.1a}$$

and of operators

$$O'(\chi\theta\varphi) \equiv \begin{bmatrix} \theta(\chi) & \theta(\chi) \\ \theta(-\chi) & \theta(-\chi) \end{bmatrix} O(\chi\theta\varphi). \tag{5.1b}$$

χ, θ, and φ have been defined above [Eq. (2.5)]. $\theta(\chi)$ is a unit step function. The scalar product is then invariant under the substitution (5.1a). Furthermore one has, from (5.1b), $O'_{\Lambda_1}O'_{\Lambda_2} = O'_{\Lambda_1\Lambda_2}$, where O_Λ is an operator corresponding to a Lorentz transformation Λ. Thus the substitutions (5.1) yield the identical representation.

Utilizing $[\theta(\chi)]^2 = \theta(\chi)$, $\theta(\chi)\theta(-\chi) = 0$, we have

$$O'\psi' = \begin{bmatrix} O(|\chi|\ \theta\varphi) & 0 \\ 0 & O(-|\chi|\ \theta\varphi) \end{bmatrix} \begin{bmatrix} \theta(\chi)\psi \\ \theta(-\chi)\psi \end{bmatrix} + \begin{bmatrix} [\theta(\chi), O]\psi \\ [\theta(-\chi), O]\psi \end{bmatrix}. \tag{5.2}$$

Now

$$\begin{aligned} [\theta(\chi), O]\psi(\chi\theta\varphi) &= I(D_1(\Lambda))\psi(\Lambda^{-1}[\theta, \varphi, \chi]), \\ [\theta(-\chi), O]\psi(\chi\theta\varphi) &= I(D_2(\Lambda))\psi(\Lambda^{-1}[\theta, \varphi, \chi]), \end{aligned} \tag{5.3}$$

where $I(D(\Lambda))$ equals unity if the point (χ, θ, φ) lies in the domain $D(\Lambda)$, and equals zero otherwise. $D_1(\Lambda)$ consists of all those points whose zero component changes from positive-definite to non-positive-definite under the Lorentz transformation Λ. $D_2(\Lambda)$ is the domain over which the change is from negative-definite to non-negative-definite. For a pure Lorentz transformation along the x-axis, the situation is illustrated in Fig. 1.

Clearly, only pure Lorentz transformations need be considered since $[\theta(\chi), \mathbf{M}] = 0$, the sign of the energy being unaffected by space rotations. If now $\psi(\chi, \theta, \varphi)$ is zero for all points in D_1 or D_2, i.e., if $D' = D_1 \cup D_2$, and

$$\text{supp } \psi(\chi\theta\varphi) \cap D'(\Lambda) = 0, \tag{5.4}$$

then the second term in (5.2) is zero. Now let $u = \sinh^{-1}[v/(1 - v^2)^{\frac{1}{2}}]$ characterize the Lorentz transformation. Its magnitude determines the magnitude of Ω in Fig. 1 and hence the size of the domains D_1 and D_2. If (5.4) holds for $\Lambda(u)$ then it holds for all $u' < u$ (smaller angle Ω). Let $u_{\max}$ be the largest u producible in a laboratory. If (5.4) holds for $\Lambda(u_{\max})$, the last term in (5.2) is zero for all

Lorentz transformations *executable in a laboratory* and the representation becomes "physically reducible" into a direct sum of irreducible representations. If $D''(u)$ is the sum of the domains $D'(\Lambda(u))$ and $D'(\Lambda(-u))$, then all points with

$$|p_0| = \Pi\,|\sinh\chi| < \Pi\,|\sinh u| \tag{5.5}$$

are in $D''(u)$ (see Fig. 2). By (5.5), if the wave-function ψ equals zero for $(\chi, \theta, \varphi) \subset D''(u_{\max})$, the representation is physically reducible. In the limit as Π goes to zero, (5.5) states that the top and bottom light cones are open sets (i.e., excluding the origin).

To examine the zero-mass limit of the nonscalar, imaginary-mass representations, we combine the substitutions (5.1) with a transformation to a helicity diagonal form. To achieve the latter, we seek a unitary transformation to diagonalize $p\cdot M$ (cf. Sec. 3). Since, for $p_\mu p^\mu < 0$, $p\cdot M \neq p\cdot S$, the transformation is no longer simply a rotation in spin space. However, by analogy, we consider the ansatz

$$U = e^{iT_0\beta}e^{-iT_1\alpha}e^{-iT_0\beta}, \tag{5.6}$$

where T_0, T_1, T_2 satisfy (2.7). Accordingly

$$\begin{aligned}
e^{iT_0\beta}T_1e^{-iT_0\beta} &= T_1\cos\beta - T_2\sin\beta,\\
e^{iT_0\beta}T_2e^{-iT_0\beta} &= T_2\cos\beta + T_1\sin\beta,\\
e^{iT_1\alpha}T_0e^{-iT_1\alpha} &= T_0\cosh\alpha + T_2\sinh\alpha,\\
e^{iT_1\alpha}T_0e^{-iT_1\alpha} &= T_2\cosh\alpha + T_0\sinh\alpha.
\end{aligned} \tag{5.7}$$

Substituting U into the equation

$$U\mathbf{p}\cdot\mathbf{M}U^{-1} = |\mathbf{p}|\,T_0, \tag{5.8}$$

one obtains via (5.7) and much tedious algebra

$$\begin{aligned}
\alpha &= \tanh^{-1}\left[\frac{p_0(p_-p_+)^{\frac{1}{2}}}{(|p|^2 + p_3\Pi)}\right]\\
&= \tanh^{-1}\left[\frac{\sinh\chi\sin\theta}{\cosh\chi + \cos\theta}\right],\\
\beta &= \tan^{-1}(p_1/p_2) = \tfrac{1}{2}\pi - \varphi,\\
M_1' &= UM_{23}U^{-1} = -i(\mathbf{p}\times\nabla_p)_1\\
&\quad + T_0(p_1\,|\mathbf{p}|/p_-p_+)(1 - p_3/|\mathbf{p}|),\\
M_2' &= -i(\mathbf{p}\times\nabla_p)_2 + T_0(p_2\,|\mathbf{p}|/p_-p_+)(1 - p_3/|\mathbf{p}|),\\
M_3' &= -i(\mathbf{p}\times\nabla_p)_3 + T_0,\\
N_1' &= ip_0\frac{\partial}{\partial p_1} + T_0\left(\frac{-p_2p_3}{p_-p_+} + \frac{\Pi^2p_2}{|\mathbf{p}|^2(p_3+\Pi)}\right)\\
&\quad + T_1\frac{p_1}{(p_-p_+)^{\frac{1}{2}}}\left(1 - \frac{(|\mathbf{p}|^2 - p_3\Pi)}{|\mathbf{p}|\,p_0}\right)
\end{aligned} \tag{5.9}$$

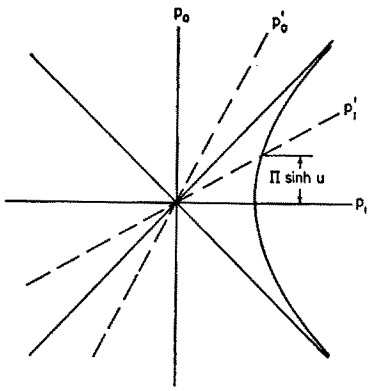


FIG. 2. A graphic illustration of the inequality (5.5). Only $D'(\Lambda(u))$ is shown. The curve on the right is the positive branch of the hyperbola $p_0^2 - p_1^2 = -\Pi^2$.

$$\begin{aligned}
&\quad + T_2\frac{p_2}{(p_-p_+)^{\frac{1}{2}}}\left(\frac{p_0}{p_3+\Pi} - \frac{(|\mathbf{p}|^2 + p_3\Pi)}{|\mathbf{p}|\,(p_3+\Pi)}\right),\\
N_2' &= ip_0\frac{\partial}{\partial p_2} + T_0\left(\frac{p_1p_3}{p_-p_+} + \frac{\Pi^2p_1}{|\mathbf{p}|^2(p_3+\Pi)}\right)\\
&\quad + T_1\frac{p_2}{(p_-p_+)^{\frac{1}{2}}}\left(1 - \frac{(|\mathbf{p}|^2 - p_3\Pi)}{|\mathbf{p}|\,p_0}\right)\\
&\quad + T_2\frac{p_1}{(p_-p_+)^{\frac{1}{2}}}\left(\frac{p_0}{p_3+\Pi} - \frac{(|\mathbf{p}|^2 + p_3\Pi)}{|\mathbf{p}|\,(p_3+\Pi)}\right),\\
N_3' &= ip_0\frac{\partial}{\partial p_3} + T_1\frac{(p_-p_+)^{\frac{1}{2}}}{p_0}\left(\frac{|\mathbf{p}| - p_0}{p_3+\Pi} - \frac{\Pi}{|\mathbf{p}|}\right).
\end{aligned}$$

To cast (5.9) into a form suitable for the transition to the limit, and take into account the double valuedness of the energy within the irreducible representation, we must transform to the four-dimensional angle variables (2.5), and utilize the substitution (5.1). Thus, for example,

$$\begin{aligned}
N_2' = M_{02}' &= \begin{pmatrix}\theta(\chi) & \theta(\chi)\\ \theta(-\chi) & \theta(-\chi)\end{pmatrix}\\
&\times\left[i\sin\theta\sin\varphi\frac{\partial}{\partial\chi} + i\tanh\chi\cos\theta\sin\varphi\frac{\partial}{\partial\theta}\right.\\
&+ i\tanh\chi\frac{\cos\varphi}{\sin\theta}\frac{\partial}{\partial\varphi}\\
&+ T_0\left(\cot\theta\cos\varphi + \frac{\sin\theta\cos\varphi}{\cosh\chi[1+\cosh\chi\cos\theta]}\right)\\
&+ T_1\sin\varphi\left(1 - \left[\frac{\cosh\chi - \cos\theta}{\sinh\chi}\right]\right)\\
&\left.+ T_2\cos\varphi\left(\frac{\cos\theta - \sinh\chi - \cosh\chi}{1+\cosh\chi\cos\theta}\right)\right].
\end{aligned} \tag{5.10}$$

One disadvantage of the form (5.10), however, is that the passage to the mass-zero limit is not as transparent as in the form (5.9). From the expressions (5.9) we see that all coefficients of T_1, T_2 approach zero as Π approaches zero, and the representation reduces into a direct sum of an infinite

number of one-component, helicity-diagonal representations. Except for the energy ambiguity these are identical with (2.4). Taking into account (5.2) and (5.8), one obtains the direct sum of a double infinity of representations—that is, two irreducible representations for each sign of the helicity, corresponding to the two different signs of the energy.

We note several interesting facts. First, the representations (5.9) and (5.10) for $P_{\Pi}^{\pm l}$, prior to the taking of the limit, are in a reduced form, corresponding to the direct sum $P_{\Pi}^{l} \oplus P_{\Pi}^{-l}$. Second, *all* the (unitary) P_{Π} representations reduce in the limit to a direct sum of the *physical*, one-component, mass-zero representations.[14] It is clear that this must be so. The unphysical mass-zero representations are characterized by an "infinite" spin (i.e., P equals zero whereas W is finite). All the irreducible representations for P less than zero, however, are characterized by a fixed, finite value of the spin (which may, of course, be arbitrarily large, though finite). In the limit the spin ceases to be an observable but it remains constant as one lets the squared mass approach zero from below. It becomes identifiable, so to speak, with a 0/0 type operator (i.e., P and W are both zero), rather than a divergent one.

Since the coefficients of T_1 and T_2 in (5.9) approach zero continuously with the mass, it is clear that one may always find a small enough value of the (imaginary) mass such that below this value the particle is indistinguishable from a zero-mass particle. The zero-mass particles which exist in nature can be considered as continuous limits of particles with real or imaginary mass. Owing to the fact that the unphysical, infinite spin, mass-zero representations do not admit a continuous description either from above or below it is perhaps not surprising that the "particles" to which they correspond do not exist.

6. SUMMARY AND CONCLUSIONS

We have investigated the mass-zero limits of the non-mass-zero representations and have found that both physically and mathematically the mass-zero limit is in no way singular. Indeed it is only the infinite-spin representations corresponding to $W \neq 0$, $P = 0$, which do not admit of a continuous description. That is, neither the $P > 0$ nor $P < 0$ representations reduce to them physically or mathematically.

The spin of a mass-zero particle is physically ambiguous, though for a finite-mass particle the spin is *definable* right up to zero mass. The nonphysical nature of the spin comes on in a continuous manner, inasmuch as it becomes increasingly difficult to perform the measurement of spin as the mass approaches zero. Since all experimental apparatus have finite parameters of resolution, one can always find a value of the mass below which the concept of spin has no experimental meaning.

The helicity of a zero-mass particle does not determine its spin. A zero-mass particle with helicity q (q must be an integer or half-integer—cf. Sec. 2), may, with equal validity, be considered as the zero-mass limit of a helicity q state of (a) Any massive particle with spin $= q + n$, where n is any nonnegative integer. (b) Any imaginary mass "particle" with spin $= q - (n + 1)$, where n is any nonnegative integer. (c) Any imaginary mass "particle" with $P^{-1}W < O(-\frac{1}{4})$ for single (double)-valued representations.

The position operator for a nonzero (real)-mass particle loses much of its meaning in the mass-zero limit. It can still be *defined* but the indicated measurement cannot be carried out. The mathematical expression of this fact takes the form of a superselection rule which is turned on, as it were, in a continuous manner as the mass goes to zero.

Finally, we have seen that particles of "sufficiently small" imaginary mass are indistinguishable from zero-mass particles, and hence cannot be excluded on experimental grounds from a physical theory. By "sufficiently small" we mean sufficiently small in comparison with the resolution parameters of a measuring apparatus, as explained in the text. The possible inclusion of imaginary mass particles with finite imaginary mass into a physical theory has been discussed elsewhere.[10,11]

ACKNOWLEDGMENTS

We should like to express our thanks to Professor K. W. Ford for reviewing the manuscript and to Professor D. Falk for several valuable suggestions. We are also grateful to Lockheed Missiles and Space Company for their support during the early part of this work. Finally, we are deeply indebted to Dr. R. A. Berg for many stimulating and illuminating discussions.

APPENDIX A

In this appendix we sketch the calculation leading to Eq. (4.2). The effect of a Lorentz transformation

[14] This result implies keeping the spin fixed and varying the mass, which is usually the case of physical interest. Using the method of group contractions, Robinson has considered a limit of a sequence of representations with different spin. D. W. Robinson, Helv. Phys. Acta 35, 98 (1962).

on the wavefunction takes the form

$$e^{iN\cdot u}\psi(p) = e^{iS_3\theta}\psi(\Lambda^{-1}(u)p), \tag{A1}$$

where θ is momentum-dependent and is determined by Eq. (9) of Ref. 15. In the helicity diagonal representation we may rewrite (A1) to read

$$\begin{aligned} e^{iS_3\theta'}e^{iN\cdot u}\psi(p) &= e^{iS_3(\theta'+\theta)}\psi(\Lambda^{-1}(u)p) \qquad \text{(A2)} \\ &= e^{iS_3\theta''}\psi(\Lambda^{-1}(u)p), \end{aligned}$$

where $\cos\theta' = p_3/|\mathbf{p}|$. In the limit one finds

$$\lim_{m\to 0} e^{iS_3\theta''} = (1 + iS_2mu/|\mathbf{p}|) \tag{A3}$$

from which (4.2) follows immediately.

APPENDIX B

Particles with zero mass travel only with the velocity of light whereas particles with finite (real) mass may travel with any velocity less than that of light. How then can mass-zero particles be considered as the continuous limit of non-mass-zero particles?

Let $v_{\max}$ be the *maximum* velocity resolvable from that of light. Then the maximum resolvable momentum is

$$p_{\max} = mv_{\max}/(1 - v_{\max}^2)^{\frac{1}{2}}. \tag{B1}$$

If, however

$$m < (1 - v_{\max}^2)^{\frac{1}{2}}/\mathfrak{a}v_{\max}, \tag{B2}$$

where $\mathfrak{a}$ is the dimension of the measuring apparatus, then

$$p_{\max} < 1/\mathfrak{a}, \tag{B3}$$

and all states of the particle with velocities resolvable from that of light are undetectable.

The limits on the mass in (B2), (4.1a), and (4.3) are all different. There are many more limits possible, depending both on which experiment we are performing and which property of a zero-mass particle we wish a particle with infinitesimal mass to simulate. If the particle is to be indistinguishable from a zero-mass particle in every way, then the mass must be less than the lowest of these limits.

[15] Yu. M. Shirokhov, Zh. Eksperim. i Teor. Fiz. 35, 1005 (1958) [English transl.: Soviet Phys.—JETP 8, 703 (1959)].

Continuous-Representation Theory. IV. Structure of a Class of Function Spaces Arising from Quantum Mechanics

JAMES MCKENNA AND JOHN R. KLAUDER
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey
(Received 15 November 1963)

A rigorous development of the continuous representation of Hilbert space by bounded, continuous, multidimensional phase-space functions $\psi(p, q)$ is presented. It is shown that these functions form a closed subspace of $L^2(p, q)$ whose elements are functions and not equivalence classes. Differential properties are investigated and it is pointed out that there are a multitude of definitions whereby $\psi(p, q)$ possesses continuous derivatives of all orders. In one of these definitions, each $\psi(p, q)$ is proportional to a multidimensional, entire function $f(q - ip)$, establishing a connection between Bargmann's Hilbert space of entire functions and one example of a continuous representation. Attention is devoted to the purely functional characterization of the continuous representation by means of the reproducing kernel as a special case of Aronszajn's general theory. Properties of various operators in a continuous representation are carefully defined.

1. INTRODUCTION

CONTINUOUS representations of Hilbert space have been introduced abstractly in Part I,[1] and a few of their quantum mechanical applications have been discussed in Parts II and III.[2] However, irrespective of any possible applications, functional representations of Hilbert space certainly have an intrinsic interest all their own. In this paper, therefore, we examine one important class of functional representations afforded by continuous representations in considerable detail and rigor, and relate these representations to more familiar ones. As our example we consider the phase-space continuous representations pertinent to the description of N quantum-mechanical degrees of freedom. This example serves both to illustrate the machinery needed to discuss a continuous representation in an infinite-dimensional Hilbert space, as well as to form an important preliminary for a continuous representation of boson fields as functionals on test functions to be discussed in a subsequent paper.

In a certain sense, the function spaces we discuss constitute generalizations of the Hilbert space of entire functions first considered by Segal,[3] and more recently studied in great detail by Bargmann,[4] and subsequently employed by him in a discussion of representations of the rotation group.[5] Additional applications of this space have been made by Schweber to Feynman quantization.[6] More recently, Sudarshan and Glauber have employed a closely related Hilbert space in order to study light beams.[7] These applications provide additional motivation for our present study.

Fundamental for the construction of a continuous representation is the notion of an overcomplete family of states (OFS) which generates the continuous representation.[1] In order to make this paper self-contained, we restate the properties defining an OFS, and analyze in detail the OFS pertinent to our continuous representations. In so doing considerable information regarding the associated continuous representations is also gleaned.

Prior to defining an OFS and outlining the contents of the paper, we settle a few basic questions of notation. Throughout the paper, the term Hilbert space means an infinite-dimensional, separable Hilbert space defined over the field of complex numbers. If $\mathfrak{H}$ is a Hilbert space and $\Psi' \in \mathfrak{H}$, $\Psi \in \mathfrak{H}$, then the inner product is denoted by (Ψ', Ψ), which, conforming to quantum-mechanical usage, is linear in the second term and conjugate linear in the first. The norm of Ψ is $||\Psi|| = (\Psi, \Psi)^{\frac{1}{2}}$. The commutator of two operators, X and Y, defined on $\mathfrak{H}$ is always defined as

$$[X, Y] = XY - YX. \tag{1.1}$$

The postulates defining an OFS can now be listed.[1] Let $\mathfrak{H}$ be a Hilbert space, and let $\mathfrak{T}$ be the set of all unit vectors in $\mathfrak{H}$. Then a subset $\mathfrak{S} \subset \mathfrak{T}$

[1] J. R. Klauder, J. Math. Phys. 4, 1055 (1963), referred to as I.
[2] J. R. Klauder, J. Math. Phys. 4, 1058 (1963), referred to as II; J. Math. Phys., 5, 177 (1964), referred to as III.
[3] I. E. Segal, *Proceedings of the Summer Seminar, Boulder, Colorado, 1960. Vol. II. Mathematical Problems of Relativistic Physics* (American Mathematical Society, Providence, Rhode Island, 1960).
[4] V. Bargmann, Communs. Pure Appl. Math. 14, 187 (1961).
[5] V. Bargmann, Rev. Mod. Phys. 34, 829 (1962).
[6] S. S. Schweber, J. Math. Phys. 3, 831 (1962).
[7] E. C. G. Sudarshan, Phys. Rev. Letters 10, 277 (1963); R. J. Glauber, Phys. Rev. Letters 10, 84 (1963); Phys. Rev. 131, 2766 (1963).

is called an OFS if it satisfies the following three postulates.

Postulate 1. (Local density and continuity.) For each $\Phi \in \mathfrak{S}$ and every $\delta > 0$, there exists a vector $\Phi' \in \mathfrak{S}$ different from Φ for which $||\Phi - \Phi'|| < \delta$. The set $\mathfrak{S}$ is an arcwise connected subset of $\mathfrak{H}$ or a union thereof.

Postulate 2. (Label continuity.) There is a Hausdorff space $\mathfrak{L}$ (called the "label" space) and a mapping $\mathfrak{M}$ of $\mathfrak{L}$ onto $\mathfrak{S}$. This mapping is weakly continuous, by which we mean the following. If $l \in \mathfrak{L}$ and $\mathfrak{M}$ maps l onto $\Phi[l] \in \mathfrak{S}$, then $(\Phi[l], \Psi)$ is a continuous function with respect to the topology of $\mathfrak{L}$ for every $\Psi \in \mathfrak{H}$.

Postulate 3. (Completeness and resolution.) The set $\mathfrak{S}$ spans the space $\mathfrak{H}$, i.e., the completion in norm of the set of all linear combinations of elements in $\mathfrak{S}$ yields $\mathfrak{H}$. The identity operator in $\mathfrak{H}$ can be resolved into an integral over projection operators onto individual vectors in $\mathfrak{S}$.

In this paper, we study in detail the OFS defined roughly as follows. Let $\mathfrak{H}$ be an arbitrary Hilbert space and let Q and P be two linear, self-adjoint operators defined on $\mathfrak{H}$, which satisfy the commutation relations (we set $\hbar = 1$):

$$[Q, P] = iI, \tag{1.2}$$

where I is the unit operator. If $\Phi_0 \in \mathfrak{H}$ is an arbitrary, fixed unit vector, we define the set $\mathfrak{S}$ to be the collection of vectors

$$e^{-iqP}e^{ipQ}\Phi_0 = \Phi[p, q], \tag{1.3}$$

for all real p and q. The corresponding continuous representation of $\mathfrak{H}$ is the space of functions

$$\psi(p, q) = (\Phi[p, q], \Psi), \tag{1.4}$$

for all $\Psi \in \mathfrak{H}$.[2] Throughout the paper a continuous representation always implies a phase-space continuous representation of the character of (1.3) and (1.4).

The contents of the body of this paper can now be summarized. In Sec. 2 we generalize and make mathematically rigorous the definition of $\mathfrak{S}$ given by (1.3). In particular, as regards the generalization, the label space is taken to be $2N$-dimensional rather then 2-dimensional, corresponding to N degrees of freedom. The proof that the set $\mathfrak{S}$ defined in Sec. 2 constitutes an OFS is given in Sec. 3. Section 4 is devoted to constructing the continuous representation of $\mathfrak{H}$ generated by $\mathfrak{S}$. The main topic of this section is how the differentiability of the functions $\psi(p, q)$ in the continuous representation depends on the choice of Φ_0. The intimate connection between the theory of continuous representations and Aronszajn's theory of Hilbert spaces with kernel functions is developed in Sec. 5.[8,9] Finally, Sec. 6 is devoted to the form taken by linear operators in a continuous representation of $\mathfrak{H}$. Examples of several important operators are given.

Before proceeding to the main part of the paper, we should like to comment on the methods we have used in proving some of our results. It will soon become apparent that a theorem of von Neumann's, which provides us with a canonical map of our abstract space $\mathfrak{H}$ onto the space of square integrable functions, plays a crucial role in many of our proofs.[10] Most importantly, it allows us to use the powerful machinery of the theory of Fourier transforms in L^2 in proving Theorem 3.1, which is the central theorem of the paper. In a certain sense, we feel that this is a drawback to our paper, for von Neumann's theorem severely restricts us in attempts to generalize our results to the case where the label space is a space of test functions. Nevertheless, we were unable to prove several crucial results without the aid of this theorem. Whenever we found that we could replace von Neumann's theorem by more general techniques, however, we did so for the sake of generality, even if the cost was a slightly longer proof.

2. DEFINITION OF THE OVERCOMPLETE FAMILY OF STATES

As stated in the introduction, we are interested in considering an arbitrary Hilbert space $\mathfrak{H}$, and the OFS $\mathfrak{S}$, $\mathfrak{S} \subset \mathfrak{H}$, consisting of all unit vectors of the form $e^{-iqP}e^{ipQ}\Phi_0$, where Φ_0 is an arbitrary, but fixed, unit vector in $\mathfrak{H}$. [We do not impose here the constraints $(\Phi_0, P\Phi_0) = (\Phi_0, Q\Phi_0) = 0$ required in II for specific applications.] This corresponds to a quantum-mechanical system with one degree of freedom, and to generalize this to a system of N degrees of freedom, we first introduce $2N$ self-adjoint operators Q_α and P_α, $\alpha = 1, 2, \cdots, N$. We assume that these operators satisfy the commutation relations

$$[Q_\alpha, P_\beta] = i\delta_{\alpha\beta}I, \quad [Q_\alpha, Q_\beta] = 0, \quad [P_\alpha, P_\beta] = 0, \tag{2.1}$$

where $\delta_{\alpha\beta}$ is the Kronecker delta function. We then consider the set of vectors $\mathfrak{S}$, defined for all values of the real variables q_α and p_α,

$$\prod_{\alpha=1}^{N} e^{-iq_\alpha P_\alpha}e^{ip_\alpha Q_\alpha}\Phi_0. \tag{2.2}$$

[8] N. Aronszajn, Proc. Cambridge Phil. Soc. **39**, 133 (1943).
[9] N. Aronszajn, Trans. Am. Math. Soc. **68**, 337 (1950).
[10] J. von Neumann, Math. Ann. **104**, 570 (1931).

As is well known, the operators Q_α and P_α are unbounded, and as a consequence are somewhat awkward to handle directly. The operators $e^{ip_\alpha Q_\alpha}$ and $e^{-iq_\alpha P_\alpha}$, however, when appropriately defined are unitary operators, and hence much nicer to work with. Keeping this in mind, we reformulate the definition of $\mathfrak{S}$ by starting *ab initio* with an appropriate set of $2N$ unitary operators.

To facilitate this construction, we first recall the following definition. A one-parameter group of unitary operators on $\mathfrak{H}$ is a function $U[t]$ defined on the whole real line and taking on values in the set of unitary operators mapping $\mathfrak{H}$ onto itself. In addition, $U[t]$ satisfies the relations

$$U[0] = I, \qquad U[s]U[t] = U[s + t], \tag{2.3}$$

where I is the unit operator.[11] The one-parameter group of unitary operators $U[t]$ is said to be strongly continuous if, for all $\Psi \in \mathfrak{H}$,

$$\lim_{t\to 0} ||(U[t] - I)\Psi|| = 0. \tag{2.4}$$

Let there be $2N$ strongly continuous, one-parameter groups of unitary operators $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$, $\alpha = 1, 2, \cdots, N$. Let these operators satisfy the following relations:

$$\begin{gathered}[V_\alpha[q_\alpha], V_\beta[q_\beta]] = 0, \quad [W_\alpha[p_\alpha], W_\beta[p_\beta]] = 0, \\ V_\alpha[q_\alpha]W_\beta[p_\beta] = e^{-ip_\beta q_\alpha \delta_{\alpha\beta}} W_\beta[p_\beta]V_\alpha[q_\alpha].\end{gathered} \tag{2.5}$$

If $\mathbf{q}$ denotes the set $\{q_\alpha\}$ and $\mathbf{p}$ denotes $\{p_\beta\}$, then we define

$$\begin{gathered}\mathfrak{V}[\mathbf{q}] \equiv \prod_{\alpha=1}^{N} V_\alpha[q_\alpha], \quad \mathfrak{W}[\mathbf{p}] \equiv \prod_{\beta=1}^{N} W_\beta[p_\beta], \\ U[\mathbf{p}, \mathbf{q}] \equiv \mathfrak{V}[\mathbf{q}]\mathfrak{W}[\mathbf{p}].\end{gathered} \tag{2.6}$$

Now if $\Phi_0 \in \mathfrak{H}$ is an arbitrary unit vector, set

$$\Phi[\mathbf{p}, \mathbf{q}] = U[\mathbf{p}, \mathbf{q}]\Phi_0, \tag{2.7}$$

and $\mathfrak{S}$ is the set of all $\Phi[\mathbf{p}, \mathbf{q}]$ as $(\mathbf{p}, \mathbf{q})$ varies over $R_N \times R_N$, where R_N denotes the N-fold direct product of the real line with itself, supplied with the product topology. We take $R_N \times R_N$, supplied with the product topology, as the label space $\mathfrak{L}$, and the map $\mathfrak{M}$ is defined by (2.7), namely $\mathfrak{M}$: $(\mathbf{p}, \mathbf{q}) \to \Phi[\mathbf{p}, \mathbf{q}]$. The replacement of the ordinary commutation relations (2.1) by the relations (2.5) is due to Weyl.[12]

The set $\mathfrak{S}$ defined by (2.7) depends on the choice of the unit vector Φ_0 (hereafter referred to as the "fiducial vector"), and some of the consequences of this dependence will be the subject of later sections of this paper. The set $\mathfrak{S}$ also depends on the representation of the operator $U[\mathbf{p}, \mathbf{q}]$, but this dependence is much less fundamental and can be settled now. The problem of determining all possible representations of $U[\mathbf{p}, \mathbf{q}]$ was completely solved in a celebrated paper by von Neumann.[10] We summarize in the following theorem those results of Von Neumann's paper which will be of basic importance for our work.

Theorem (von Neumann). Let $\mathfrak{H}$ be an arbitrary, separable Hilbert space, and $V_\alpha[q_\alpha]$, $W_\alpha[p_\alpha]$, $\alpha = 1, 2, \cdots, N$, be $2N$ strongly continuous,[13] one-parameter groups of unitary operators on $\mathfrak{H}$ which satisfy relations (2.5). Then $\mathfrak{H}$ can be decomposed into the direct sum of a finite or countably infinite number of mutually orthogonal, closed subspaces, $\mathfrak{H} = \sum_n \oplus \mathfrak{H}_n$, such that each subspace is simultaneously invariant under the groups of operators $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$, $\alpha = 1, 2, \cdots, N$. The representations of these groups induced on $\mathfrak{H}_n$ and $\mathfrak{H}_m$ are unitarily equivalent to one another for all n and m. Furthermore, there is a unitary map T_n of each $\mathfrak{H}_n$ onto the Hilbert space $L^2(R_N)$ of Lebesgue-measurable, complex-valued functions of N real variables, $f(x_1, \cdots, x_N)$, which are square integrable over R_N. This unitary map T_n has the important property that it maps the representation of the operators $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$ induced on $\mathfrak{H}_n$ onto the operators defined as follows:

$$\begin{aligned}&T_n V_\alpha[q_\alpha]T_n^{-1}f(x_1, \cdots, x_N) \\ &\quad = f(x_1, \cdots, x_{\alpha-1}, x_\alpha - q_\alpha, x_{\alpha+1}, \cdots, x_N),\end{aligned} \tag{2.8}$$

$$\begin{aligned}&T_n W_\alpha[p_\alpha]T_n^{-1}f(x_1, \cdots, x_N) \\ &\quad = e^{ip_\alpha x_\alpha} f(x_1, \cdots, x_N).\end{aligned} \tag{2.9}$$

The representation given by (2.8) and (2.9) is called the Schrödinger representation, and we call the unitary map T_n the Schrödinger map. When dealing with an irreducible representation of the operators, the Schrödinger map is denoted simply by T.

It follows from von Neumann's theorem, that if we demand that the family of operators $U[\mathbf{p}, \mathbf{q}]$ be

[11] F. Riesz and B. Sz.-Nagy, *Functional Analysis* (Frederick Ungar Publishing Company, New York, 1955), p. 380.

[12] H. Weyl, *The Theory of Groups and Quantum Mechanics* (Dover Publications, Inc., New York), p. 274.

[13] In his paper, von Neumann only required a condition much weaker than strong continuity, namely weak measurability. A one-parameter group of unitary operators $U[\tau]$ is said to be weakly measurable if, given any Ψ and Φ in $\mathfrak{H}$, the function $(\Phi, U[\tau]\Psi)$ is a Lebesgue-measurable function of τ. His theorem then showed that in this case weak measurability implies strong continuity. Since this is a technical point which seems to have little bearing on what follows, we have assumed strong continuity from the start.

irreducible, then up to an unitary equivalence, there is really only one representation. We always assume that we are dealing with an irreducible representation unless otherwise indicated. In Secs. 4 and 5, we discuss some of the consequences of using a reducible representation of $U[\mathbf{p}, \mathbf{q}]$, and indicate there our reasons for sticking to an irreducible representation.

3. THE DEMONSTRATION THAT $\mathfrak{S}$ IS AN OVERCOMPLETE FAMILY OF STATES

A. Local Density and Label Continuity

We now show that $\mathfrak{S}$ as defined in Sec. 2 is indeed an OFS. The first step in this direction is to prove that the vectors $\Phi[\mathbf{p}, \mathbf{q}] \in \mathfrak{H}$ are strongly continuous functions of their arguments $(\mathbf{p}, \mathbf{q})$. To this end we first establish

Lemma 3.1. The function $\mathfrak{U}[\mathbf{q}] = \prod_{\alpha=1}^{N} V_\alpha[q_\alpha]$, defined on R_N and taking on values in the set of unitary operators mapping $\mathfrak{H}$ onto itself, is a strongly continuous, N-parameter group of unitary operators. The same statement is true of the function $\mathfrak{W}[\mathbf{p}]$.

Proof: For fixed $\mathbf{q}$, $\mathfrak{U}[\mathbf{q}]$ is unitary since it is the product of unitary operators, and the properties $\mathfrak{U}[0] = I$ and $\mathfrak{U}[\mathbf{q}]\mathfrak{U}[\mathbf{q}'] = \mathfrak{U}[\mathbf{q} + \mathbf{q}']$ are a direct consequence of the facts that $V_\alpha[q_\alpha]$ is a one-parameter group for $\alpha = 1, \cdots, N$, and that the relations (2.5) are satisfied. The strong continuity follows from the fact that for arbitrary $\Psi \in \mathfrak{H}$, $\mathfrak{U}[\mathbf{q}]\Psi$, considered as a function of a single q_α is strongly continuous, uniformly with respect to the remaining $N - 1$ variables. Let $\mathbf{u}_\alpha$ be the vector with components $(\mathbf{u}_\alpha)_\beta = \delta_{\alpha\beta}$. Then because the $V_\alpha[q_\alpha]$ are unitary and commute with each other, we have

$$||\mathfrak{U}[\mathbf{q} + h\mathbf{u}_i]\Psi - \mathfrak{U}[\mathbf{q}]\Psi|| = ||V_i[q_i + h]\Psi - V_i[q_i]\Psi||. \tag{3.1}$$

Given any $\epsilon > 0$, we can pick $\delta > 0$ so that for all h satisfying $|h| < \delta$, the right-hand side of (3.1) is less than ϵ, and the choice of δ is independent of the remaining $N - 1$ coordinates. By adding and subtracting terms, and using the triangle inequality, $||\mathfrak{U}[\mathbf{q}']\Psi - \mathfrak{U}[\mathbf{q}]\Psi||$ can be bounded by N terms of the form (3.1). As an immediate consequence of the definition of the norm in R_N,

$$|\mathbf{q}| = \left(\sum_{\alpha=1}^{N} q_\alpha^2\right)^{\frac{1}{2}},$$

we find $|q_\alpha| \leq |\mathbf{q}|$. Hence given $\epsilon > 0$ we can find a $\delta > 0$ so that for all $\mathbf{q}'$ satisfying $|\mathbf{q} - \mathbf{q}'| < \delta$, we have $||\mathfrak{U}[\mathbf{q}']\Psi - \mathfrak{U}[\mathbf{q}]\Psi|| < \epsilon$. The conclusions of the theorem clearly hold for $\mathfrak{W}[\mathbf{p}]$ also.

With the aid of Lemma 3.1 we can now easily establish

Lemma 3.2. The vector valued function $\Phi[\mathbf{p}, \mathbf{q}]$, defined on $R_N \times R_N$, is strongly continuous in the product topology of $R_N \times R_N$.

Proof: Let $(\mathbf{p}_0, \mathbf{q}_0)$ and $\epsilon > 0$ be given. We show that there is a $\delta > 0$ such that for all $(\mathbf{p}, \mathbf{q})$ satisfying $|\mathbf{p} - \mathbf{p}_0| < \delta$, $|\mathbf{q} - \mathbf{q}_0| < \delta$ we have $||\Phi[\mathbf{p}_0, \mathbf{q}_0] - \Phi[\mathbf{p}, \mathbf{q}]|| < \epsilon$. For

$$\begin{aligned}
&||\Phi[\mathbf{p}_0, \mathbf{q}_0] - \Phi[\mathbf{p}, \mathbf{q}]|| \\
&\quad = ||\mathfrak{U}[\mathbf{q}_0]\mathfrak{W}[\mathbf{p}_0]\Phi_0 - \mathfrak{U}[\mathbf{q}]\mathfrak{W}[\mathbf{p}]\Phi_0|| \\
&\quad \leq ||\mathfrak{U}[\mathbf{q}_0]\mathfrak{W}[\mathbf{p}_0]\Phi_0 - \mathfrak{U}[\mathbf{q}]\mathfrak{W}[\mathbf{p}_0]\Phi_0|| \\
&\qquad + ||\mathfrak{U}[\mathbf{q}]\mathfrak{W}[\mathbf{p}_0]\Phi_0 - \mathfrak{U}[\mathbf{q}]\mathfrak{W}[\mathbf{p}]\Phi_0|| \\
&\quad = ||\mathfrak{U}[\mathbf{q}_0]\mathfrak{W}[\mathbf{p}_0]\Phi_0 - \mathfrak{U}[\mathbf{q}]\mathfrak{W}[\mathbf{p}_0]\Phi_0|| \\
&\qquad + ||\mathfrak{W}[\mathbf{p}_0]\Phi_0 - \mathfrak{W}[\mathbf{p}]\Phi_0||.
\end{aligned} \tag{3.2}$$

Since $\mathbf{p}_0$ is fixed, the strong continuity of $\mathfrak{U}[\mathbf{q}]$ and $\mathfrak{W}[\mathbf{p}]$ allows us to find a $\delta > 0$ such that when $|\mathbf{p} - \mathbf{p}_0| < \delta$ and $|\mathbf{q} - \mathbf{q}_0| < \delta$, each of the two terms on the right of (3.2) is less than $\frac{1}{2}\epsilon$.

It is clear that Lemma 3.2 shows that $\mathfrak{S}$ satisfies Postulate 1. Also, since strong continuity implies weak continuity, the mapping $\mathfrak{M}$ is weakly continuous, and Postulate 2 is satisfied by $\mathfrak{S}$.

B. Completeness and Resolution of Unity

To prove that $\mathfrak{S}$ satisfies Postulate 3, that is, to prove completeness and the existence of a resolution of the identity, is a less trivial job. In outline, our first step is to give meaning to vector-valued integrals of the form

$$\Psi = \int_{R_N}\int_{R_N} \psi(\mathbf{p}, \mathbf{q})\Phi[\mathbf{p}, \mathbf{q}] \frac{d^N p}{(2\pi)^{\frac{1}{2}N}} \frac{d^N q}{(2\pi)^{\frac{1}{2}N}}, \tag{3.3}$$

and to operator-valued integrals of the form

$$B = \int_{R_N}\int_{R_N} b(\mathbf{p}, \mathbf{q})\Phi[\mathbf{p}, \mathbf{q}]\Phi[\mathbf{p}, \mathbf{q}]^\dagger \times \frac{d^N p}{(2\pi)^{\frac{1}{2}N}} \frac{d^N q}{(2\pi)^{\frac{1}{2}N}}. \tag{3.4}$$

In (3.3) and (3.4), $\psi(\mathbf{p}, \mathbf{q})$ and $b(\mathbf{p}, \mathbf{q})$ are scalar-valued functions,

$$d^N p = \prod_{\alpha=1}^{N} dp_\alpha, \qquad d^N q = \prod_{\alpha=1}^{N} dq_\alpha,$$

and $\Phi[\mathbf{p}, \mathbf{q}]\Phi^\dagger[\mathbf{p}, \mathbf{q}]$ is the projection operator defined

for every $\Psi \in \mathfrak{H}$ by

$$\Phi[\mathbf{p}, \mathbf{q}]\Phi^{\dagger}[\mathbf{p}, \mathbf{q}]\Psi = (\Phi[\mathbf{p}, \mathbf{q}], \Psi)\Phi[\mathbf{p}, \mathbf{q}]. \tag{3.5}$$

In order to simplify the notation in the future, an integral sign with no limits always denotes an N-fold integral over R_N, and we often employ the abbreviation

$$d\mu(\mathbf{p}, \mathbf{q}) \equiv \frac{d^N p}{(2\pi)^{\frac{1}{2}N}} \frac{d^N q}{(2\pi)^{\frac{1}{2}N}}.$$

The second and last steps are to show that every vector $\Psi \in \mathfrak{H}$ can be represented in the form (3.3) and that the unit operator can be represented in the form (3.4).

In order to define integrals of the form (3.3) we must first study all functions of the form

$$\psi(\mathbf{p}, \mathbf{q}) = (\Phi[\mathbf{p}, \mathbf{q}], \Psi), \tag{3.6}$$

where Ψ is an arbitrary vector in $\mathfrak{H}$. Since $\Phi[\mathbf{p}, \mathbf{q}]$ is a strongly continuous function of $(\mathbf{p}, \mathbf{q})$ it is *a fortiori* weakly continuous. Hence $\psi(\mathbf{p}, \mathbf{q})$ is a continuous function of $(\mathbf{p}, \mathbf{q})$. However, it can be seen from the proof of Lemma 3.2 that, in general, $\psi(\mathbf{p}, \mathbf{q})$ is not uniformly continuous. An application of Schwartz's inequality shows that

$$|\psi(\mathbf{p}, \mathbf{q})| \leq ||\Psi||. \tag{3.7}$$

A more remarkable property of each of the functions (3.6) is that they are square integrable:

$$\iint |\psi(\mathbf{p}, \mathbf{q})|^2 \, d\mu(\mathbf{p}, \mathbf{q}) < \infty. \tag{3.8}$$

In order to prove this statement, we appeal to von Neumann's theorem. Let $\varphi_0(\mathbf{x})$ and $\psi(\mathbf{x})$ be the functions in $L^2(R_N)$ corresponding to Φ_0 and Ψ under the Schrödinger map T. Then the function corresponding to $\Phi[\mathbf{p}, \mathbf{q}]$ under T is

$$T\Phi[\mathbf{p}, \mathbf{q}] = e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0(\mathbf{x} - \mathbf{q}), \tag{3.9}$$

and since T is unitary, we must have

$$\psi(\mathbf{p}, \mathbf{q}) = \int e^{-i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0^*(\mathbf{x} - \mathbf{q})\psi(\mathbf{x}) \, d^N x. \tag{3.10}$$

We first note that after making the change of variables $\mathbf{x} \to \mathbf{x} + \mathbf{q}$ in the integral (3.10), $\psi(\mathbf{p}, \mathbf{q})$, for fixed $\mathbf{q}$, can be considered the Fourier transform of the integrable function

$$h(\mathbf{x}, \mathbf{q}) = (2\pi)^{\frac{1}{2}N}\varphi_0^*(\mathbf{x})\psi(\mathbf{x} + \mathbf{q}). \tag{3.11}$$

Furthermore, since $\varphi_0(\mathbf{x})$ and $\psi(\mathbf{x})$ are both in $L^2(R_N)$, $h(\mathbf{x}, \mathbf{q})$ is a measurable, square integrable function of $\mathbf{x}$ and $\mathbf{q}$ on $R_N \times R_N$. This can be proved as follows. Since $\psi(\mathbf{x})$ and $\varphi_0(\mathbf{x})$ are measurable in R_N, $h(\mathbf{x}, \mathbf{q})$ is measurable in $R_N \times R_N$. In addition, we have

$$\begin{aligned}\int |h(\mathbf{x}, \mathbf{q})|^2 \, d^N q &= (2\pi)^N \int |\varphi_0^*(\mathbf{x})|^2 \, |\psi(\mathbf{x} + \mathbf{q})|^2 \, d^N q \\ &= (2\pi)^N \, |\varphi_0^*(\mathbf{x})|^2 \int |\psi(\mathbf{r})|^2 \, d^N r \\ &= (2\pi)^N \, |\varphi_0^*(\mathbf{x})|^2 \, ||\Psi||^2. \end{aligned} \tag{3.12}$$

Hence it is clear that

$$\begin{aligned}\int d^N x \int |h(\mathbf{x}, \mathbf{q})|^2 \, d^N q &= (2\pi)^N \, ||\Psi||^2 \int |\varphi_0^*(\mathbf{x})|^2 \, d^N x \\ &= (2\pi)^N \, ||\Psi||^2 \, ||\Phi_0||^2 = (2\pi)^N \, ||\Psi||^2, \end{aligned} \tag{3.13}$$

and by the theorems of Tonelli and Fubini[14] it follows that $|h(\mathbf{x}, \mathbf{q})|^2$ is integrable, the integration can be performed in any order [thus yielding the result (3.13)], and also

$$\int |h(\mathbf{x}, \mathbf{q})|^2 \, d^N x$$

exists for all $\mathbf{q}$ except possibly for those forming a set of measure zero in R_N. Therefore, for all fixed $\mathbf{q}$, except those forming a set of measure zero, $\psi(\mathbf{p}, \mathbf{q})$ is the Fourier transform of a function in $L^1(R_N) \cap L^2(R_N)$; and so as a function of $\mathbf{p}$, $\psi(\mathbf{p}, \mathbf{q}) \in L^2(R_N)$ for almost all $\mathbf{q}$. Now we can employ Parseval's theorem[15] and for almost all $\mathbf{q}$ we obtain

$$\int |\psi(\mathbf{p}, \mathbf{q})|^2 \, d^N p = \int |h(\mathbf{k}, \mathbf{q})|^2 \, d^N k. \tag{3.14}$$

We have already proved, however, that the right-hand side of (3.14) is an integrable function of $\mathbf{q}$, and the value of the repeated integral is given in (3.13). A final appeal to Tonelli's theorem concludes the proof that $\psi(\mathbf{p}, \mathbf{q}) \in L^2(R_N \times R_N)$. As a by-product of these manipulations, we have shown that

$$\iint |\psi(\mathbf{p}, \mathbf{q})|^2 \, d\mu(\mathbf{p}, \mathbf{q}) = ||\Psi||^2. \tag{3.15}$$

If we set $\lambda(\mathbf{p}, \mathbf{q}) = (\Phi[\mathbf{p}, \mathbf{q}], \Lambda)$ for some $\Lambda \in \mathfrak{H}$, then the same methods quickly yield the result

$$\iint \psi^*(\mathbf{p}, \mathbf{q})\lambda(\mathbf{p}, \mathbf{q}) \, d\mu(\mathbf{p}, \mathbf{q}) = (\Psi, \Lambda). \tag{3.16}$$

[14] E. J. McShane, *Integration* (Princeton University Press, Princeton, New Jersey, 1944), pp. 137, 145.

[15] S. Bochner and K. Chandrasekharan, *Fourier Transforms* (Princeton University Press, Princeton, New Jersey, 1949), p. 120.

Let us summarize these results in

Theorem 3.1. If $\Psi \in \mathfrak{H}$ is an arbitrary vector, define the function $\psi(\mathbf{p}, \mathbf{q}) = (\Phi[\mathbf{p}, \mathbf{q}], \Psi)$. Then $\psi(\mathbf{p}, \mathbf{q})$ is a continuous, bounded function, with an upper bound $||\Psi||$. It is furthermore a square integrable function: the integral of its square, according to (3.15), is the square of the norm of Ψ. If $\lambda(\mathbf{p}, \mathbf{q}) = (\Phi[\mathbf{p}, \mathbf{q}], \Lambda)$ is another such function, then the inner product (Ψ, Λ) is given by (3.16).

We now give a meaning to the integral written down formally in (3.3). Let $\psi_0(\mathbf{p}, \mathbf{q})$ be an arbitrary, measurable, complex-valued, square integrable function. Then the integral

$$\Psi_0 = \iint \psi_0(\mathbf{p}, \mathbf{q})\Phi[\mathbf{p}, \mathbf{q}]\, d\mu(\mathbf{p}, \mathbf{q}) \qquad (3.17)$$

is defined to be that element of $\mathfrak{H}$ whose inner product, (Ψ_0, Ψ), with an arbitrary $\Psi \in \mathfrak{H}$, is given by

$$(\Psi_0, \Psi) = \iint \psi_0^*(\mathbf{p}, \mathbf{q})\psi(\mathbf{p}, \mathbf{q})\, d\mu(\mathbf{p}, \mathbf{q}), \qquad (3.18)$$

in which

$$\psi(\mathbf{p}, \mathbf{q}) = (\Phi[\mathbf{p}, \mathbf{q}], \Psi).$$

Theorem 3.1 ensures that such $\psi(\mathbf{p}, \mathbf{q})$ are square integrable so that the integral (3.18) exists for all $\Psi \in \mathfrak{H}$. The integral (3.17), defined by (3.18) is known as a Pettis integral and its properties have been investigated.[16] In particular, it has been shown that the existence of (3.18) for all $\Psi \in \mathfrak{H}$ is a necessary and sufficient condition that (3.17) be a uniquely defined vector in $\mathfrak{H}$.

We are now in a position to prove

Theorem 3.2. Let $\Psi \in \mathfrak{H}$. Then

$$\Psi = \iint (\Phi[\mathbf{p}, \mathbf{q}], \Psi)\Phi[\mathbf{p}, \mathbf{q}]\, d\mu(\mathbf{p}, \mathbf{q}). \qquad (3.19)$$

Proof: From Theorem 3.1, we know that $\psi(\mathbf{p}, \mathbf{q}) = (\Phi[\mathbf{p}, \mathbf{q}], \Psi)$ is square integrable, so the right-hand side of (3.19) is a well-defined vector in $\mathfrak{H}$. Denote the right-hand side of (3.19) provisionally by Ψ'. If $\Lambda \in \mathfrak{H}$ is any other vector, we have

$$(\Lambda, \Psi') = \iint (\Lambda, \Phi[\mathbf{p}, \mathbf{q}])(\Phi[\mathbf{p}, \mathbf{q}], \Psi)\, d\mu(\mathbf{p}, \mathbf{q}). \qquad (3.20)$$

Equation (3.20) is just the definition of the vector-valued integral. But according to Theorem 3.1, the right-hand side of (3.20) is just (Λ, Ψ). Therefore $(\Lambda, \Psi - \Psi') = 0$ for all $\Lambda \in \mathfrak{H}$, and hence $\Psi' = \Psi$.

An immediate and important consequence of Theorem 3.2 is that the closed subspace spanned by $\mathfrak{S}$ (the closure of the set consisting of all finite linear combinations of elements of $\mathfrak{S}$) is just $\mathfrak{H}$ itself. In other words $\mathfrak{S}$ is complete in $\mathfrak{H}$. Let $[\mathfrak{S}]$ denote the subspace spanned by $\mathfrak{S}$. Then if $[\mathfrak{S}] \neq \mathfrak{H}$, let $\Psi \in \mathfrak{H} - [\mathfrak{S}]$, where $\mathfrak{H} - [\mathfrak{S}]$ is the orthogonal complement of $[\mathfrak{S}]$ in $\mathfrak{H}$. This implies in particular that $(\Phi[\mathbf{p}, \mathbf{q}], \Psi) \equiv 0$, but then by Theorem 3.2, it follows that $\Psi = 0$. Therefore $[\mathfrak{S}] = \mathfrak{H}$. We list this result as

Lemma 3.3. $\mathfrak{S}$ is complete in $\mathfrak{H}$, that is, the closed subspace spanned by $\mathfrak{S}$ is $\mathfrak{H}$ itself.

It is now a straightforward matter to define a certain type of operator-valued integral and to prove the resolution of the identity formula. Let $b(\mathbf{p}, \mathbf{q})$ be a bounded, measurable function. Then the integral

$$B = \iint b(\mathbf{p}, \mathbf{q})\Phi[\mathbf{p}, \mathbf{q}]\Phi^\dagger[\mathbf{p}, \mathbf{q}]\, d\mu(\mathbf{p}, \mathbf{q}) \qquad (3.21)$$

is defined to be that operator which maps the vector $\Psi \in \mathfrak{H}$ onto

$$B\Psi = \iint b(\mathbf{p}, \mathbf{q})(\Phi[\mathbf{p}, \mathbf{q}], \Psi)\Phi[\mathbf{p}, \mathbf{q}]\, d\mu(\mathbf{p}, \mathbf{q}). \qquad (3.22)$$

Since $b(\mathbf{p}, \mathbf{q})$ is bounded, $B\Psi$ is well-defined. Further, if $\Psi, \Lambda \in \mathfrak{H}$ and α is any complex number, then because $(\Phi[\mathbf{p}, \mathbf{q}], \alpha\Psi + \Lambda) = \alpha(\Phi[\mathbf{p}, \mathbf{q}], \Psi) + (\Phi[\mathbf{p}, \mathbf{q}], \Lambda)$, and because the integral is a linear operation, it is clear that the operator B, as we have defined it, is a linear operator. In particular if we set $b(\mathbf{p}, \mathbf{q}) \equiv 1$, a reference to Theorem 3.2 shows that we obtain the unit operator. We list this result as

Lemma 3.4. The unit operator may be expressed as

$$I = \iint \Phi[\mathbf{p}, \mathbf{q}]\Phi^\dagger[\mathbf{p}, \mathbf{q}]\, d\mu(\mathbf{p}, \mathbf{q}). \qquad (3.23)$$

This completes the proof that the set $\mathfrak{S}$ is indeed an overcomplete family of states. Lemma 3.2 shows that $\mathfrak{S}$ satisfies Postulates 1 and 2, and Theorem 3.2 and Lemmas 3.3 and 3.4 demonstrate that $\mathfrak{S}$ satisfies Postulate 3.

4. THE CONTINUOUS REPRESENTATION OF $\mathfrak{H}$ GENERATED BY $\mathfrak{S}$

A. The Definition of $\mathfrak{C}$

In the first part of this section we use the results of Sec. 3 to construct the continuous representation, $\mathfrak{C}$, of $\mathfrak{H}$ generated by $\mathfrak{S}$. Since the properties of

[16] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups* (American Mathematical Society, Providence, Rhode Island, 1957), pp. 76–78.

the functions forming $\mathfrak{C}$ depend very strongly on the choice of the fiducial vector used in the definition of $\mathfrak{S}$, we are in effect going to construct a large number of function spaces.

We assume here, as in earlier sections of this paper, that we have an irreducible representation of the family of operators $U[\mathbf{p}, \mathbf{q}]$ defined on the separable Hilbert space $\mathfrak{H}$. We have seen that if $\Psi \in \mathfrak{H}$, then the relationship

$$C\Psi = \psi(\mathbf{p}, \mathbf{q}) = (\Phi[\mathbf{p}, \mathbf{q}], \Psi)$$

provides us with a map, C, of $\mathfrak{H}$ onto a class of bounded, continuous, square integrable functions. This family of functions is clearly a linear vector space, which we denote by $\mathfrak{C}$. As we soon see, $\mathfrak{C}$ depends strongly on the fiducial vector Φ_0 which enters in the definition of C. When it is necessary to emphasize this point, we refer to the space $\mathfrak{C}$ corresponding to the fiducial vector Φ_0.

We can supply $\mathfrak{C}$ with an inner product in a natural fashion. If $\psi'(\mathbf{p}, \mathbf{q})$ and $\psi(\mathbf{p}, \mathbf{q})$ are element of $\mathfrak{C}$, then set

$$(\psi', \psi)_c = \iint \psi'^*(\mathbf{p}, \mathbf{q})\psi(\mathbf{p}, \mathbf{q})\, d\mu(\mathbf{p}, \mathbf{q}). \tag{4.1}$$

From now on, all inner products and norms in $\mathfrak{C}$ have a subscript c appended in order to distinguish them from inner products and norms in $\mathfrak{H}$. The null element of $\mathfrak{C}$ is $\psi(\mathbf{p}, \mathbf{q}) \equiv 0$, and because of the continuity of the functions in $\mathfrak{C}$, $(\psi, \psi)_c = 0$ if and only if $\psi(\mathbf{p}, \mathbf{q}) \equiv 0$. There are thus no difficulties in showing that the inner product in $\mathfrak{C}$ defines a norm. Theorem 3.2 can now be used to show that the map C is one-one. For suppose $C\Psi' = C\Psi$, i.e., $\psi'(\mathbf{p}, \mathbf{q}) \equiv \psi(\mathbf{p}, \mathbf{q})$ and hence $0 \equiv (\Phi[\mathbf{p}, \mathbf{q}], \Psi' - \Psi)$. Then Theorem 3.2 shows that $\Psi' - \Psi = 0$. Furthermore, Theorem 3.1 shows that the map C is isometric, and Theorem 3.2 shows that C^{-1} is defined on all of $\mathfrak{C}$. It follows that $\mathfrak{C}$, supplied with the inner product (4.1), is the unitary image of $\mathfrak{H}$, and hence in particular it must be complete. We sum these results up.

Theorem 4.1. The set of functions, $\mathfrak{C}$, given by $\psi(\mathbf{p}, \mathbf{q}) = (\Phi[\mathbf{p}, \mathbf{q}], \Psi)$ for all $\Psi \in \mathfrak{H}$, is a family of bounded, continuous, and square integrable functions. When supplied with the inner product (4.1) the set $\mathfrak{C}$ is a complete Hilbert space which is unitarily equivalent to the original space $\mathfrak{H}$ under the unitary mapping C

$$C\Psi = (\Phi[\mathbf{p}, \mathbf{q}], \Psi) = \psi(\mathbf{p}, \mathbf{q}), \tag{4.2}$$

$$C^{-1}\psi(\mathbf{p}, \mathbf{q}) = \int \psi(\mathbf{p}, \mathbf{q})\Phi[\mathbf{p}, \mathbf{q}]\, d\mu(\mathbf{p}, \mathbf{q}) = \Psi. \tag{4.3}$$

$\mathfrak{C}$ is called a continuous representation of $\mathfrak{H}$.

An important characteristic of the continuous representations $\mathfrak{C}$ should be noted here. Namely, the elements of $\mathfrak{C}$ are single functions, and are *not* equivalence classes of functions, as are the elements of $L^2(R_N)$, for example.

As an application of this theorem, let us pick $\mathfrak{H}$ to be $L^2(R_N)$ and $\Phi[\mathbf{p}, \mathbf{q}] = e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0(\mathbf{x} - \mathbf{q})$. Then we get the transform pair

$$\psi(\mathbf{p}, \mathbf{q}) = \int e^{-i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0^*(\mathbf{x} - \mathbf{q})\psi(\mathbf{x})\, d^N x, \tag{4.4}$$

$$\psi(\mathbf{x}) = \int \psi(\mathbf{p}, \mathbf{q})e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0(\mathbf{x} - \mathbf{q})\, d\mu(\mathbf{p}, \mathbf{q}), \tag{4.5}$$

where the integral (4.5) is to be defined in the sense of a Pettis integral.

B. Differentiability and Growth Properties

We now show that by suitably choosing the fiducial vector Φ_0, each $\psi(\mathbf{p}, \mathbf{q}) \in \mathfrak{C}$ can be guaranteed a certain minimum number of derivatives. In order to do this, we first examine the infinitesimal generators of the groups $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$. A well-known theorem due to Stone[17] states that every strongly continuous, one-parameter group of unitary transformations $\{T[t]\}$ is generated by an infinitesimal transformation iA, where A is a self-adjoint transformation which is, in general, unbounded:

$$T[t] = e^{itA}, \quad iA = \lim_{t\to 0} \frac{1}{t}(T[t] - 1). \tag{4.6}$$

The limit in (4.6) is taken in the strong sense. Furthermore, an element $\Psi \in \mathfrak{H}$ is in the domain of A if and only if

$$\lim_{t\to 0} \left\{\frac{1}{t}(T[t] - 1)\Psi\right\} \tag{4.7}$$

exists in the strong sense.

We can apply Stone's theorem to our one-parameter groups and write

$$V_\alpha[q_\alpha] = e^{-iq_\alpha P_\alpha}, \qquad W_\alpha[p_\alpha] = e^{ip_\alpha Q_\alpha}. \tag{4.8}$$

(The minus sign is inserted in the application of Stone's theorem to $V_\alpha[q_\alpha]$ so that P_α will have its conventional sign in the Schrödinger representation.) It is well known that in the Schrödinger representation, the operators P_α and Q_α are given by[10]

$$P_\alpha = -i\,\partial/\partial x_\alpha, \qquad Q_\alpha = x_\alpha. \tag{4.9}$$

It is further known that the linear manifold $\mathfrak{D}$ consisting of all infinitely differentiable functions having compact support is dense in $L^2(R_N)$, and

[17] F. Riesz and B. Sz.-Nagy, *Functional Analysis* (Frederick Ungar Publishing Co., New York, 1955), p. 383.

that this set of functions is in the domain of all polynomials in the P_α and Q_α in the Schrödinger representation.[18] On $\mathfrak{D}$, the operators P_α and Q_α satisfy their well-known commutation relation.

The final fact we need regarding the infinitesimal transformations is the following. Let $\mathfrak{D}_{P_\alpha}$ and $\mathfrak{D}_{Q_\alpha}$ denote the domains of P_α and Q_α, $\alpha = 1, 2, \cdots, N$. Then $U[\mathbf{p}, \mathbf{q}]\mathfrak{D}_{P_\alpha} \subset \mathfrak{D}_{P_\alpha}$, $U[\mathbf{p}, \mathbf{q}]\mathfrak{D}_{Q_\alpha} \subset \mathfrak{D}_{Q_\alpha}$, $\alpha = 1, 2, \cdots, N$. Further, if $\Phi \in \mathfrak{D}_{P_\alpha}$ and $\Psi \in \mathfrak{D}_{Q_\alpha}$, then

$$P_\alpha U[\mathbf{p}, \mathbf{q}]\Phi = U[\mathbf{p}, \mathbf{q}](P_\alpha + p_\alpha I)\Phi, \tag{4.10a}$$

$$Q_\alpha U[\mathbf{p}, \mathbf{q}]\Psi = U[\mathbf{p}, \mathbf{q}](Q_\alpha + q_\alpha I)\Psi, \tag{4.10b}$$

where I denotes the unit operator.

To prove these results, we note first that if B is any bounded operator on $\mathfrak{H}$, and

$$iA = \lim_{t\to 0} \frac{1}{t}(T[t] - 1),$$

then

$$\lim_{t\to 0} B\frac{1}{t}(T[t] - 1) = B \lim_{t\to 0} \frac{1}{t}(T[t] - 1) = iBA. \tag{4.11}$$

For let Φ be in the domain of A, then

$$\|[B(1/t)(T[t] - 1) - iBA]\Phi\| \leq \|B\| \, \|[(1/t)(T[t] - 1) - iA]\Phi\| \to 0$$

as $t \to 0$. Now suppose $\Phi \in \mathfrak{D}_{P_\alpha}$, then by using the relations (2.5), we can write

$$\begin{aligned}(i/q'_\alpha)(V_\alpha[q'_\alpha] - 1)U[\mathbf{p}, \mathbf{q}]\Phi &= e^{-ip_\alpha q_\alpha'} U[\mathbf{p}, \mathbf{q}](i/q'_\alpha)(V_\alpha[q'_\alpha] - 1)\Phi \\ &\quad + U[\mathbf{p}, \mathbf{q}](i/q'_\alpha)(e^{-iq_\alpha' p_\alpha} - 1)\Phi. \end{aligned} \tag{4.12}$$

With the aid of the remark above and Stone's theorem, we see that the right-hand side of (4.12) has the strong limit $U[\mathbf{p}, \mathbf{q}](P_\alpha + p_\alpha I)\Phi$, as $q'_\alpha \to 0$. Again from Stone's theorem, it follows that $U[\mathbf{p}, \mathbf{q}]\Phi \in \mathfrak{D}_{P_\alpha}$, and that (4.10a) is true. The proof of the remaining part of the statement is the same. These last results are well known, although rigorous proofs of them do not seem to be readily available.[19]

Suppose, now, that the fiducial vector Φ_0 is in the domain of P_α and Q_α, $\alpha = 1, 2, \cdots, N$. Then for each $\psi(\mathbf{p}, \mathbf{q}) \in \mathfrak{C}$, all the first partial derivatives exist and are continuous. Consider, for example,

$$\begin{aligned}\frac{\partial \psi}{\partial p_\alpha} &= \lim_{\Delta p_\alpha \to 0} \frac{\psi(p_1, \cdots, p_{\alpha-1}, p_\alpha + \Delta p_\alpha, p_{\alpha+1}, \cdots, p_N, \mathbf{q}) - \psi(\mathbf{p}, \mathbf{q})}{\Delta p_\alpha} \\ &= \lim_{\Delta p_\alpha \to 0} \left(U[\mathbf{p}, \mathbf{q}]\left\{\frac{W_\alpha[\Delta p_\alpha] - 1}{\Delta p_\alpha}\right\}\Phi_0, \Psi\right).\end{aligned}$$

To get this last equality we have used the relations (2.5). From Stone's theorem, we know that

$$\lim_{\Delta p_\alpha \to 0} \frac{W_\alpha[\Delta p_\alpha] - 1}{\Delta p_\alpha} = iQ_\alpha$$

in the strong sense, and hence the limit exists weakly. Therefore

$$\partial\psi/\partial p_\alpha = -i(U[\mathbf{p}, \mathbf{q}]Q_\alpha\Phi_0, \Psi). \tag{4.13}$$

Notice that from Schwartz's inequality

$$|\partial\psi/\partial p_\alpha| \leq \|Q_\alpha\Phi_0\| \, \|\Psi\|. \tag{4.14}$$

A similar calculation yields, with the help of (4.10), the result

$$\begin{aligned}\partial\psi/\partial q_\alpha &= i(P_\alpha U[\mathbf{p}, \mathbf{q}]\Phi_0, \Psi) \\ &= i(U[\mathbf{p}, \mathbf{q}]P_\alpha\Phi_0, \Psi) + ip_\alpha\psi(\mathbf{p}, \mathbf{q}).\end{aligned} \tag{4.15}$$

Again note that

$$|\partial\psi/\partial q_\alpha| \leq \{\|P_\alpha\Phi_0\| + |p_\alpha|\} \, \|\Psi\|. \tag{4.16}$$

The continuity of the derivatives follows directly from the strong continuity of the family of operators $U[\mathbf{p}, \mathbf{q}]$.

It is clear that even if Φ_0 is not in the domain of all second-degree polynomials in P_α and Q_α some of the functions $\psi(\mathbf{p}, \mathbf{q})$ still possess higher-order continuous derivatives besides the first-order derivatives guaranteed them by our choice of Φ_0. In particular, when Ψ is in the domain of all P_α and Q_α, then the corresponding $\psi(\mathbf{p}, \mathbf{q})$ possesses continuous second-order derivatives. On the other hand, if the fiducial vector Φ_0 does not belong to the domain of any second-degree polynomial in P_α and Q_α, then there exist Ψ which are not in the domain of any P_α or Q_α, and consequently the corresponding $\psi(\mathbf{p}, \mathbf{q})$ does not possess continuous second derivatives.

These considerations can easily be extended to fiducial vectors, Φ_0, which are in the domain of all polynomials in the P_α and Q_α of degree $\leq n$. We state these results as a theorem.

Theorem 4.2. Let the fiducial vector Φ_0 be in the

[18] L. Schwartz, *Theorie Des Distributions* (Hermann & Cie, Paris, 1957), Vol. I. The fact that $\mathfrak{D}$ is dense in $L^2(R_N)$ can be deduced from Theorem I, p. 22. A more extended discussion of this point is given in the Princeton Thesis of J. S. Lew, "The Structure of Representations of the Canonical Commutation Relations," Princeton University, Princeton, New Jersey, 1960 (unpublished).

[19] J. S. Lew, Ref. 18, and J. R. Klauder, Ref. 2.

domain of all polynomials in the P_α and Q_α of degree $\leq n$, and let $\mathfrak{C}$ be the continuous representation of $\mathfrak{H}$ corresponding to Φ_0. Then all the mixed partial derivatives, of order up to and including n, of each $\psi(\mathbf{p}, \mathbf{q}) \in \mathfrak{C}$ exist. Each such derivative is a continuous function and is bounded by a polynomial in $|p_1|, \cdots, |p_N|$ of degree at most n.

We can state an even stronger result:

Theorem 4.3. There exist in $\mathfrak{H}$ dense linear manifolds of vectors which are in the domain of all polynomials in P_α and Q_α, $\alpha = 1, 2, \cdots, N$; for example, the image $T^{-1}\mathfrak{D}$ of all infinitely differentiable functions of compact support in $L^2(R_N)$ under the inverse of the Schrödinger map T^{-1}. If we choose the fiducial vector Φ_0 from this set, then all the functions, $\psi(\mathbf{p}, \mathbf{q})$, of the space $\mathfrak{C}$ corresponding to Φ_0, possess mixed continuous derivatives of *all* orders.

C. Connection to Representations by Entire Functions

It is now natural to ask whether the fiducial vector Φ can be chosen so that the corresponding continuous representation $\mathfrak{C}$ exhibits some special relationship with a Hilbert space of entire functions. [For simplicity in what follows, we consider only the case $N = 1$, i.e., $\psi(p, q)$ is a function of just two scalar variables, p and q.] We cannot require the elements of $\mathfrak{C}$ themselves to be entire functions of $q \pm ip$, for each $\psi(p, q) \in \mathfrak{C}$ is bounded, whereas by Liouville's theorem a nonconstant entire function cannot be bounded.

However, it has been pointed out in Part II that Φ_0 can be chosen so that up to a common multiplicative function each $\psi(p, q)$ is an entire function of $q - ip$. We now extend this investigation and determine all those continuously differentiable functions $a(p, q)$ and associated fiducial vectors Φ_0 so that for fixed $a(p, q)$ and Φ_0, the product $a(p, q)\psi(p, q)$ is an entire function of $q - ip$, or possibly of $q + ip$, for all $\psi(p, q) \in \mathfrak{C}$. Our chief tool is the fact that if a function $f(x + iy)$ is defined at every point $x + iy$, then the necessary and sufficient conditions that it be an entire function are that its real and imaginary parts have continuous, first-order partial derivatives at each point, and that at each point it satisfies the Cauchy–Reimann equations.[20]

[20] E. C. Titchmarch, *The Theory of Functions* (Oxford University Press, Oxford, England, 1949), p. 68. The condition that the derivatives be continuous is actually superfluous. Cf. S. Saks, *Theory of the Integral* (G. E. Stechert, New York, 1937), p. 199.

We first show that one *cannot* choose $a(p, q)$ and Φ_0 so that $a(p, q)\psi(p, q)$ is an entire function of $q + ip$ for all $\psi(p, q) \in \mathfrak{C}$. If $a\psi$ is an entire function of $q + ip$, it must satisfy the Cauchy–Reimann equations

$$\partial(a\psi)/\partial q + i\,\partial(a\psi)/\partial p = 0. \tag{4.17}$$

With the aid of Eqs. (4.13) and (4.15) we see that (4.17) can be written as

$$(U[p, q]\{(\partial a^*/\partial q - i\,\partial a^*/\partial p)\Phi_0 + a^*(Q - iP - ipI)\Phi_0\}, \Psi) = 0, \tag{4.18}$$

where a^* is the complex conjugate of a. Since (4.18) must hold for all $\Psi \in \mathfrak{H}$ and since $U[p, q]$ has an inverse for all (p, q), Eq. (4.18) is equivalent to

$$(\partial a^*/\partial q - i\,\partial a^*/\partial p)\Phi_0 + a^*(Q - iP - ipI)\Phi_0 = 0. \tag{4.19}$$

Since Φ_0 is independent of p and q and a cannot vanish throughout a region, Eq. (4.19) is equivalent to the two equations

$$(Q - iP + \lambda I)\Phi_0 = 0, \tag{4.20}$$

$$\partial a/\partial q + i\,\partial a/\partial p + (ip - \lambda^*)a = 0, \tag{4.21}$$

where $\lambda = \mu + i\nu$ is an arbitrary complex number, independent of p and q. If we go to the Schrödinger representation, Eq. (4.20) reads

$$(d/dx - x - \lambda)\varphi_0(x) = 0, \tag{4.22}$$

which has the solution $\varphi_0(x) = Ae^{\frac{1}{2}(x+\lambda)^2}$. However, this function is not an element of $L^2(R)$ for any value of λ, and we can conclude that Eq. (4.20) has no solution in $\mathfrak{H}$.

On the other hand, if $a\psi$ is to be an entire function of $q - ip$, the Cauchy–Riemann equations become

$$\partial(a\psi)/\partial p + i\,\partial(a\psi)/\partial q = 0, \tag{4.23}$$

or equivalently

$$(iQ - P + \lambda I)\Phi_0 = 0, \tag{4.24}$$

$$\partial a/\partial p + i\,\partial a/\partial q - (p + \lambda^*)a = 0, \tag{4.25}$$

where again $\lambda = \mu + i\nu$ is a complex constant independent of p and q. Let $\Phi_{0,\lambda}$ denote a solution of (4.24) and define the vector $\Phi_{0,0}$ by the equation

$$\Phi_{0,\lambda} = U[\mu, -\nu]\Phi_{0,0}, \tag{4.26}$$

then if we substitute (4.26) into (4.24) and use the commutation relations (2.5), we see that $\Phi_{0,0}$ must satisfy

$$(iQ - P)\Phi_{0,0} = 0. \tag{4.27}$$

This is the well-known equation for the ground state of an harmonic oscillator, and in the Schrödinger representation the unit vector solution of (4.27) is[21]

$$\varphi_{0,0}(x) = \pi^{-\frac{1}{4}} e^{-\frac{1}{2}x^2}. \tag{4.28}$$

This solution is unique up to a constant multiple of modulus 1, and so (4.26) gives the essentially unique solution of (4.24). Since this fiducial vector, and hence any $\Phi_{0,\lambda}$, is in the domain of all polynomials in P and Q, it follows that $\mathfrak{C}$ consists of functions having continuous derivatives of all orders.

It is further easily verified that

$$b_\lambda(p, q) = \exp\{\tfrac{1}{4}[(p + \lambda^*)^2 + q^2] - i\tfrac{1}{2}(p + \lambda^*)q + i\nu\mu - \tfrac{1}{2}\mu^2 - i\lambda(q - ip)\} \tag{4.29}$$

is a particular solution of (4.25). It should be noted that $b_\lambda(p, q)$ never vanishes, satisfies Eq. (4.25) for all values of p and q, and has continuous derivatives of all orders. An arbitrary solution of (4.25) can be written in the form $a_\lambda(p, q) = b_\lambda(p, q)c(p, q)$, where $c(p, q)$ satisfies the Cauchy–Reimann equations (4.23). In other words, every continuously differentiable solution of (4.25) is a product of $b_\lambda(p, q)$ times an entire function $q - ip$. If we denote by $\mathfrak{C}_\lambda$ the continuous representation of $\mathfrak{H}$ corresponding to the fiducial vector $\Phi_{0,\lambda}$ and $\psi_\lambda(p, q)$ the elements of this space, then we can conclude from the above results [by setting $c(p, q) \equiv 1$] that every $\psi_\lambda(p, q) \in \mathfrak{C}_\lambda$ can be written as

$$\psi_\lambda(p, q) = b_\lambda^{-1}(p, q) f_\lambda(q - ip). \tag{4.30}$$

In (4.30), $f_\lambda(q - ip)$ is an entire function because it satisfies the Cauchy–Riemann equations everywhere and has continuous derivatives everywhere.

To determine the relationship between the spaces $\mathfrak{C}_\lambda$ for different values of λ, we employ (4.26). If $\Psi \in \mathfrak{H}$ and $\psi_\lambda(p, q)$ is the corresponding function in $\mathfrak{C}_\lambda$, then

$$\begin{aligned}\psi_\lambda(p, q) &= (U[p, q]U[\mu, -\nu]\Phi_{0,0}, \Psi) \\ &= e^{i\nu p}(U[p + \mu, q - \nu]\Phi_{0,0}, \Psi) \\ &= e^{i\nu p}\psi_0(p + \mu, q - \nu).\end{aligned} \tag{4.31}$$

Thus every function in $\mathfrak{C}_\lambda$ is obtained from the corresponding function in $\mathfrak{C}_0$ by a translation and multiplication by $e^{i\nu p}$. A further calculation shows that

$$\beta_\lambda(p, q) = e^{-i\nu p - \frac{1}{2}|\lambda|^2 - i\lambda(q - ip)}\beta_0(p + \mu, q - \nu). \tag{4.32}$$

If we combine (4.30), (4.31), and (4.32) we see that the entire functions $f_\lambda(q - ip)$ and $f_0(q - ip)$ satisfy

$$f_\lambda(q - ip) = e^{-i\lambda(q - ip) - \frac{1}{2}|\lambda|^2} f_0(q - ip - i\lambda^*). \tag{4.33}$$

Now $\mathfrak{C}_0$ can be considered as a space of entire functions, $f_0(q - ip)$, with an inner product given by

$$(2\pi)^{-1}\iint e^{-\frac{1}{2}(p^2 + q^2)} f_0(q - ip)^* g_0(q - ip)\, dq\, dp. \tag{4.34}$$

However, this is just the Hilbert space of entire functions, $\mathfrak{F}$, studied by Segal[3] and Bargmann.[4] Furthermore, it has been shown by Bargmann[4] that Eq. (4.33) relating $f_\lambda(q - ip)$ to $f_0(q - ip)$ is a unitary map of $\mathfrak{F}$ onto itself. In other words the class of functions $f_\lambda(q - ip)$ is identical with the class of functions $f_0(q - ip)$. Thus while the choice of the separation constant λ superfically changes the class of functions $\mathfrak{C}_\lambda$, the underlying class of entire functions is the same in every case. Also, the space $\mathfrak{F}$ is the only space of entire functions which arises in a natural fashion from a continuous representation of the type discussed in this paper.

D. Reducibility and Irreducibility of $U[p, q]$

We conclude Sec. 4 with a brief discussion of the effects of choosing a reducible representation of the operator family $U[\mathbf{p}, \mathbf{q}]$. Let us assume that we can write $\mathfrak{H}$ as the direct sum of two orthogonal subspaces, $\mathfrak{H} = \mathfrak{H}_1 \oplus \mathfrak{H}_2$, such that $\mathfrak{H}_1$ and $\mathfrak{H}_2$ are both invariant under $U[\mathbf{p}, \mathbf{q}]$ for all $(\mathbf{p}, \mathbf{q})$. If we pick the fiducial vector Φ_0 in $\mathfrak{H}_1$, call it Φ_{01}, then for all $\Psi \in \mathfrak{H}_2$, we clearly have $(U[\mathbf{p}, \mathbf{q}]\Phi_{01}, \Psi) \equiv 0$. In fact, it can be seen that the continuous representation arising from this choice of Φ_0 is just the unitary image of the subspace $\mathfrak{H}_1$, and not the whole space $\mathfrak{H}$. If we pick as the fiducial vector $\Phi_0 = \Phi_{01} \oplus \Phi_{02}$, the direct sum of $\Phi_{01} \in \mathfrak{H}_1$ and $\Phi_{02} \in \mathfrak{H}_2$, then the resulting space of functions, $\psi(\mathbf{p}, \mathbf{q})$, does not generally form a continuous representation as we use the term, because when supplied with the inner product (4.1) it is no longer a Hilbert space isometric with $\mathfrak{H}_1 \oplus \mathfrak{H}_2$ when the isometry is defined by C and C^{-1} in (4.2) and (4.3), respectively. (In special cases, this isometry may be secured merely by scaling the measure $d\mu$; this is touched on in Sec. 5.) Since it can be shown that in most cases this space is isomorphic to the direct sum of the continuous representations of $\mathfrak{H}_1$ and $\mathfrak{H}_2$, it suffices to consider only continuous representations generated by irreducible representations of $U[\mathbf{p}, \mathbf{q}]$.

[21] In the Schrödinger representation, $\Phi_{0,\lambda}$ is given by $\varphi_{0,\lambda}(x) = \pi^{-\frac{1}{4}} e^{i\mu(x+\nu) - \frac{1}{2}(x+\nu)^2}$.

5. CONTINUOUS REPRESENTATIONS AND REPRODUCING KERNELS

A. Definition and Properties of the Reproducing Kernel

Let $\mathfrak{C}$ be a continuous representation of $\mathfrak{H}$ corresponding to the fiducial vector Φ_0. We associate with $\mathfrak{C}$ a function, $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$, called the reproducing kernel, which is defined as follows:

$$\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}) = (U[\mathbf{p}', \mathbf{q}']\Phi_0, U[\mathbf{p}, \mathbf{q}]\Phi_0). \tag{5.1}$$

For fixed $(\mathbf{p}, \mathbf{q})$, $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ is an element of $\mathfrak{C}$ when considered as a function of $(\mathbf{p}', \mathbf{q}')$, as follows directly from the definition, (5.1). From this last statement we can deduce the property of the reproducing kernel which accounts for its name. Let $\psi(\mathbf{p}', \mathbf{q}') = (U[\mathbf{p}', \mathbf{q}']\Phi_0, \Psi) \in \mathfrak{C}$. Then for every $(\mathbf{p}, \mathbf{q})$ it follows from Theorem 3.1 that

$$\begin{aligned}(\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}), \psi(\mathbf{p}', \mathbf{q}'))_c &= \iint \mathcal{K}^*(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})\psi(\mathbf{p}', \mathbf{q}')\, d\mu(\mathbf{p}', \mathbf{q}') \\ &= (U[\mathbf{p}, \mathbf{q}]\Phi_0, \Psi) = \psi(\mathbf{p}, \mathbf{q}). \end{aligned} \tag{5.2}$$

The same reasoning yields the relation

$$\iint \mathcal{K}^*(\mathbf{p}'', \mathbf{q}''; \mathbf{p}', \mathbf{q}')\mathcal{K}(\mathbf{p}'', \mathbf{q}''; \mathbf{p}, \mathbf{q})\, d\mu(\mathbf{p}'', \mathbf{q}'') = \mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}). \tag{5.3}$$

If $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ is viewed as a collection of functions in $\mathfrak{C}$, each function corresponding to a choice of $(\mathbf{p}, \mathbf{q})$, then it is seen that $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ is just the image in $\mathfrak{C}$ of the OFS $U[\mathbf{p}, \mathbf{q}]\Phi_0$.

In addition, the reproducing kernel has a number of special properties of which we list several of the more important.

1. If α_k, $k = 1, \cdots, M$, is an arbitrary, finite set of M complex numbers and $(\mathbf{p}_k, \mathbf{q}_k)$, $k=1, \cdots, M$, is an arbitrary set of M points in $R_N \times R_N$, then

$$\sum_{j=1}^{M}\sum_{k=1}^{M} \alpha_j^*\alpha_k\mathcal{K}(\mathbf{p}_j, \mathbf{q}_j; \mathbf{p}_k, \mathbf{q}_k) \geq 0. \tag{5.4a}$$

The sum is (5.4a) is just $||\sum_{k=1}^{M} \alpha_k U[\mathbf{p}_k, \mathbf{q}_k]\Phi_0||^2$, which is obviously nonnegative. From the inequality (5.4a) the following relations can be deduced[8]:

$$\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = \mathcal{K}^*(\mathbf{r}, \mathbf{s}; \mathbf{p}, \mathbf{q}), \tag{5.4b}$$

$$\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}, \mathbf{q}) \geq 0. \tag{5.4c}$$

2. In addition to (5.4c), our kernel function satisfies the stronger condition

$$\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}, \mathbf{q}) \equiv 1. \tag{5.5}$$

This follows trivially from the definition of $\mathcal{K}$.

3. $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ is a continuous function of its variables. It is continuous in $(\mathbf{p}', \mathbf{q}')$ uniformly with respect to $(\mathbf{p}, \mathbf{q})$, and similarly it is continuous in $(\mathbf{p}, \mathbf{q})$ uniformly with respect to $(\mathbf{p}', \mathbf{q}')$. This is a consequence of the strong continuity of the family of vectors $\Phi[\mathbf{p}, \mathbf{q}] = U[\mathbf{p}, \mathbf{q}]\Phi_0$, a result which we proved in Lemma 3.2.

4. As a consequence of the definition (5.1) and the commutation relations (2.5), $\mathcal{K}$ satisfies the relation

$$\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}) \equiv e^{i\mathbf{p}\cdot(\mathbf{q}'-\mathbf{q})}\mathcal{K}(\mathbf{p}' - \mathbf{p}, \mathbf{q}' - \mathbf{q}; 0, 0). \tag{5.6}$$

Since $\mathcal{K}(\mathbf{p}, \mathbf{q}; 0, 0)$ is a function in $\mathfrak{C}$, it is square integrable. Thus, up to the phase factor, the kernel function is a square integrable, difference kernel.

5. The arguments used in Sec. 4 can be applied equally well to $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ to show that by suitably choosing the fiducial vector, the kernel function will have continuous derivatives. In particular, a dense set of Φ_0 exist such that the kernel function will be infinitely differentiable.

B. Construction of the Continuous Representation from the Kernel Function

The significance of the kernel function lies in the fact that it completely determines the corresponding continuous representation. That this should be true is not at all surprising. In fact, a general theory of Hilbert spaces of functions that possess a reproducing kernel has been developed by Aronszajn,[8,9] and our continuous representations are a special case of his spaces. However, the continuous representations are so rich in structure, that they possess many important additional properties not shared by all of Aronszajn's spaces.

Aronszajn has shown that starting with a function $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ that satisfies only condition (5.4a), it is possible to construct a Hilbert space of functions for which $\mathcal{K}$ is the reproducing kernel. We outline Aronszajn's construction, and then afterwards we show how the other properties of $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$, namely Eq. (5.3) and properties 1 through 4 listed above, provide the space of functions so constructed with those properties which characterize it as a continuous representation.

Let $\mathfrak{C}_1$ be the set of all functions of the form

$$\psi(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{M} \alpha_k\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}_k, \mathbf{q}_k), \tag{5.7}$$

where α_k is an arbitrary, finite set of M complex

numbers, and $(\mathbf{p}_k, \mathbf{q}_k)$ is an arbitrary, finite set of M points in $R_N \times R_N$. $\mathfrak{C}_1$ is clearly a linear vector space. The null element of $\mathfrak{C}_1$ is defined as the function $\psi(\mathbf{p}, \mathbf{q}) \equiv 0$. This last requirement ensures that the elements of $\mathfrak{C}_1$ (and its completion $\mathfrak{C}$) are functions and not equivalence classes of functions. This point has been emphasized earlier, and it might be noted here that this is a common characteristic of Aronszajn's function spaces.

If

$$\varphi(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{M'} \beta_k \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}'_k, \mathbf{q}'_k), \tag{5.8}$$

then the inner product of φ and ψ is defined to be

$$(\psi, \varphi)_c = \sum_{k=1}^{M} \sum_{l=1}^{M'} \alpha_k^* \beta_l \mathcal{K}(\mathbf{p}_k, \mathbf{q}_k; \mathbf{p}'_l, \mathbf{q}'_l). \tag{5.9}$$

Since $\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}', \mathbf{q}')$ is an element of $\mathfrak{C}_1$ for fixed $(\mathbf{p}', \mathbf{q}')$, Eqs. (5.7) and (5.9) together yield the reproducing property

$$\psi(\mathbf{p}', \mathbf{q}') = (\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}', \mathbf{q}'), \psi(\mathbf{p}, \mathbf{q}))_c. \tag{5.10}$$

Since the kernel satisfies (5.4a), the inner product $\mathfrak{C}_1$ defines a nonnegative norm, $||\psi||_c^2 = (\psi, \psi)_c$. The norm of ψ can be expressed as

$$||\psi||_c^2 = \sum_{j=1}^{M} \sum_{k=1}^{M} \alpha_j^* \alpha_k \mathcal{K}(\mathbf{p}_j, \mathbf{q}_j; \mathbf{p}_k, \mathbf{q}_k) = \sum_{j=1}^{M} \alpha_j^* \psi(\mathbf{p}_j, \mathbf{q}_j), \tag{5.11}$$

which shows that if $\psi(\mathbf{p}, \mathbf{q}) \equiv 0$, then $||\psi||_c = 0$. Conversely, an application of Schwartz's inequality to (5.10) yields the inequality

$$|\psi(\mathbf{p}, \mathbf{q})| \leq ||\psi||_c \, [\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}, \mathbf{q})]^{\frac{1}{2}}, \tag{5.12}$$

which shows that if $||\psi||_c = 0$, then $\psi(\mathbf{p}, \mathbf{q}) \equiv 0$. $\mathfrak{C}_1$ thus has all the properties of a Hilbert space except that in general it is not complete, i.e., it is a pre-Hilbert space.

The standard technique of completing a pre-Hilbert space is to embed it in the Hilbert space consisting of equivalence classes of Cauchy sequences from the original space.[22] However, for pre-Hilbert spaces such as $\mathfrak{C}_1$, Aronszajn has shown that to each Cauchy sequence of $\mathfrak{C}_1$ there corresponds a unique function. If $\mathfrak{C}_1$ is enlarged by the addition of this set of functions, the resulting space is complete. We indicate how this is done, and refer the interested reader to Aronszajn's paper for full details.

Let $\psi_n(\mathbf{p}, \mathbf{q})$ be a Cauchy sequence in $\mathfrak{C}_1$, i.e., $||\psi_n - \psi_m||_c \to 0$ as $n, m \to \infty$. Then, applying inequality (5.12) to $\psi_n(\mathbf{p}, \mathbf{q}) - \psi_m(\mathbf{p}, \mathbf{q})$, we obtain

$$|\psi_n(\mathbf{p}, \mathbf{q}) - \psi_m(\mathbf{p}, \mathbf{q})| \leq ||\psi_n - \psi_m||_c \, [\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}, \mathbf{q})]^{\frac{1}{2}}, \tag{5.13}$$

which shows that the sequence $\psi_n(\mathbf{p}, \mathbf{q})$ converges pointwise to a function $\psi(\mathbf{p}, \mathbf{q})$. Equivalent Cauchy sequences yield the same limit function. An application of the triangle inequality yields

$$|\,||\psi_n||_c - ||\psi_m||_c\,| \leq ||\psi_n - \psi_m||_c, \tag{5.14}$$

which shows the existence of the $\lim_{n\to\infty} ||\psi_n||_c$. The norm of $\psi(\mathbf{p}, \mathbf{q})$ is then defined to be

$$||\psi||_c = \lim_{n\to\infty} ||\psi_n||_c. \tag{5.15}$$

The space $\mathfrak{C}_1$, augmented by the addition of all limit functions $\psi(\mathbf{p}, \mathbf{q})$ will be denoted by $\mathfrak{C}$. It is a Hilbert space of functions allowing the kernel function $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$.

It is clear from this outline of the construction of $\mathfrak{C}$ that until further properties of the kernel function are specified besides (5.4a), little can be said about the functions comprising $\mathfrak{C}$. We now examine the consequences for $\mathfrak{C}$ of the additional properties enjoyed by our kernel function.

First, conditions 2 and 3 imply that all the functions of $\mathfrak{C}$ are continuous and bounded. For since $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ is a continuous function of $(\mathbf{p}', \mathbf{q}')$ for fixed $(\mathbf{p}, \mathbf{q})$, all the functions of $\mathfrak{C}_1$ are clearly continuous. Further, with the aid of Eq. (5.5), inequality (5.13) can be sharpened to

$$|\psi_n(\mathbf{p}, \mathbf{q}) - \psi_m(\mathbf{p}, \mathbf{q})| \leq ||\psi_n - \psi_m||_c. \tag{5.16}$$

This implies that the Cauchy sequence of continuous functions $\psi_n(\mathbf{p}, \mathbf{q})$ converges uniformly to the limit function $\psi(\mathbf{p}, \mathbf{q})$, which therefore also must be a continuous function. The sharpened form of inequality (5.12) reads

$$|\psi(\mathbf{p}, \mathbf{q})| \leq ||\psi||_c. \tag{5.17}$$

This holds for all $\psi(\mathbf{p}, \mathbf{q}) \in \mathfrak{C}_1$ and from the construction of $\mathfrak{C}$ it holds for every element of $\mathfrak{C}$ as well. Since $R_N \times R_N$ is a separable space, a theorem of Aronszajn's says that $\mathfrak{C}$, being a space of continuous functions, must be separable.[23]

We next show that Eq. (5.6) allows us to get simple representations of the one-parameter groups of unitary operators $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$, $\alpha = 1, \cdots, N$, which satisfy the commutation relations (2.5), and from which in turn we can reconstruct the kernel

[22] S. Bochner and K. Chandrasekharan, Ref. 15, p. 86.

[23] N. Aronszajn, Ref. 8, p. 141.

function. Let the operators $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$ be defined as follows:

$$(V_\alpha[q_\alpha]\psi)(\mathbf{p}', \mathbf{q}') = \psi(\mathbf{p}', \mathbf{q}' - q_\alpha \mathbf{u}_\alpha), \tag{5.18}$$

$$(W_\alpha[p_\alpha]\psi)(\mathbf{p}', \mathbf{q}') = e^{ip_\alpha q_\alpha'}\psi(\mathbf{p}' - p_\alpha \mathbf{u}_\alpha, \mathbf{q}'), \tag{5.19}$$

where $\mathbf{u}_\alpha$ is the vector with components $(\mathbf{u}_\alpha)_\beta = \delta_{\alpha\beta}$. In the first place, it can be verified that $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$ are linear maps of $\mathfrak{C}_1$ onto $\mathfrak{C}_1$. Two easily derived consequences of Eq. (5.6) are the identities

$$\mathcal{K}(\mathbf{p}, \mathbf{q} - \mathbf{r}; \mathbf{p}', \mathbf{q}') \equiv \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}', \mathbf{q}' + \mathbf{r}), \tag{5.20}$$

$$e^{i\mathbf{s}\cdot\mathbf{q}}\mathcal{K}(\mathbf{p} - \mathbf{s}, \mathbf{q}; \mathbf{p}', \mathbf{q}') \equiv e^{i\mathbf{s}\cdot\mathbf{q}'}\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}' + \mathbf{s}, \mathbf{q}'). \tag{5.21}$$

Then if $\psi \in \mathfrak{C}_1$, we have

$$\begin{aligned}(V_\alpha[q_\alpha]\psi)(\mathbf{p}', \mathbf{q}') &= \psi(\mathbf{p}', \mathbf{q}' - q_\alpha \mathbf{u}_\alpha) \\ &= \sum_{j=1}^{N} \beta_j \mathcal{K}(\mathbf{p}', \mathbf{q}' - q_\alpha \mathbf{u}_\alpha; \mathbf{p}_j, \mathbf{q}_j) \\ &= \sum_{j=1}^{N} \beta_j \mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}_j, \mathbf{q}_j + q_\alpha \mathbf{u}_\alpha) \in \mathfrak{C}_1,\end{aligned} \tag{5.22}$$

$$\begin{aligned}(W_\alpha[p_\alpha]\psi)(\mathbf{p}', \mathbf{q}') &= e^{ip_\alpha q_\alpha'}\psi(\mathbf{p}' - p_\alpha \mathbf{u}_\alpha, \mathbf{q}') \\ &= e^{ip_\alpha q_\alpha'} \sum_{j=1}^{N} \beta_j \mathcal{K}(\mathbf{p}' - p_\alpha \mathbf{u}_\alpha, \mathbf{q}'; \mathbf{p}_j, \mathbf{q}_j) \\ &= \sum_{j=1}^{N} \beta_j e^{ip_\alpha (q_j)_\alpha} \mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}_j + p_\alpha \mathbf{u}_\alpha, \mathbf{q}_j) \in \mathfrak{C}_1.\end{aligned} \tag{5.23}$$

It is now readily verified that on $\mathfrak{C}_1$, $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$ form $2N$, weakly continuous, one-parameter groups of unitary operators, $\alpha = 1, \cdots, N$, which in addition satisfy the commutation relations (2.5). The weak continuity follows directly from the form of the inner product on $\mathfrak{C}_1$, (5.8), and the continuity of the kernel function.

Since $\mathfrak{C}_1$ is dense in $\mathfrak{C}$ and the operators $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$ are bounded, their unique extension to bounded operators defined on all of $\mathfrak{C}$ is immediately obtained by a standard construction.[24] Furthermore, the extended operators have the same bounds as the unextended operators. However, since each element of $\mathfrak{C}$ is the limit of a Cauchy sequence of elements in $\mathfrak{C}_1$, which converges to its limit pointwise, as well as in norm, (5.18) and (5.19) are seen to give the definition of $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$ everywhere on $\mathfrak{C}$. The extended operators, $V_\alpha[q_\alpha]$ and $W_\alpha[p_\alpha]$, are again readily seen to form two, weakly continuous, one-parameter groups of unitary operators which satisfy the commutation relations (2.5). In addition, in the case of a one-parameter group of unitary operators, weak continuity is known to imply strong continuity.[25]

Finally let $U[\mathbf{p}, \mathbf{q}] = \prod_{\alpha=1}^{N} V_\alpha[q_\alpha]W_\alpha[p_\alpha]$, and let

$$\varphi_0(\mathbf{p}, \mathbf{q}) = \mathcal{K}(\mathbf{p}, \mathbf{q}; 0, 0). \tag{5.24}$$

so that $\varphi_0(\mathbf{p}, \mathbf{q}) \in \mathfrak{C}$. Then it follows that

$$\begin{aligned}&(U[\mathbf{p}', \mathbf{q}']\varphi_0(\mathbf{p}'', \mathbf{q}''), U[\mathbf{p}, \mathbf{q}]\varphi_0(\mathbf{p}'', \mathbf{q}''))_c \\ &= (e^{i\mathbf{p}'\cdot(\mathbf{q}''-\mathbf{q}')}\mathcal{K}(\mathbf{p}'' - \mathbf{p}', \mathbf{q}'' - \mathbf{q}'; 0, 0), \\ &\quad \times e^{i\mathbf{p}\cdot(\mathbf{q}''-\mathbf{q})}\mathcal{K}(\mathbf{p}'' - \mathbf{p}, \mathbf{q}'' - \mathbf{q}; 0, 0))_c \\ &= (\mathcal{K}(\mathbf{p}'', \mathbf{q}''; \mathbf{p}', \mathbf{q}'), \mathcal{K}(\mathbf{p}'', \mathbf{q}''; \mathbf{p}, \mathbf{q}))_c \\ &= \mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}).\end{aligned} \tag{5.25}$$

The next step is to show that all the functions in $\mathfrak{C}$ are square integrable with respect to the Lebesgue measure $d\mu(\mathbf{p}, \mathbf{q}) \equiv (2\pi)^{-N} d^N p\, d^N q$, and that the inner product in $\mathfrak{C}$ is given by

$$(\psi, \varphi)_c = \iint \psi^*(\mathbf{p}, \mathbf{q})\varphi(\mathbf{p}, \mathbf{q})\, d\mu(\mathbf{p}, \mathbf{q}). \tag{5.26}$$

From property 4 of the kernel function, it follows that all the functions of $\mathfrak{C}_1$ are square integrable, and since the kernel function satisfies Eq. (5.3), it is also true that the inner product in $\mathfrak{C}_1$ is given by (5.26). Thus $\mathfrak{C}$ is also the L^2 completion of $\mathfrak{C}_1$. Consequently, all the functions in $\mathfrak{C}$ are square integrable, and their inner product is given by (5.26).

Finally, with the aid of (5.3), (5.4), (5.5) and (5.6), we can construct a unitary map of $\mathfrak{C}$ onto $L^2(R_N)$, which explicitly exhibits the irreducibility of the family of operators $U[\mathbf{p}, \mathbf{q}]$. Consider first the function

$$P(\mathbf{x}, \mathbf{y}) = \text{l.i.m.}\, \frac{1}{(2\pi)^N} \int \mathcal{K}(\mathbf{p}, \mathbf{x} - \mathbf{y}; 0, 0)e^{i\mathbf{p}\cdot\mathbf{y}}\, d^N p, \tag{5.27}$$

which we rigorously define in a moment. We show that $P(\mathbf{x}, \mathbf{y})$ can be written as

$$P(\mathbf{x}, \mathbf{y}) = \varphi_0(\mathbf{x})\varphi_0^*(\mathbf{y}), \tag{5.28}$$

where $\varphi_0(\mathbf{x})$ is a square integrable function of unit norm. The function $\varphi_0(\mathbf{x})$ is uniquely determined by $P(\mathbf{x}, \mathbf{y})$ up to a constant, complex multiple of magnitude one. We then show by mapping $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ onto $e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0(\mathbf{x} - \mathbf{q})$ that the desired unitary map of $\mathfrak{C}$ onto $L^2(R_N)$ is obtained.

The function $P(\mathbf{x}, \mathbf{y})$ is defined to be the limit in the mean of the sequence of functions

[24] S. Bochner and K. Chandrasekharan, Ref. 15, p. 92.

[25] F. Riesz and B. Sz.-Nagy, Ref. 11, p. 381.

$$P_n(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^N} \int_{|\mathbf{p}| \leq n} \mathcal{K}(\mathbf{p}, \mathbf{x} - \mathbf{y}; 0, 0) e^{i\mathbf{p}\cdot\mathbf{y}} \, d^N p,$$

$$n = 1, 2, \cdots, \qquad (5.29)$$

where the region of integration is over the sphere $|\mathbf{p}| \leq n$. The functions $P_n(\mathbf{x}, \mathbf{y})$ can be shown to be measurable functions in R_N in a straightforward manner. Furthermore, employing the fact that for almost all $\mathbf{x}$, $P_n(\mathbf{x} + \mathbf{y}, \mathbf{y})$ is the Fourier transform of a function of $\mathbf{y}$ which is in $L^1(R_N) \cap L^2(R_N)$, we can invoke Parseval's theorem to write

$$\int |P_n(\mathbf{x} + \mathbf{y}, \mathbf{y}) - P_m(\mathbf{x} + \mathbf{y}, \mathbf{y})|^2 \, d^N y$$

$$= \frac{1}{(2\pi)^N} \int_{m \leq |\mathbf{p}| \leq n} |\mathcal{K}(\mathbf{p}, \mathbf{x}; 0, 0)|^2 \, d^N p. \qquad (5.30)$$

Since the right-hand side of (5.30) is an integrable function of $\mathbf{x}$, we can integrate both sides with respect to $\mathbf{x}$, make a linear change of variables in the left-hand side, and invoke Fubini's theorem to interchange orders of integration, and obtain

$$\iint |P_n(\mathbf{x}, \mathbf{y}) - P_m(\mathbf{x}, \mathbf{y})|^2 \, d^N x \, d^N y$$

$$= \frac{1}{(2\pi)^N} \int_{m \leq |\mathbf{p}| \leq n} d^N p \int_{R_N} |\mathcal{K}(\mathbf{p}, \mathbf{x}; 0, 0)|^2 \, d^N x. \qquad (5.31)$$

Equation (5.31) shows that $P_n(\mathbf{x}, \mathbf{y})$ is a Cauchy sequence in $L^2(R_N \times R_N)$ and $P(\mathbf{x}, \mathbf{y})$ is defined to be its limit in $L^2(R_N \times R_N)$. It then follows that

$$\iint |P(\mathbf{x}, \mathbf{y})|^2 \, d^N x \, d^N y$$

$$= \frac{1}{(2\pi)^N} \iint |\mathcal{K}(\mathbf{p}, \mathbf{x}; 0, 0)|^2 \, d^N p \, d^N x$$

$$= \mathcal{K}(0, 0; 0, 0) = 1, \qquad (5.32)$$

where the last two equalities follow from (5.3) and (5.4b).

Furthermore, when considered as a function of $\mathbf{p}$ with $\mathbf{x}$ held fixed, $f_{\mathbf{x}}(\mathbf{p}) = \mathcal{K}(\mathbf{p}, \mathbf{x}; 0, 0) \in L^2(R_N)$ for almost all $\mathbf{x}$. Therefore, $f_{\mathbf{x}}(\mathbf{p})$ has a Fourier transform for almost all $\mathbf{x}$ which is just the limit in the mean of $P_n(\mathbf{x} + \mathbf{y}, \mathbf{y})$ considered as a function of $\mathbf{y}$ with $\mathbf{x}$ held fixed. By making use of the fact that every sequence of functions in L^2 which converges in the mean has a subsequence which converges almost everywhere, it can be shown that the Fourier transform of $f_{\mathbf{x}}(\mathbf{p})$ is equal to $P(\mathbf{x} + \mathbf{y}, \mathbf{y})$ for almost all $\mathbf{y}$.[26]

Two additional properties of $P(\mathbf{x}, \mathbf{y})$ are important. Because of identities (5.4b) and (5.6), it can be shown that $P_n(\mathbf{x}, \mathbf{y})^* = P_n(\mathbf{y}, \mathbf{x})$, $n = 1, 2, \cdots$; hence, for almost all $(\mathbf{x}, \mathbf{y})$,

$$P(\mathbf{x}, \mathbf{y})^* = P(\mathbf{y}, \mathbf{x}). \qquad (5.33)$$

Also, for almost all $(\mathbf{x}, \mathbf{y})$, we have

$$P(\mathbf{x}, \mathbf{y}) = \int P(\mathbf{x}, \mathbf{z}) P(\mathbf{z}, \mathbf{y}) \, d^N z. \qquad (5.34)$$

This can be proved by making use of the fact that $P(\mathbf{x} + \mathbf{y}, \mathbf{y})$ can be considered the Fourier transform of $\mathcal{K}(\mathbf{p}, \mathbf{x}; 0, 0)$. With the aid of (5.3) and (5.6), we can write

$$\mathcal{K}(\mathbf{p}, \mathbf{x}; 0, 0) = \frac{1}{(2\pi)^N} \int d^N s \int d^N r \, e^{-i\mathbf{r}\cdot(\mathbf{s}-\mathbf{x})}$$

$$\times \mathcal{K}(\mathbf{p} - \mathbf{r}, \mathbf{x} - \mathbf{s}; 0, 0) \mathcal{K}(\mathbf{r}, \mathbf{s}; 0, 0). \qquad (5.35)$$

Using Parseval's theorem, the inner integral on the right-hand side of (5.35) can be written as the convolution of the Fourier transforms of $\mathcal{K}(\mathbf{p} - \mathbf{r}, \mathbf{x} - \mathbf{s}; 0, 0)$ and $\mathcal{K}(\mathbf{r}, \mathbf{s}; 0, 0)$. Making a linear change of variables in the resulting integral, for almost all $(\mathbf{p}, \mathbf{x})$ we obtain

$$\mathcal{K}(\mathbf{p}, \mathbf{x}; 0, 0)$$

$$= \int d^N v \, e^{-i\mathbf{v}\cdot\mathbf{p}} \int d^N u \, P(\mathbf{v} + \mathbf{x}, \mathbf{u}) P(\mathbf{u}, \mathbf{v}). \qquad (5.36)$$

In Eqs. (5.35) and (5.36) the order of integration is immaterial since the integrand is an integrable function on $R_N \times R_N$. Now the inner integral in (5.36) is also a square integrable function of $\mathbf{v}$ for almost all $\mathbf{x}$. This is proved by applying Schwartz's inequality to the inner integral and then using arguments similar to those used in discussing the function defined in (3.11). It therefore follows that if $\mathbf{x}$ is held fixed, the inner integral in (5.36) is just the Fourier transform of $\mathcal{K}(\mathbf{p}, \mathbf{x}; 0, 0)$. It then follows that Eq. (5.34) is valid for almost all $(\mathbf{x}, \mathbf{y})$.

Since $P(\mathbf{x}, \mathbf{y})$ is square integrable and satisfies (5.33), it is a Hilbert–Schmidt kernel.[27] Furthermore, relation (5.34) shows that $P(\mathbf{x}, \mathbf{y})$ can only have the eigenvalues 0 or 1. Hilbert–Schmidt theory then says that all nonzero eigenvalues of $P(\mathbf{x}, \mathbf{y})$ are of finite multiplicity, and if $\varphi_j(\mathbf{x})$, $j = 0, \cdots, R$ are the orthonormal set of eigenfunctions corresponding to the eigenvalue one, then

$$P(\mathbf{x}, \mathbf{y}) = \sum_{j=0}^{R} \varphi_j(\mathbf{x}) \varphi_j^*(\mathbf{y}). \qquad (5.37)$$

[26] The authors are indebted to L. A. Shepp for demonstrating the truth of this statement.

[27] F. Riesz and B. Sz.-Nagy, Ref. 11, p. 242.

But Eq. (5.32) can only be satisfied if the eigenvalue one has multiplicity one and hence $R = 0$. This completes the proof that $P(\mathbf{x}, \mathbf{y})$ can be written in the form (5.28) and identifies $\varphi_0(\mathbf{x})$ as the eigenfunction of $P(\mathbf{x}, \mathbf{y})$ corresponding to the eigenvalue one. Finally, if we substitute Eq. (5.28) into Eq. (5.36) and make use of (5.6), we can obtain

$$\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = \int e^{-i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0^*(\mathbf{x}-\mathbf{q})e^{i\mathbf{r}\cdot(\mathbf{x}-\mathbf{s})}\varphi_0(\mathbf{x}-\mathbf{s})\, d^N x. \quad (5.38)$$

We now proceed to construct the unitary map L of $\mathfrak{C}$ onto $L^2(R_N)$. For fixed $(\mathbf{p}, \mathbf{q})$, $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ is a function in $\mathfrak{C}$, and we set

$$L\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}) = e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0(\mathbf{x}-\mathbf{q}). \quad (5.39)$$

Also for any $\psi(\mathbf{p}', \mathbf{q}') \in \mathfrak{C}_1$, we have

$$\begin{aligned} L\psi(\mathbf{p}', \mathbf{q}') &= L\Big(\sum_{k=1}^{N} \alpha_k \mathcal{K}(\mathbf{p}', \mathbf{q}'; p_k, q_k)\Big) \\ &= \sum_{k=1}^{N} \alpha_k L\mathcal{K}(\mathbf{p}', \mathbf{q}'; p_k, q_k) \\ &= \sum_{k=1}^{N} \alpha_k e^{i\mathbf{p}_k\cdot(\mathbf{x}-\mathbf{q}_k)}\varphi_0(\mathbf{x}-\mathbf{q}_k). \end{aligned} \quad (5.40)$$

Let us verify that L is isometric on $\mathfrak{C}_1$. From (5.39) it follows that

$$(L\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}), L\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{r}, \mathbf{s})) = \int e^{-i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0^*(\mathbf{x}-\mathbf{q})e^{i\mathbf{r}\cdot(\mathbf{x}-\mathbf{s})}\varphi_0(\mathbf{x}-\mathbf{s})\, d^N x, \quad (5.41)$$

where the inner product is taken in $L^2(R_N)$. If (5.38) and (5.41) are combined, there results

$$\begin{aligned} (L\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}), L\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{r}, \mathbf{s})) &= \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) \\ &= (\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}), \mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{r}, \mathbf{s}))_c. \end{aligned} \quad (5.42)$$

Because of the way in which the inner product is defined for elements in $\mathfrak{C}_1$, it is readily seen that L is a linear map of $\mathfrak{C}_1$ into $L^2(R_N)$ which preserves inner products. The representation (5.7) of an element of $\mathfrak{C}_1$ is not unique. However, the isometry of L on $\mathfrak{C}_1$ shows that it maps two different representations of the same element in $\mathfrak{C}_1$ onto the same element in $L^2(R_N)$, i.e., L is single-valued. Since $\mathfrak{C}_1$ is dense in $\mathfrak{C}$, L can now be uniquely extended in standard fashion to a map of all of $\mathfrak{C}$ into $L^2(R_N)$ which preserves inner products.[23]

A linear map which preserves inner products is one-to-one, and so to complete the proof that L is a unitary map of $\mathfrak{C}$ onto $L^2(R_N)$, we must show that the image of $\mathfrak{C}$ under L is all of $L^2(R_N)$. However, the image of $\mathfrak{C}$ under L is just the subspace spanned by the set of elements $e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0(\mathbf{x}-\mathbf{q})$ for all $\mathbf{p}$, $\mathbf{q}$, and by Lemma 3.3 this subspace is all of $L^2(R_N)$. This completes the proof that L is a unitary map of $\mathfrak{C}$ onto $L^2(R_N)$.

It is now a straightforward matter to determine the image under L of the family of operators $U[\mathbf{p}, \mathbf{q}]$ which we have defined on $\mathfrak{C}$. If $f(\mathbf{x}) \in L^2(R_N)$, then

$$LU[\mathbf{p}, \mathbf{q}]L^{-1}f(\mathbf{x}) = e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}f(\mathbf{x}-\mathbf{q}). \quad (5.43)$$

However, von Neumann's theorem asserts that the family of operators $LU[\mathbf{p}, \mathbf{q}]L^{-1}$ is irreducible, and hence the family of operators $U[\mathbf{p}, \mathbf{q}]$, defined on $\mathfrak{C}$, is irreducible. The unitary map L is just the Schrödinger map of $\mathfrak{C}$ onto $L^2(R_N)$.

This completes the reconstruction of the continuous representation starting from a given kernel function. We summarize these results in

Theorem 5.1. Let $\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})$ be a function which satisfies Eq. (5.3) and possesses properties 1 through 4 of part A of this section. Then $\mathcal{K}$ uniquely determines a continuous representation of $L^2(R_N)$. The fiducial vector $\varphi_0(x)$ is uniquely determined up to a constant, complex multiple of absolute value one. An irreducible representation of the family of operators $U[\mathbf{p}, \mathbf{q}]$ is uniquely determined, and the corresponding OFS is uniquely determined up to a unitary equivalence.

C. The Significance of the Inner Product in $\mathfrak{C}$

In concluding this discussion about Aronszajn spaces and continuous representations, we want to emphasize the role played by the inner product. In the first place, the fact that each $\mathfrak{C}$ is a subspace of $L^2(R_N \times R_N)$ is nontrivial, because examples of Aronszajn spaces exist for which this is not the case. For example, let[28]

$$\mathcal{K}(p; p') = e^{-\frac{1}{2}(p-p')^2}. \quad (5.44)$$

It is not difficult to show that this is indeed a kernel function, but that there is no measure $d\mu(p)$ with which the idempotent character of $\mathcal{K}(p; p')$ can be expressed as

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}(p'-p'')^2}e^{-\frac{1}{2}(p''-p)^2}\, d\mu(p'') = e^{-\frac{1}{2}(p'-p)^2}. \quad (5.45)$$

Thus the Aronszajn space corresponding to this kernel is *not* a subspace of $L^2(R)$ for any measure μ.

In the second place, there is an intimate relationship between the measure $[d^N p/(2\pi)^{\frac{1}{2}N}][d^N q/(2\pi)^{\frac{1}{2}N}]$ and the irreducibility of the representation of the

[28] This is a special case of a class of functions studied by Aronszajn, Ref. 8, p. 152.

operators $U[\mathbf{p}, \mathbf{q}]$ constructed on $\mathfrak{C}$. For example let Eq. (5.3) be replaced by the relation

$$\iint \mathcal{K}^*(\mathbf{p}'', \mathbf{q}''; \mathbf{p}, \mathbf{q})\mathcal{K}(\mathbf{p}'', \mathbf{q}''; \mathbf{p}', \mathbf{q}') \times R^{\frac{1}{2}} \frac{d^N p''}{(2\pi)^{\frac{1}{2}N}} R^{\frac{1}{2}} \frac{d^N q''}{(2\pi)^{\frac{1}{2}N}} = \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{p}', \mathbf{q}'), \tag{5.46}$$

where R is a positive integer. Then it can be shown by the same methods used in part B of this section, that the representation of $U[\mathbf{p}, \mathbf{q}]$, given by (5.18) and (5.19), is no longer irreducible. In fact a unitary map of $\mathfrak{C}$ onto the R-fold direct sum of $L^2(R_N)$ with itself can be constructed using the techniques of part B, such that the image of $U[\mathbf{p}, \mathbf{q}]$ is completely reduced by each of the R subspaces, $L^2(R_N)$. The kernel itself may be decomposed as

$$\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}) = R^{-1} \sum_{i=1}^{R} \mathcal{K}_i(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}) \tag{5.47}$$

in which each $\mathcal{K}_i$ is generated by an irreducible representation of $U[\mathbf{p}, \mathbf{q}]$ and the corresponding $\mathfrak{C}_i$ are orthogonal subspaces of $L^2(R_N \times R_N)$. These kernels arose in connection with the analysis in III, Sec. 2.

More generally, however, most kernels based on reducible representations of the operator family $U[\mathbf{p}, \mathbf{q}]$ admit no expression of the inner product with the aid of one measure solely on $R_N \times R_N$ since a decomposition in the manner of (5.47) is not into orthogonal subspaces of $L^2(R_N \times R_N)$. Such a decomposition may include a finite number of terms as in (5.47) or may be of the form

$$\mathcal{K}(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q}) = \int \mathcal{K}_\alpha(\mathbf{p}', \mathbf{q}'; \mathbf{p}, \mathbf{q})\, d\sigma(\alpha), \qquad 1 = \int d\sigma(\alpha), \tag{5.48}$$

where each $\mathcal{K}_\alpha$ corresponds to an irreducible representation, and where $\mathcal{K}_\alpha$ is a σ-measurable function, σ being a nondecreasing real function of bounded variation. In either case, the properties of such kernels may be found from a knowledge of their irreducible components.

6. THE REPRESENTATION OF LINEAR OPERATORS ON $\mathfrak{C}$

A. Bounded Operators on $\mathfrak{C}$

In Eq. (4.1) we defined the linear map C, which maps the Hilbert space $\mathfrak{H}$ onto its continuous representation $\mathfrak{C}$. In this section we examine the images under C of all bounded, linear operators mapping $\mathfrak{H}$ into itself. Let $\Psi \in \mathfrak{H}$ and let B be a bounded, linear operator on $\mathfrak{H}$. Then, making use of (3.19),

$$CB\Psi = (\Phi[\mathbf{p}, \mathbf{q}], B\Psi) = \iint (\Phi[\mathbf{p}, \mathbf{q}], B\Phi[\mathbf{r}, \mathbf{s}])(\Phi[\mathbf{r}, \mathbf{s}], \Psi)\, d\mu(\mathbf{r}, \mathbf{s}) = \iint B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})\psi(\mathbf{r}, \mathbf{s})\, d\mu(\mathbf{r}, \mathbf{s}). \tag{6.1}$$

Thus the image of B is the integral operator whose kernel is

$$B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = (\Phi[\mathbf{p}, \mathbf{q}], B\Phi[\mathbf{r}, \mathbf{s}]). \tag{6.2}$$

Since every bounded, linear operator on $\mathfrak{C}$ is the image of a bounded, linear operator on $\mathfrak{H}$, we can conclude that every bounded, linear operator on $\mathfrak{C}$ can be represented by an integral operator of the form (6.2). Notice that the kernel corresponding to the unit operator is just $\mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$.

We state without proof a number of conditions satisfied by the kernel (6.2). In what follows $\varphi(\mathbf{p}, \mathbf{q})$ and $\psi(\mathbf{p}, \mathbf{q})$ are arbitrary elements of $\mathfrak{C}$.

1. $B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$ is a continuous, bounded function. It is continuous in $(\mathbf{p}, \mathbf{q})$ uniformly with respect to $(\mathbf{r}, \mathbf{s})$, and it is continuous in $(\mathbf{r}, \mathbf{s})$ uniformly with respect to $(\mathbf{p}, \mathbf{q})$.

2. For fixed $(\mathbf{p}, \mathbf{q})$, $B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$ is square integrable with respect to $d\mu(\mathbf{r}, \mathbf{s})$, and for fixed $(\mathbf{r}, \mathbf{s})$ it is square integrable with respect to $d\mu(\mathbf{p}, \mathbf{q})$.

3. The functions

$$\lambda(\mathbf{p}, \mathbf{q}) \equiv \iint B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})\varphi(\mathbf{r}, \mathbf{s})\, d\mu(\mathbf{r}, \mathbf{s}) \tag{6.3}$$

and

$$\chi^*(\mathbf{r}, \mathbf{s}) \equiv \iint \psi^*(\mathbf{p}, \mathbf{q})B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})\, d\mu(\mathbf{p}, \mathbf{q}) \tag{6.4}$$

are both square integrable, and

$$\iint \psi^*(\mathbf{p}, \mathbf{q})\lambda(\mathbf{p}, \mathbf{q})\, d\mu(\mathbf{p}, \mathbf{q}) = \iint \chi^*(\mathbf{r}, \mathbf{s})\varphi(\mathbf{r}, \mathbf{s})\, d\mu(\mathbf{r}, \mathbf{s}). \tag{6.5}$$

Furthermore, there is a constant $||B||$, such that

$$\left|\iint \psi^*(\mathbf{p}, \mathbf{q})\lambda(\mathbf{p}, \mathbf{q})\, d\mu(\mathbf{p}, \mathbf{q})\right| \leq ||B|| \cdot ||\psi||_c \cdot ||\varphi||_c. \tag{6.6}$$

4. The kernel for B obeys

$$\iint B(\mathbf{p}, \mathbf{q}; \mathbf{t}, \mathbf{u})\mathcal{K}(\mathbf{t}, \mathbf{u}; \mathbf{r}, \mathbf{s})\, d\mu(\mathbf{t}, \mathbf{u})$$

$$= \iint \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{t}, \mathbf{u})B(\mathbf{t}, \mathbf{u}; \mathbf{r}, \mathbf{s})\, d\mu(\mathbf{t}, \mathbf{u})$$

$$= B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}). \tag{6.7}$$

5. If $B^\dagger$ is the adjoint transformation of B, its kernel is

$$B^\dagger(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = B(\mathbf{r}, \mathbf{s}; \mathbf{p}, \mathbf{q})^*. \tag{6.8}$$

If B is self-adjoint ($B^\dagger = B$), then

$$B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = B(\mathbf{r}, \mathbf{s}; \mathbf{p}, \mathbf{q})^*. \tag{6.9}$$

6. If A and B are bounded, linear operator on $\mathfrak{H}$, then so is AB, and their respective kernels satisfy

$$(AB)(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$$

$$= \iint A(\mathbf{p}, \mathbf{q}; \mathbf{t}, \mathbf{u})B(\mathbf{t}, \mathbf{u}; \mathbf{r}, \mathbf{s})\, d\mu(\mathbf{t}, \mathbf{u}). \tag{6.10}$$

7. If A is unitary, then

$$\iint A(\mathbf{p}, \mathbf{q}; \mathbf{t}, \mathbf{u})A(\mathbf{r}, \mathbf{s}; \mathbf{t}, \mathbf{u})^*\, d\mu(\mathbf{t}, \mathbf{u})$$

$$= \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}), \tag{6.11a}$$

$$\iint A(\mathbf{t}, \mathbf{u}; \mathbf{p}, \mathbf{q})^*A(\mathbf{t}, u; \mathbf{r}, \mathbf{s})\, d\mu(\mathbf{t}, \mathbf{u})$$

$$= \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}). \tag{6.11b}$$

8. If tr A exists, it is given by

$$\operatorname{tr} A = \iint A(\mathbf{p}, \mathbf{q}; \mathbf{p}, \mathbf{q})\, d\mu(\mathbf{p}, \mathbf{q}). \tag{6.12}$$

We now list several examples of bounded operators. In the first three cases we choose $\mathfrak{H}$ as $L^2(R_N)$.

a. Let T be the Fourier transform,

$$Tf(\mathbf{x}) = \text{l.i.m. } (2\pi)^{-\frac{1}{2}N} \int e^{i\mathbf{x}\cdot\mathbf{y}} f(\mathbf{y})\, d^N y. \tag{6.13}$$

Then if $\varphi_0(\mathbf{x})$ represents the fiducial vector and $\tilde{\varphi}_0(\mathbf{y})$ its Fourier Transform,

$$T(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$$

$$= \int e^{-i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0^*(\mathbf{x} - \mathbf{q})e^{i\mathbf{x}\cdot\mathbf{s}}\tilde{\varphi}_0(\mathbf{x} + \mathbf{r})\, d^N x. \tag{6.14}$$

Since T is unitary, $T(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$ satisfies condition 7.

b. Let E_{λ_α} be the projection operator defined as follows:

$$E_{\lambda_\alpha} f(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & x_\alpha \leq \lambda_\alpha, \\ 0, & x_\alpha > \lambda_\alpha. \end{cases} \tag{6.15}$$

Here x_α is the αth component of $\mathbf{x}$. Then

$$E_{\lambda_\alpha}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = \int e^{-i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})}\varphi_0^*(\mathbf{x} - \mathbf{q})$$

$$\times e^{i\mathbf{r}\cdot(\mathbf{x}-\mathbf{s})}\varphi_0(\mathbf{x} - \mathbf{s})\theta(\lambda_\alpha - x_\alpha)\, d^N x, \tag{6.16}$$

where

$$\theta(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{6.17}$$

c. Let F_{λ_α} be the projection operator defined as follows: If $\tilde{f}(\mathbf{k})$ is the Fourier transform of $f(\mathbf{x})$, then

$$F_{\lambda_\alpha} f(x) = \text{l.i.m. } (2\pi)^{-\frac{1}{2}N}$$

$$\times \int \theta(\lambda_\alpha - k_\alpha)\tilde{f}(-\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}\, d^N k, \tag{6.18}$$

where k_α is the αth component of $\mathbf{k}$. Then

$$F_{\lambda_\alpha}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = \int e^{-i\mathbf{k}\cdot\mathbf{q}}\tilde{\varphi}_0^*(\mathbf{k} + \mathbf{p})$$

$$\times e^{i\mathbf{k}\cdot\mathbf{s}}\tilde{\varphi}_0(\mathbf{k} + \mathbf{r})\theta(\lambda_\alpha + k_\alpha)\, d^N k. \tag{6.19}$$

d. In Sec. 3 we defined operators of the form

$$B = \iint b(\mathbf{t}, \mathbf{u})\Phi[\mathbf{t}, \mathbf{u}]\Phi^\dagger[\mathbf{t}, \mathbf{u}]\, d\mu(\mathbf{t}, \mathbf{u}), \tag{6.20}$$

where $b(\mathbf{t}, \mathbf{u})$ is a bounded, measurable function. Then

$$B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = \iint \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{t}, \mathbf{u})b(\mathbf{t}, \mathbf{u})$$

$$\times \mathcal{K}(\mathbf{t}, \mathbf{u}; \mathbf{r}, \mathbf{s})\, d\mu(\mathbf{t}, \mathbf{u}). \tag{6.21}$$

e. Weyl[29] introduced a class of operators of the form

$$F = \iint U[\mathbf{t}, \mathbf{u}]f(\mathbf{t}, \mathbf{u})\, d\mu(\mathbf{t}, \mathbf{u}). \tag{6.22}$$

If $f(\mathbf{t}, \mathbf{u})$ is a square integrable function, and F is defined by its matrix elements, as is the analogous operator (6.20), then F is a bounded, linear operator for which

$$F(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = \iint e^{i\mathbf{t}\cdot\mathbf{s}}$$

$$\times \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{r} + \mathbf{t}, \mathbf{s} + \mathbf{u})f(\mathbf{t}, \mathbf{u})\, d\mu(\mathbf{t}, \mathbf{u}). \tag{6.23}$$

We now show that every operator kernel function satisfying conditions 2, 3, and 4 defines a bounded, linear operator on $\mathfrak{E}$, and at the same time we construct the image under C^{-1} of this operator in $\mathfrak{H}$. Let $B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$ satisfy conditions 2–4, and consider

[29] H. Weyl, Ref. 12, p. 274.

the formal, operator-valued (on $\mathfrak{H}$) integral

$$B = \iiiint \Phi[\mathbf{p}, \mathbf{q}]\Phi^{\dagger}[\mathbf{r}, \mathbf{s}] \times B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})\, d\mu(\mathbf{p}, \mathbf{q})\, d\mu(\mathbf{r}, \mathbf{s}). \quad (6.24)$$

We define B as that operator on $\mathfrak{H}$ which maps the vector Ψ onto

$$B\Psi = \iint \Phi[\mathbf{p}, \mathbf{q}](B\psi)(\mathbf{p}, \mathbf{q})\, d\mu(\mathbf{p}, \mathbf{q}), \quad (6.25)$$

where

$$(B\psi)(\mathbf{p}, \mathbf{q}) = \iint B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})\psi(\mathbf{r}, \mathbf{s})\, d\mu(\mathbf{r}, \mathbf{s}). \quad (6.26)$$

Since $\psi(\mathbf{r}, \mathbf{s}) = (\Phi[\mathbf{r}, \mathbf{s}], \Psi)$ is an element of $\mathfrak{C}$, it follows from conditions 2 and 3 that $(B\psi)(\mathbf{p}, \mathbf{q})$ is square integrable. Then according to the results of Sec. 3, the integral (6.25) is a well-defined vector in $\mathfrak{H}$. The linearity of the operator defined in (6.24)–(6.26) follows directly from the relation

$$(\Phi[\mathbf{p}, \mathbf{q}], \alpha\Psi + \Lambda) = \alpha(\Phi[\mathbf{p}, \mathbf{q}], \Psi) + (\Phi[\mathbf{p}, \mathbf{q}], \Lambda), \quad (6.27)$$

for all $\Psi, \Lambda \in \mathfrak{H}$, and all complex numbers α, and the fact that the integral is a linear operation. Furthermore, condition 3 shows that B is a bounded operator. Finally, we construct the kernel function of the operator on $\mathfrak{C}$ corresponding to B under the map C. With the aid of condition 4, we get

$$CBC^{-1} = (\Phi[\mathbf{p}, \mathbf{q}], B\Phi[\mathbf{r}, \mathbf{s}]) = B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}). \quad (6.28)$$

This completes the demonstration that any operator kernel function satisfying conditions 2–4 is the image under C of a bounded, linear operator on $\mathfrak{H}$. The operator B on $\mathfrak{H}$ corresponding to the kernel function $B(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$ under C^{-1} is given by (6.24).

It should be noticed that the unit operator has the representation

$$I = \iiiint \Phi[\mathbf{p}, \mathbf{q}]\Phi^{\dagger}[\mathbf{r}, \mathbf{s}] \times \mathcal{K}(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})\, d\mu(\mathbf{p}, \mathbf{q})\, d\mu(\mathbf{r}, \mathbf{s}), \quad (6.29)$$

which in view of (3.19) and the definition of $\mathcal{K}$ is equivalent to (3.23).

B. Unbounded Operators on $\mathfrak{C}$

If O is an unbounded, linear operator mapping $\mathfrak{H}$ into itself, then considerably less can be said about its image in $\mathfrak{C}$. In particular, it could happen that none of the vectors $\Phi[\mathbf{p}, \mathbf{q}]$ is in the domain of O, in which case a representation of the image of O as an integral operator on $\mathfrak{C}$, as defined in part A, would be impossible.

However, in the important case of self-adjoint operators more can be said. It is well known that corresponding to each self-adjoint operator O defined on $\mathfrak{H}$, there is a family of projection operators on $\mathfrak{H}$, $\{E_\lambda\}$, $-\infty < \lambda < \infty$, called a resolution of the identity, which satisfies the relation[30]

$$O = \int_{-\infty}^{\infty} \lambda\, dE_\lambda. \quad (6.30)$$

In certain applications, it is the resolution of the identity which is of most interest; this is the case in the quantum mechanical theory of measurements, for example.[31] Since all projections are bounded operators, the results of part A can be applied to the projections comprising a resolution of the identity.

Two important resolutions of the identity are given in examples b and c.[32] The two families of projections, $\{E_{\lambda_\alpha}\}$ of example b and $\{F_{\lambda_\alpha}\}$ of example c, $-\infty < \lambda_\alpha < \infty$ are the resolutions of the identity of the operators Q_α and P_α, respectively, defined in (4.9).

Finally, in those cases where the set $\mathfrak{S}$ (all vectors of the form $\Phi[\mathbf{p}, \mathbf{q}]$) is included in the domain, $\mathfrak{D}_O$, of the unbounded operator O, it is sometimes possible to represent O as an integral operator on $\mathfrak{C}$. This is always the case when O is self-adjoint and $\mathfrak{S} \subset \mathfrak{D}_O$. For then

$$O(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = (\Phi[\mathbf{p}, \mathbf{q}], O\Phi[\mathbf{r}, \mathbf{s}]) = (O\Phi[\mathbf{p}, \mathbf{q}], \Phi[\mathbf{r}, \mathbf{s}]). \quad (6.31)$$

Again the results of Sec. 3 show that $O(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$ is a continuous, square integrable function of $(\mathbf{p}, \mathbf{q})$ for fixed $(\mathbf{r}, \mathbf{s})$ [this is also true with $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{r}, \mathbf{s})$ interchanged]. Then from Eq. (3.16), if $\psi(\mathbf{p}, \mathbf{q}) \in \mathfrak{C}$,

$$\iint O(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})\psi(\mathbf{r}, \mathbf{s})\, d\mu(\mathbf{r}, \mathbf{s}) = \int (O\Phi[\mathbf{p}, \mathbf{q}], \Phi[\mathbf{r}, \mathbf{s}])(\Phi[\mathbf{r}, \mathbf{s}], \Psi)\, d\mu(\mathbf{r}, \mathbf{s}) = (O\Phi[\mathbf{p}, \mathbf{q}], \Psi). \quad (6.32)$$

If $\Psi \in \mathfrak{D}_O$, then

$$(O\Phi[\mathbf{p}, \mathbf{q}], \Psi) = (\Phi[\mathbf{p}, \mathbf{q}], O\Psi). \quad (6.33)$$

In other words $O(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s})$ maps the representative of Ψ in $\mathfrak{C}$ onto the representative of $O\Psi$ in $\mathfrak{C}$. However, if $\Psi \notin \mathfrak{D}_O$ then the right-hand side of (6.32) is not a function in $\mathfrak{C}$.

---

[30] F. Riesz and B. Sz.-Nagy, Ref. 11.

[31] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, New Jersey, 1955), Chap. III.

[32] J. von Neumann, Ref. 31, pp. 128–136.

Two important examples are the operators Q_α and P_α. In Sec. 4 we showed that if $\Phi_0 \in \mathfrak{D}_{Q_\alpha}$, then $\mathfrak{S} \subset \mathfrak{D}_{Q_\alpha}$; and also if $\Phi_0 \in \mathfrak{D}_{P_\alpha}$, then $\mathfrak{S} \subset \mathfrak{D}_{P_\alpha}$. In a straightforward fashion one can show that if

$$Q_\alpha(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = \int x_\alpha e^{-i\mathbf{p}\cdot(\mathbf{x}-\mathbf{q})} \times \varphi_0^*(\mathbf{x} - \mathbf{q}) e^{i\mathbf{r}\cdot(\mathbf{x}-\mathbf{s})} \varphi_0(\mathbf{x} - \mathbf{s})\, d^N x, \quad (6.34)$$

and similarly if $\Phi_0 \in \mathfrak{D}_{P_\alpha}$, then

$$P_\alpha(\mathbf{p}, \mathbf{q}; \mathbf{r}, \mathbf{s}) = -\int k_\alpha e^{-i\mathbf{k}\cdot\mathbf{q}} \tilde{\varphi}_0^*(\mathbf{k} + \mathbf{p}) e^{i\mathbf{k}\cdot\mathbf{s}} \tilde{\varphi}_0(\mathbf{k} + \mathbf{r})\, d^N k. \quad (6.35)$$

In concluding Sec. 6, it should be pointed out that the fact that bounded operators can be represented by kernel functions on $\mathfrak{C}$ is typical of Aronszajn spaces. However, we were able to say more about the structure of these kernels in the case of continuous representations than can be said in the case of a general Aronszajn space.

7. ACKNOWLEDGMENTS

We are indebted to several of our colleagues for helpful discussions on various topics in this paper. In particular, we wish to thank H. J. Landau and L. A. Shepp.

Uniqueness of 4- and 8-Dimensional Spaces

D. C. PEASLEE
Australian National University, Canberra, Australia
(Received 3 December 1963)

Among rotation groups R_n, the cases $n = 4$ and 8 are unique in having two inequivalent $n \times n$ representations. Mathematically this is related to the uniqueness of quaternions and octonions; physically these groups seem to underlie the real and charge-space symmetries of elementary particles. An attempt is made to interpret this fact by assuming a lack of inherent geometrical preference between Fermi–Dirac and Bose–Einstein statistics. Corollaries are the identity of real and charge-space statistics and the complete disjointness of real and charge-space coordinates.

INTRODUCTION

AMONG n-dimensional rotation groups R_n the cases $n=4$ and $n=8$ are specially distinguished:

(i) They have two inequivalent representations of dimension n, all other cases having only one;

(ii) They seem to determine the symmetries of elementary particles, R_4 for real space (with one imaginary axis), R_8 for charge space.[1]

The following note seeks to relate (i) and (ii) in terms of the simultaneous possibility of Bose–Einstein and Fermi–Dirac statistics. As corollaries:

(a) real and charge-space statistics are completely correlated;

(b) real and charge-space coordinates are completely disjoint;

(c) five- and six-dimensional treatments of Dirac and Maxwell equations appear as formalities without special physical significance.

1. ALGEBRAIC RELATIONS

For any dimension, the regular representation of R_n is given by the matrices $\Sigma_{AB} = -i(1_{AB} - 1_{BA})$, $A, B = 0, 1, \cdots, n-1$; here 1_{AB} is an $n \times n$ matrix with a unit entry in the Ath row and Bth column, zeros elsewhere. Then

$$\Sigma_{AB}\Sigma_{CD} = \delta_{AC}1_{BD} + \delta_{BD}1_{AC} - \delta_{AD}1_{BC} - \delta_{BC}1_{AD}, \quad (1)$$

so that Σ_{AB} and Σ_{CD} commute if they have no common indices; and $\Sigma_{AB}^2 = 1_{AA} + 1_{BB}$, whence the eigenvalues of Σ_{AB} are $+1, -1, (0)^{n-2}$.

We obtain an inequivalent $n \times n$ representation of R_n if it is possible to form linear combinations of the Σ_{AB} that behave like corresponding Dirac matrices Γ_A $(A = 1 \cdots n-1)$: namely,

$$\Gamma_A^2 = 1, \qquad \Gamma_A\Gamma_B = -\Gamma_B\Gamma_A, \quad A \neq B. \quad (2)$$

The independent representation is then

$$\sigma_{AB} = (1/4i)[\Gamma_A, \Gamma_B], \quad A, B \neq 0, \quad (3)$$
$$\sigma_{0A} = \pm\Gamma_A = -\sigma_{A0}.$$

The σ_{AB} satisfy the same commutation rules as the Σ_{AB} and hence represent R_n; but their eigenvalue distribution is $(\frac{1}{2})^{\frac{1}{2}n}$, $(-\frac{1}{2})^{\frac{1}{2}n}$, so they are inequivalent to Σ_{AB}. This incidentally shows that n must be even.

The relation between the eigenvalues of Γ_A and Σ_{AB} requires the linear combination to have the form

$$\Gamma_A = \tfrac{1}{2}g_{ABC}\Sigma_{BC} \pm \Sigma_{0A} \quad (4)$$

where the coefficients g_{ABC} are totally antisymmetric, run only over the values $A, B, C \neq 0$, and have magnitude unity or zero. The factor of $\frac{1}{2}$ in Eq. (4) is to correct for counting each Σ_{BC} twice in the usual convention for summing over repeated indices. For each choice of $A, B \neq 0$ there must be just one nonvanishing g_{ABC}. Then each Γ_A will consist of $\frac{1}{2}n$ different Σ_{AB} with no common indices, so the eigenvalue distribution $(-1)^{n/2}$, $(+1)^{n/2}$ for Γ_A is assured.

Consider the expression

$$g_{ABC}g_{ADE} + g_{ABE}g_{ADC} = 2\delta_{BD}\delta_{CE} - \delta_{BC}\delta_{DE} - \delta_{BE}\delta_{CD} + \gamma_{BCDE}, \quad (5)$$

with summation over A on the left-hand side, and γ_{BCDE} a totally antisymmetric function of its indices. The conditions of antisymmetry and uniqueness already imposed on the g_{ABC} imply the form on the right-hand side of Eq. (5); it will not be necessary for present purposes to determine γ_{BCDE} any more closely, although in fact it vanishes identically when $n = 4, 8$ (in the latter case the g_{ABC} become just the e_{ABC} of Ref. 1). Now

$$\Gamma_A\Gamma_B = \tfrac{1}{4}g_{ACD}g_{BEF}\Sigma_{CD}\Sigma_{EF} \pm \tfrac{1}{2}\{g_{ACD}\Sigma_{CD}\Sigma_{0B} + g_{BEF}\Sigma_{0A}\Sigma_{EF}\} + \Sigma_{0A}\Sigma_{0B}$$
$$= g_{ACD}g_{BCF}1_{DF} \pm ig_{ABC}\Sigma_{0C} + 1_{AB} + \delta_{AB}1_{00}; \quad (6)$$

[1] Arguments favoring R_8 for charge space are presented, for example, by D. C. Peaslee, J. Math. Phys. 4, 910 (1963).

but by insertion of Eq. (5) this becomes

$$\Gamma_A\Gamma_B = -g_{ABC}g_{CDF}1_{DF} - 2(1_{BA}) + 1_{AB} + \delta_{AB}1_{FF}(1 - \delta_{0F}) - \gamma_{ABDF}1_{DF} \pm ig_{ABC}\Sigma_{0C} + 1_{AB} + \delta_{AB}1_{00} = \delta_{AB} \pm ig_{ABC}\Sigma_{0C} + 2i\Sigma_{AB}. \quad (7)$$

This last form establishes both parts of Eq. (2).

We now examine whether any g_{ABC} can in fact satisfy the requirements imposed. Suppose the set to start with the members

$$g_{123}, g_{145}, g_{167}, \cdots g_{1(n-2)(n-1)} = +1, \quad (8)$$

which can be assumed without loss of generality. There are $\frac{1}{2}(n-2)$ such terms. The additional nonvanishing g_{ABC} with $A = 2$ must have $B, C > 3$, and likewise for $A = 3$. For both $A = 2$ and $A = 3$ there are $\frac{1}{2}(n-2) - 1$ new, nonvanishing g_{ABC}; for $A = 4, 5$ there are $\frac{1}{2}(n-2) - 3$ terms each; for $A = 2m, 2m+1$, there are $\frac{1}{2}(n-2) - (2m-1)$ terms each. This sequence clearly stops in general before the ultimate limit of $2m + 1 = n - 1$; in fact, there is zero contribution when $(2m+1) = \frac{1}{2}(m-2)$, or

$$n = 4m \quad (9)$$

where m is a positive integer. The total number of independent g_{ABC} collected up to the cutoff of Eq. (9) is $(2m-1)\frac{1}{2}(n-2) - 2(m-1)^2 = 2m^2 - 1$. If one regards the g_{ABC} as the BC elements of square matrices G_A, then there is clearly a one-to-one correspondence between Γ_A and G_A; since there are only $n-1$ quantities Γ_A, the same is true of g_{ABC}. Thus,

$$2m^2 - 1 = n - 1 = 4m - 1, \quad m = 0, 2. \quad (10)$$

The only nonempty case therefore appears to be $n = 4m = 8$. A special degenerate case can be added, however, by considering $g_{ABC}, g_{BCA}, g_{CAB}$ as independent, so that

$$3(2m^2 - 1) = 4m - 1, \qquad m = -\tfrac{1}{3}, 1, \quad (11)$$

and $n = 4$. From the present point of view, quaternions appear as degenerate (hence associative) octonions.[2]

Of course it has not been shown that constructions of the type in Eq. (4) exhaust the possibilities of finding inequivalent $n \times n$ representations of R_n. At this point we recall the group theory result[3] that this situation exists only for $n = 4, 8$; since Eqs. (10) and (11) yield the same result, Eq. (4) must be perfectly general.

The connection of the above with conventional definitions of octonions and quaternions[4] is immediate. The matrices $(i\Gamma_A)$ are a satisfactory representation of octonions ($A = 1 \cdots 7$) and quaternions ($A = 1 \cdots 3$); the coefficients g_{ABC} can be read immediately from the definition

$$\epsilon_A\epsilon_B = g_{ABC}\epsilon_C, \quad (12)$$

where ϵ_A is an octonion or quaternion.

2. GEOMETRIC INTERPRETATION

For any space of n orthogonal axes the basic geometrical object is the coordinatelike vector x_A, $A = 0 \cdots, n-1$; the elementary transformation is the rotation of a vector, specified by the $n \times n$ matrices Σ_{AB}. In general, objects that transform under the σ_{AB} are called half vectors. For the special cases $n = 4$ and 8, however, a different interpretation is possible. Since the σ_{AB} are of the right dimensions to apply to a vector, one can say that they generate a vector transformation $x_A \rightarrow x'_A$ that is not rotation but "half rotation". If Γ_A^{BC} is the BC element of the matrix Γ_A, then the bilinear form[5]

$$\Gamma_A^{BC}y_B^*x_C \equiv (y^\dagger\Gamma_A x) \quad (13)$$

behaves under σ_{AB} as the vector form x_A does under Σ_{AB}. The coordinates y_B and x_C are taken as real; complex conjugation of one member is required for operation of the σ_{AB}. The transformation of Eq. (13) under Σ_{AB} is singular. We take now the point of view that all transforming objects x, y are vectors, but that there exist two inequivalent rotation operators, Σ_{AB} for linear and σ_{AB} for bilinear forms. Inequivalence means that their roles cannot be interchanged.

Consider now the transformation of a wavefunction $\varphi_{ABC} \cdots (x)$. The intrinsic geometrical properties of the wavefunction are specified by the indices $A, B, \cdots$; in fact, the indices cannot have any other meaning than geometrical, if the properties of the particle all reside in one or another of several "spaces". The indices on φ thus play the same role as those on the coordinates x_A, the basic wavefunction in n-dimensional space being φ_A. To preserve the point of view in the preceding paragraph, one may say that for $n = 4$ and $n = 8$ there exist two inequivalent operations of a rotational character,

[2] For $n = 4$ the corresponding Dirac $\Gamma_A (A = 0, 1, 2, 3)$ must be constructed by recognizing that the usual spin matrices σ and ρ are given by

$$\sigma_A = \tfrac{1}{2}g_{ABC}\Sigma_{BC} + \Sigma_{0A}, \quad \rho_A = \tfrac{1}{2}g_{ABC}\Sigma_{BC} - \Sigma_{0A},$$

with $A = 1, 2, 3$.

[3] E.g., F. D. Murnaghan, *The Theory of Group Representations* (Johns Hopkins Press, Baltimore, Maryland, 1938).

[4] L. E. Dickson, Ann. Math. 20, 155 (1918).

[5] for $n = 8$ the role of Γ_0 is filled by the quantity $1_{00} - \sum_{F\neq 0} 1_{FF}$; see Refs. 4 and 1, in which latter Γ_0 is called U.

specified by Σ_{AB} and σ_{AB}, to apply to φ_A. Under Σ_{AB} the φ_A itself transforms like an elementary n-vector; under σ_{AB} the bilinear form $\Gamma_A^{BC}\varphi_B^*\varphi_C = (\varphi^+\Gamma_A\varphi)$ transforms in this way. Since application of σ_{AB} to the linear or Σ_{AB} to the bilinear form yields geometrical nonsense, we may classify all our wavefunctions into two groups: the φ_A subject only to Σ_{AB}, and the ψ_A subject only to σ_{AB}. This classification is obviously in one-to-one correspondence with what we ordinarily call the *statistics* of the wavefunction: Any number of identical φ_A can be created or destroyed (Bose–Einstein statistics); but the ψ_A must always come in creation–destruction pairs[6] $(\psi^+\Gamma_A\psi)$, so that no more than one of any particular variety can exist (Fermi–Dirac statistics). The possibility for *both* types of statistics to exist without some geometrical preference for one or the other—such as a basic representation of smaller dimensions—is a unique feature of spaces with $n = 4$ or 8 axes.

Reversing the order of the argument, one may postulate the existence of both commutative and anticommutative statistics, and the lack of any distinguishing preference inherent in the geometry of the embedding space. Then $n = 4$ or 8; and as these values are observed in practice, one may suspect the correctness of the postulate.

For real coordinate space these arguments simply reflect the well-known correlation of spin and statistics; it strengthens our conclusion to note that the correlation can be proved only with fully relativistic wavefunctions ($n = 4$) and not with their nonrelativistic approximations ($n = 3$). We have tacitly excluded the indefinite metric, which is the only way presently known to avoid the spin-statistics correlation.

The same arguments carry over immediately to charge space: in the absence of an indefinite metric, the correlation of σ_{AB} with F–D and Σ_{AB} with B–E statistics. This has been called the "correlation between real and charge-space statistics" and earlier references are given in Ref. 1; it now appears that "identity" would be a better word than "correlation" in this phrase.

It is interesting to note that this identity precludes the possibility that real and charge space are two subsections of one manifold; for this manifold must then be R_n with $n > 8$, and the basic postulate of nonpreference between B–E and F–D statistics would be violated. If charge space had only $n = 4$, then charge and real space together might compose an acceptable 8-dimensional manifold. It is fortunate that nature does not present us with this problem, although it is not obvious why.

It is known that the Maxwell and Dirac equations can be displayed in six-dimensional form; this appears to have no special significance in the light of the arguments above.

[6] The appearance of γ_0 in $\bar{\psi} = \psi^\dagger\gamma_0$ for real space is directly associated with the metric requiring the 0 axis to be pure imaginary; no need is yet apparent for a charge-space analog.

Lower Bounds on the Lehmann Weights in Spin-Zero Meson Theory

C. R. HAGEN
Department of Physics and Astronomy, University of Rochester, Rochester, New York
(Received 14 January 1964)

It is shown within the framework of conventional spin-zero meson theory that, if all renormalizations are assumed finite, the Lehmann weights of the fermion and meson Green's functions cannot decrease arbitrarily rapidly as a function of energy.

IN a recent publication[1] an attempt has been made to free quantum electrodynamics from the divergences which have long been its most serious flaw. The central feature of this work was the self-consistent calculation of the asymptotic behavior of the Lehmann weight corresponding to the electron Green's function. Since it is not at all clear at present whether those features peculiar to electrodynamics will preclude a generalization of this technique to the more interesting domain of strong interactions, it is tempting to speculate upon the possibilities inherent to such an approach.

One such effort in this direction has been made by Acharya and Han[2] who calculated the pion–nucleon coupling constant from the second-order Lehmann weight of the nucleon Green's function and the experimentally observed pion–nucleon mass ratio. Crucial to this calculation was the condition used by Johnson *et al.* that the fermion bare mass be zero. Since there has been no theoretical support to date for such an extrapolation to the realm of strong interactions, it is well to focus attention here on some general conditions which must be satisfied in possible future theories.

As in Ref. 1, we take the view that the divergence difficulties of quantum field theory are the result of an inadequate perturbation theory rather than the manifestation of any intrinsic properties of the full theory. It is shown, nonetheless, using techniques developed by Lehmann,[3] that in scalar and pseudoscalar meson theory the Lehmann weights must asymptotically possess a lower bound.

We shall consider the system described by the Lagrangian[4]

$$\mathfrak{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m_0\bar{\psi}\psi + g_0\bar{\psi}(1, \gamma_5)\psi\phi + \tfrac{1}{2}(\phi\partial_{\mu}\phi^{\mu} - \phi^{\mu}\partial_{\mu}\phi) - \tfrac{1}{2}\mu_0^2\phi^2 + \tfrac{1}{2}\phi^{\mu}\phi_{\mu} - \tfrac{1}{4}\lambda^2\phi^4, \quad (1)$$

where we have ignored inessential complications which can be introduced in the form of internal kinematical degrees of freedom. The notation $(1, \gamma_5)$ has been introduced in the meson–nucleon interaction term to permit the simultaneous consideration of the scalar and pseudoscalar cases. For complete generality we have included a direct meson–meson interaction which appears in the Lagrangian with a nonpositive coefficient. That such a condition on the sign of the ϕ^4 term is necessary to ensure the existence of a ground state can be easily demonstrated by a straightforward application of techniques used by Baym[5] to establish the inconsistency of the cubic boson interaction. Finally, we might remark that, while for a scalar ϕ it may be possible to include a ϕ^3 interaction, the reader can readily verify that its inclusion modifies none of our conclusions.

The field equations implied by (1) are

$$\left[\gamma^{\mu}\frac{1}{i}\partial_{\mu} + m_0 - g_0(1, \gamma_5)\phi(x)\right]\psi(x) = 0,$$

$$(-\partial^2 + \mu_0^2)\phi(x) = g_0\bar{\psi}(1, \gamma_5)\psi - \lambda^2\phi^3.$$

We define the meson Green's function by

$$\mathcal{G}(x - x') = i\langle 0|\,(\phi(x)\phi(x'))_+\,|0\rangle - i\langle 0|\,\phi(0)\,|0\rangle^2,$$

which by the work of Lehmann has the representation

$$\mathcal{G}(x - x') = \int \frac{dp}{(2\pi)^4} e^{ip(x-x')} \int_0^{\infty} \frac{B(\kappa)\, d\kappa^2}{p^2 + \kappa^2 - i\epsilon},$$

where $B(\kappa)$ is a positive-definite function. Similarly the fermion Green's function

$$G(x - x') = i\epsilon(x - x')\langle 0|\,(\psi(x)\bar{\psi}(x'))_+\,|0\rangle$$

can be written

[1] K. Johnson, M. Baker, and R. S. Willey, Phys. Rev. Letters 11, 518 (1963).
[2] R. Acharya and M. Y. Han (preprint).
[3] H. Lehmann, Nuovo Cimento 11, 342 (1954).
[4] We use a representation of the Dirac algebra in which γ^0 is antisymmetric and γ^k symmetric.
[5] G. Baym, Phys. Rev. 117, 886 (1960); similar remarks concerning the ϕ^4 interaction term have been made by A. Klein, "Proceedings of Seminar on Unified Theories of Elementary Particles," University of Rochester, July 1963.

$$G(x-x') = \int \frac{dp}{(2\pi)^4} e^{ip(x-x')} \times \int_0^\infty d\kappa \left[\frac{A_+(\kappa)}{\gamma p + \kappa} + \frac{A_-(\kappa)}{\gamma p - \kappa}\right].$$

The positive-definite property of the Hilbert space here also guarantees the positive-definite character of $A_+(\kappa)$ and $A_-(\kappa)$. In pion–nucleon theory, $A_\pm(\kappa)$ and $B(\kappa)$ have the form

$$A_+(\kappa) = Z_2\delta(\kappa - m) + \theta_+(\kappa - m - \mu)A'_+(\kappa),$$
$$A_-(\kappa) = \theta_+(\kappa - m - \mu)A_-(\kappa),$$
$$B(\kappa) = Z_3\delta(\kappa^2 - \mu^2) + \theta_+(\kappa^2 - 9\mu^2)B'(\kappa),$$

where Z_2 and Z_3, the wavefunction renormalization constants of the fermion and boson fields, satisfy the conditions

$$0 \leq Z_2 \leq 1, \qquad 0 \leq Z_3 \leq 1.$$

It is perhaps well to mention here that the so-called broken-symmetry theories will admit a Lehmann representation so long as they do not break manifest Lorentz invariance.

It will be more convenient to introduce the corresponding representations for the commutator and anticommutator of two field operators. Thus one has

$$\langle 0|\,[\phi(x), \phi(x')]\,|0\rangle = 2\pi \int \frac{dp}{(2\pi)^4} e^{ip(x-x')}\epsilon(p^0) \times \int_0^\infty \delta(p^2 + \kappa^2)B(\kappa)\,d\kappa^2, \qquad (2)$$

and

$$\langle 0|\,\{\psi(x), \psi^+(x')\}\,|0\rangle = 2\pi \int \frac{dp}{(2\pi)^4} e^{ip(x-x')}\epsilon(p^0) \int_0^\infty d\kappa\, \delta(p^2 + \kappa^2) \times [(\kappa - \gamma p)A_+(\kappa) - (\kappa + \gamma p)A_-(\kappa)]\beta,$$

where $\beta = \gamma^0$ and $\epsilon(p^0) = p^0/|p^0|$. The equal-time commutation relations

$$[\phi(x), \dot{\phi}(x')] = i\delta(\mathbf{x} - \mathbf{x}'),$$
$$\{\psi(x), \psi^+(x')\} = \delta(\mathbf{x} - \mathbf{x}'),$$

imply the familiar sum rules on the Lehmann weights

$$\int_0^\infty d\kappa[A_+(\kappa) + A_-(\kappa)] = 1,$$
$$\int_0^\infty d\kappa^2 B(\kappa) = 1.$$

Similarly, the use of the field equations in conjunction with Eq. (2) yields[6]

$$\int_0^\infty \kappa^2 B(\kappa)\,d\kappa^2 = \mu_0^2 + 3\lambda^2\langle 0|\,\phi^2(0)\,|0\rangle, \qquad (3a)$$

and

$$m_0 = \int_0^\infty \kappa[A_+(\kappa) - A_-(\kappa)]\,d\kappa + g_0\langle 0|\,\phi(0)\,|0\rangle. \qquad (3b)$$

For pseudoscalar ϕ, the last term in (3b) vanishes to yield a more familiar result. Similarly, in the absence of a direct meson–meson interaction, (3a) leads to the usual sum rule for the boson bare mass. Since by assumption μ_0^2 is finite, it follows from the nonfinite character of $\langle 0|\,\phi^2(0)\,|0\rangle$ that $\int_0^\infty \kappa^2 B(\kappa)\,d\kappa^2$ cannot exist for $\lambda \neq 0$. Because the boson wavefunction renormalization is finite, $\int_0^\infty B(\kappa)\,d\kappa^2$ exists, and one concludes

$$\int_{\Lambda^2}^\infty \kappa^2 B(\kappa)\,d\kappa^2 = \infty,$$

where Λ is an arbitrary mass parameter.

Let us now consider the result of a second application of the field equations to the anticommutator. Thus

$$\langle 0|\,\{(\gamma p + m_0)\psi(x), (\gamma p + m_0)\psi^+(x)\}\,|0\rangle|_{x^0=x'^0} = \delta(\mathbf{x} - \mathbf{x}')\int_0^\infty [(\kappa - m_0)^2 A_+(\kappa) + (\kappa + m_0)^2 A_-(\kappa)]\,d\kappa = g_0^2\delta(\mathbf{x} - \mathbf{x}')\langle 0|\,\phi^2(0)\,|0\rangle,$$

and we conclude by previous arguments

$$\int_\Lambda^\infty \kappa^2[A_+(\kappa) + A_-(\kappa)]\,d\kappa = \infty.$$

It is perhaps worth mentioning here that in some respects the assumptions which we have made in this paper are too stringent. For instance, if one takes the meson–meson interaction term to be

$$-\tfrac{1}{4}\lambda^2\phi^4(x) + \tfrac{3}{2}\lambda^2\langle 0|\,\phi^2(0)\,|0\rangle\phi^2(x),$$

which corresponds in fact to an infinite mass renormalization, then Eq. (3a) becomes

$$\int_0^\infty d\kappa^2 B(\kappa)\kappa^2 = \mu_0^2.$$

On the other hand, one can still show by the techniques used here that

$$\int_{\Lambda^2}^\infty \kappa^4 B(\kappa)\,d\kappa^2 = \infty$$

without any reference to the sign of λ^2.

For the meson field, a lower bound on the Lehmann weight has already been obtained for $\lambda \neq 0$.

[6] Equation (3a) has been previously considered by J. W. Moffat, Nucl. Phys. 14, 682 (1960). He did not, however, determine the sign of the meson–meson coupling constant as done above, and therefore found $\mu_0^2 = \infty$ for $\lambda^2 < 0$.

Thus we can now restrict our subsequent considerations to the exceptional case $\lambda = 0$. One has in this instance

$$\frac{1}{i}\langle 0|\,[(-\partial^2+\mu_0^2)\phi(x),(-\partial^2+\mu_0^2)\phi(x')]\,|0\rangle$$
$$= g_0^2\langle 0|\left[\bar{\psi}(1,\gamma_5)\psi,\frac{1}{i}\frac{\partial}{\partial x^{0\prime}}\bar{\psi}(1,\gamma_5)\psi\right]|0\rangle. \quad (4)$$

For $x^0 = x^{0'}$ one can express the right-hand side of Eq. (4) with the aid of the equations of motion and the canonical commutation relations as

$$-g_0^2\delta(\mathbf{x}-\mathbf{x}')\langle 0|\,2m\bar{\psi}(x)(0,1)\psi(x) + \bar{\psi}(x)\boldsymbol{\gamma}\mathbf{p}(1,1)\psi(x)\,|0\rangle.$$

Upon expanding this highly singular function on the light cone and inserting the result into (4), one obtains

$$\mu_0^4 = \int_0^\infty \kappa^4 B(\kappa)\,d\kappa^2 - 2\frac{g_0^2}{\pi}\lim_{\mathbf{x}\to 0}(-\nabla^2, m_0^2-\nabla^2) \times [\delta(\mathbf{x}^2) + \text{less singular terms}]. \quad (5)$$

The right-hand side of (5) is positive-definite and allows one to conclude as before

$$\int_{\Lambda^2}^\infty \kappa^4 B(\kappa)\,d\kappa^2 = \infty.$$

In summary then we have shown in conventional spin-zero meson theory that, if all renormalizations are finite,

$$\int_\Lambda^\infty \kappa^2[A_+(\kappa)+A_-(\kappa)]\,d\kappa = \infty,$$
$$\int_{\Lambda^2}^\infty \kappa^4 B(\kappa)\,d\kappa^2 = \infty,$$

and the stronger statement

$$\int_{\Lambda^2}^\infty \kappa^2 B(\kappa)\,d\kappa^2 = \infty$$

when there exists a direct meson–meson coupling. There are several views which one can take concerning these results among which we might mention the following.

(i) The straightforward application of the field equations and commutation relations used here is not adequate to deal with the highly singular operator products which we have encountered. This point has been discussed in quantum electrodynamics where the problems associated with gauge invariance suggest some techniques for defining these operators.

(ii) The basic couplings to the fundamental spinor fields do not involve spin-zero particles. While this view might enable one to deny the existence of lower bounds such as have been derived here, the necessity of introducing couplings to particles of higher spin would introduce considerable difficulties.

(iii) The renormalizations in conventional meson theory are not all finite. This view is, of course, at present unassailable though hardly attractive.

(iv) The requirement that the Hilbert space have a positive-definite metric should be given up. The introduction of an indefinite metric could conceivably make $\langle 0|\,\phi^2(0)\,|0\rangle$ finite and thus invalidate the bounds derived here.

(v) Finally, one can accept the bounds derived here at face value and require that they be satisfied in future theories which attempt to calculate the asymptotic Lehmann weight.

ACKNOWLEDGMENTS

The author would like to thank Dr. Okubo, Professor Marshak, and Dr. Lurié for helpful discussions and for reading the manuscript.

Lagrangian and Hamiltonian Formalisms with Supplementary Conditions*

MELVIN SCHWARTZ
Department of Physics, Adelphi University, Garden City, New York
(Received 17 October 1963)

The present note generalizes a transformation due to Takahashi. The Takahashi transformation introduces a Hamiltonian formalism in the presence of a single linear supplementary condition imposed on $n + 1$ coordinates. The Takahashi transformation is generalized to treat the case when there are N independent linear supplementary conditions imposed on the differentials of n coordinates. It is shown that the generalized transformation leads to true coordinates when the coefficients of the differentials are coordinate-independent; otherwise the transformation generally leads to quasicoordinates. More general transformations are discussed, the relation with the method of Lagrange multipliers is established, and some problems which arise in connection with quantization are pointed out.

I. INTRODUCTION

TAKAHASHI has developed a technique for introducing a Hamiltonian formalism when there is a single linear supplementary condition imposed on $n + 1$ coordinates.[1] In the present note, the more general case when there are N independent linear supplementary conditions on the differentials of n coordinates is discussed. The analysis of a well-known problem in Riemannian geometry is utilized to generalize Takahashi's transformation. In particular, it is shown that the generalized transformation can be applied to the coordinates only in the case that the coefficients in the supplementary conditions are not functions of the coordinates on a hypersurface; i.e., the supplementary conditions may be represented by N linear conditions on the coordinates as well as their differentials. When the coefficients in the supplementary conditions are functions of the coordinates, the generalized transformation generally leads to the introduction of quasicoordinates.[2] The equivalence of the present approach to the method of Lagrange multipliers is pointed out.

The case when the supplementary conditions involve the momenta as well as the coordinates is also briefly discussed. The calculational techniques are illustrated by simple examples.

The calculations are all carried out within the framework of classical dynamics. The problem of quantization is then discussed in a final section.

II. CONSTRAINTS

A. Algebraic Coordinate Conditions

Consider a real n-dimensional Euclidean space the points of which are specified by Cartesian coordinates q^μ, $(\mu = 1, \cdots, n)$.

The coordinate q^μ may represent a generalized coordinate of a classical system with a finite number of degrees of freedom or it may represent a field variable in some classical field theory. In the latter case, however, the discussion assumes that the field variable refers to a fixed value of its argument unless it is being considered in relation to a variational principle. Let there be N independent differential supplementary conditions of the form

$$dC^R = \mathbf{a}^R \cdot \mathbf{dq} = a^R_\mu dq^\mu = 0 \quad (R = 1, \cdots, N), \qquad (1)$$

where $\mathbf{a}^R$ and $\mathbf{dq}$ are vectors in the n-dimensional space with Cartesian components a^R_μ and dq^μ, respectively; $\mathbf{a}^R \cdot \mathbf{dq}$ represents a Euclidean scalar product; unless otherwise specified, repeated indices above and hereafter are to be summed over; and the coefficients a^R_μ are presumed to be real algebraic functions of the q's. Since the dC^R are all independent functions of $\mathbf{dq}$ for a given n-tuple $\mathbf{q} = (q^1, \cdots, q^n)$, the vectors $\mathbf{a}^R$ form a set of N linearly independent vectors. Thus the vectors $\mathbf{a}^R$ span a local N-dimensional Euclidean subspace at the point $\mathbf{q}$. Eq. (1) states that the vectors $\mathbf{dq}$ lie in the local $(n - N)$ dimensional Euclidean subspace orthogonal to the vectors $\mathbf{a}^R$.

When the coefficients $a^R_\mu(q)$ satisfy the integrability conditions $a^R_{\mu,\nu} = a^R_{\nu,\mu}$, where $f_{,\mu} \equiv \partial f/\partial q^\mu$, Eqs. (1) are equivalent to the equations

$$C^R(k) = 0, \qquad (R = 1, \cdots, N). \qquad (2)$$

Each of the Eqs. (2) defines an $(n - 1)$-dimensional hypersurface and $a^R_\mu = C^R_{,\mu}$ are the n Cartesian components of the gradient for each hypersurface. Each point $\mathbf{q}$ is restricted to the $(n - N)$-dimensional hypersurface representing the intersection of the N

* Supported in part by the National Science Foundation.
[1] Y. Takahashi, Physics Letters 1, 278 (1962).
[2] E. T. Whittaker, *Analytical Dynamics* (Cambridge University Press, Cambridge, England, 1937), p. 41.

hypersurfaces defined by Eqs. (2) and its local tangent spaces are locally orthogonal to the vectors $\mathbf{a}^R$.

Let $\mathcal{E}_n$ denote the n-dimensional Euclidean subspace originally referred to; let $\mathcal{E}_N$ denote the N-dimensional locally Euclidean subspace spanned by the vectors $\mathbf{a}^R$; and let $\mathcal{E}$ denote the $(n - N)$-dimensional locally Euclidean subspace orthogonal to $\mathcal{E}_N$. Let

$$g^{RS} = \mathbf{a}^R \cdot \mathbf{a}^S \tag{3}$$

denote the contravariant components of the metric tensor in $\mathcal{E}_N$ at a point $\mathbf{q}$. Then

$$D_{\mu\nu} = \delta_{\mu\nu} - g_{RS} a^R_\mu a^S_\nu, \tag{4}$$

where g_{RS} represents the covariant components of the metric tensor and satisfies $g_{RS}g^{ST} = \delta^T_R$, is the local projection operator which projects vectors in $\mathcal{E}_n$ into $\mathcal{E}$. $D_{\mu\nu}$ is a symmetric matrix with eigenvalues zero and one and hence may be diagonalized by a unitary transformation. All vectors in $\mathcal{E}_N$ are eigenvectors of $D_{\mu\nu}$ with eigenvalue zero. All vectors in $\mathcal{E}$ are eigenvectors of $D_{\mu\nu}$ with eigenvalue one. As a projection operator, $D_{\mu\nu}$ is the null operator in $\mathcal{E}_N$ and the unit operator in $\mathcal{E}$. For convenience, we represent $D_{\mu\nu}$ by means of an orthonormal basis in $\mathcal{E}$. Such an orthonormal basis may be constructed as follows. Define $(n - N)$ vectors $\mathbf{a}_i$ with Cartesian components a^μ_i by the equations

$$a^\mu_i = a_{i\mu} = \epsilon_i^{\mu\mu_1\cdots\mu_{i-1}\nu_1\cdots\nu_N} a_{1\mu_1} \cdots a_{i-1\mu_{i-1}} a^1_{\nu_1} \cdots a^N_{\nu_N},$$
$$[i = 1, \cdots, (n - N)], \tag{5}$$

where $\epsilon_i^{\mu\mu_1\cdots\mu_{i-1}\nu_1\cdots\nu_N}$ is a completely anti-symmetric Levi–Cevita density which is equal to unity whenever

$$\mu < \mu_1 < \cdots < \mu_{i-1} < \nu_1 < \cdots < \nu_N;$$

and

$$a^\mu_1 = \epsilon_1^{\mu\nu_1\cdots\nu_N} a^1_{\nu_1} \cdots a^N_{\nu_N}.$$

Hereafter it is to be understood that Greek indices run from 1 to n; capital Latin indices run from 1 to N; and lower case Latin indices run from 1 to $(n - N)$. Let $\hat{a}_i$ denote the unit vectors formed from $\mathbf{a}_i$. From Eq. (5) it obviously follows that

$$\hat{a}_i \cdot \hat{a}_j = \delta_{ij} \quad \text{and} \quad \hat{a}_i \cdot \mathbf{a}^R = 0. \tag{6}$$

Thus the vectors $\hat{a}_i$ form a suitable local orthonormal basis for $\mathcal{E}$.

At a point $\mathbf{q}$ on the hypersurface, each vector $\mathbf{dq}$ in $\mathcal{E}$ may be expressed as a linear superposition of the vectors $\hat{a}_i$.

$$\mathbf{dq} = \hat{a}_i d\eta^i, \tag{7}$$

$$d\eta^i = \hat{a}_i \cdot \mathbf{dq} = \hat{a}_{i\mu} dq^\mu. \tag{8}$$

The differentials $d\eta^i$ form a set of $(n - N)$ independent differentials in terms of which, each differential dq^μ is expressed through Eq. (7). It follows from Eqs. (6)–(7) that

$$\mathbf{dq}^2 = d\eta^i d\eta_i, \tag{9}$$

where $d\eta_i = d\eta^i$, so that $d\eta^i$'s represent the projections of an element of arclength in the hypersurface on to the tangent vectors $\hat{a}_i$.

Equation (4) represents the generalization of Takahashi's Eq. (4).[1] The vectors $\hat{a}_i$ together with the vectors $\mathbf{a}_R$, yield an explicit representation of the unitary matrix diagonalizing $D_{\mu\nu}$. In addition, Eqs. (7)–(8) represent the generalization of Takahashi's Eqs. (7) and (14).[1] We now investigate the circumstances under which the differentials $d\eta^i$ are the differentials of a permissible set of parameters for specifying the points $\mathbf{q}$ of the hypersurface; i.e., the circumstances under which Eq. (7) may be integrated to yield each q^μ as a function of $n - N$ independent η^i's.

If the integrability conditions $\hat{a}_{i,j} = \hat{a}_{j,i}$, where $f_{,i} \equiv \partial f/\partial \eta^i$, are satisfied, Eq. (7) may be integrated to yield $q^\mu = q^\mu(\eta^1, \cdots \eta^{n-N})$. It is seen from Eq. (9), however, that the metric tensor is that for a flat space so that the hypersurface is isometric to a hyperplane. If the hypersurface is a hyperplane, the vectors $\mathbf{a}_R$ must be constant along the hyperplane. Equations (2) may then be taken as representing a family of hyperplanes of dimension $(n - 1)$ each of which is orthogonal to the $(n - N)$-dimensional hyperplane representing their intersection. The equations for the hyperplanes are given by a direct integration of Eqs. (1) resulting in N linear supplementary conditions on the coordinates.

Thus we have

"The variables η^i defined in Eq. (8) are acceptable parameters for the $(n - N)$-dimensional hypersurface if and only if the hypersurface is flat. Eqs. (2) may then be taken to represent a family of $(n - 1)$-dimensional hyperplanes if the hypersurface is a hyperplane."

In general, whether or not the vectors $\hat{a}_i$ may be tangents to a family of orthogonal coordinate curves on the hypersurface will depend on the vectors $\mathbf{a}^R$ which determine the hypersurface. In any case, as is well known, given Eqs. (2), one can always find an acceptable set of parameters for the hypersurface. The main feature of transformation equations such as Eqs. (7)–(8) is that none of the q^μ are given preferential treatment. Moreover, even if the hypersurface is not flat, Eqs. (8) represent a set of independent differentials $d\eta^i$ which may be considered as quasicoordinates.[2] Such coordinates may

be used to obtain an independent set of Euler–Lagrange equations as follows.

For definiteness, consider the action integral

$$S = \int dt\, L(\mathbf{q}, \dot{\mathbf{q}}, t). \tag{10}$$

Utilizing Eq. (7), the general variation of S may be written in the form

$$\bar{\delta}S = \int dt \left\{ \left[\frac{\partial L}{\partial q^\mu} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^\mu}\right)\right] \hat{a}_i^\mu \delta\eta^i + \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}^\mu}\hat{a}_i^\mu \bar{\delta}\eta^i - \left(\frac{\partial L}{\partial \dot{q}^\mu}\hat{a}_i^\mu \dot{\eta} - L\right)\bar{\delta}t \right] \right\}, \tag{11}$$

where $\bar{\delta}f = \delta f + \dot{f}\bar{\delta}t$ is the total variation of f and a dot over a symbol denotes total time derivative. Since the $\delta\eta^i$ are independent, the Euler–Lagrange equations are

$$\left[\frac{\partial L}{\partial q^\mu} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^\mu}\right)\right] \hat{a}_i^\mu = 0. \tag{12}$$

As discussed above, the transformation equations (7)–(8), using the orthonormal set $\hat{a}_i$, leads to quasi-coordinates unless the vectors $\mathbf{a}^R$ are constant along the hypersurface. The latter remark holds whether or not Eqs. (1) may be integrated to yield Eqs. (2). Obviously however, whenever Eqs. (2) obtain, one can always find sets of vectors $\hat{a}_i$ (not necessarily orthonormal) which may be used to effect a transformation to true coordinates. If the vectors $\hat{a}_i$ are not orthogonal, then they must be coupled to a reciprocal set satisfying the condition that $\hat{a}_i \cdot \hat{a}^j = \delta_i^j$. In any case, it is clear that Eqs. (12) will still represent the Euler–Lagrange equations for the q^μ.

If Eq. (12) is multiplied on the right by $\hat{a}^{i\nu}$ and summed over i, we get

$$\left[\frac{\partial L}{\partial q^\mu} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^\mu}\right)\right] D^{\mu\nu} = 0, \tag{13}$$

where $D^{\mu\nu} = D_{\mu\nu}$ since as defined in Eq. (4), $D_{\mu\nu}$ is a Cartesian tensor. Thus, without even constructing vectors $\hat{a}_i$, one can always write Eq. (13) by using Eq. (4). Equations (12) represent $(n - N)$ equations for n variables so that appropriate use must be made of Eqs. (1); but Eqs. (13) are n equations for n variables and may be used as they stand.

Note that Eqs. (13) could also have been obtained directly using the method of Lagrange multipliers. For, upon multiplying each of Eqs. (1) by a multiplier λ_R and summing over R, one finds from the resulting Euler–Lagrange equations

$$\lambda_R = \left[\frac{\partial L}{\partial k^\mu} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^\mu}\right)\right] g_{RS} a^{S\mu}. \tag{14}$$

Thus either approach leads to the same Euler–Lagrange equations in terms of the q^μ.

Of course, the use of the orthonormal set $\hat{a}_i$ defined by Eqs. (5) is a matter of choice. Any set of $(n - N)$ linearly independent vectors $\mathbf{a}_i$ each of which is orthogonal to the vectors $\mathbf{a}^R$ may be used to represent $D_{\mu\nu}$. When it is possible to transform from the q^μ to an independent set of true coordinates, the set of vectors $\mathbf{a}_i$ will be the tangent vectors of a set of admissible coordinate curves on the hypersurface. Normalizing the vectors $\mathbf{a}_i$, the general form of the transformation equations which replace Eqs. (7)–(8) will then be

$$\mathbf{dq} = \hat{a}_i f_i(\mathbf{n}) d\eta^i, \tag{15}$$

$$f_{i|}(\mathbf{n}) d\eta^{i|} = \hat{a}^i \cdot \mathbf{dq} = g^{ij}\hat{a}_j \cdot \mathbf{dq}, \tag{16}$$

where a vertical bar following an index indicates no summation over that index; $g_{ij} = \hat{a}_i \cdot \hat{a}_j$ are the covariant components of the metric tensor in $\mathcal{E}$ and g^{ij} the contravariant components formed from the basis vectors $\hat{a}^i$; the vectors $\hat{a}^i$ are reciprocal to the vectors $\hat{a}_i$.

When the η^i are true coordinates, the momenta conjugate to η^i will be

$$\pi_i = \frac{\partial L}{\partial \dot{\eta}^i} = \frac{\partial L}{\partial \dot{q}^\mu}\frac{\partial \dot{q}^\mu}{\partial \dot{\eta}^i} = \frac{\partial L}{\partial \dot{q}^\mu}\hat{a}_{i|}^\mu f_{i|}(\mathbf{n}) = \frac{\partial L}{\partial \dot{q}^\mu} a_i^\mu. \tag{17}$$

If the matrix $\partial^2 L/\partial\dot{\eta}^i\partial\dot{\eta}^j$ is nonsingular, the momenta π_i will be independent. If $\partial^2 L/\partial\dot{\eta}^i\partial\dot{\eta}^j$ is singular, there will be constraints and we must resort to the methods of Dirac[3] or equivalently, Bergmann and Goldberg.[4] Assuming the former case and multiplying Eq. (17) by $a^{i\nu}$, the reciprocal of a_i^ν, we obtain

$$p^\mu = D^{\mu\nu}(\partial L/\partial \dot{q}^\nu) = a^{i\mu}\pi_i, \tag{18}$$

with p^μ being considered the momentum conjugate to q^μ. Assuming that η^i and π_i satisfy the standard Poisson bracket relations and using Eq. (15), we then find that

$$[q^\mu, p^\nu] = D^{\mu\nu} \tag{19}$$

is the Poisson bracket between q^μ and p^ν. In quantum mechanics, Eq. (19) would have an $i = (-1)^{\frac{1}{2}}$ on the right-hand side. The Hamiltonian will be given by the coefficient of $\bar{\delta}t$ in Eq. (11) when $\hat{a}_i$ is replaced by $\mathbf{a}_i$.

B. Differential Coordinate Conditions

In Sec. A above, the vectors $\mathbf{a}^R$ were assumed to be algebraic functions of the q^μ. When the vectors

[3] P. A. M. Dirac, Can. J. Math. **2**, 129 (1950).
[4] P. G. Bergmann and I. Goldberg, Phys. Rev. **98**, 531 (1955).

$\mathbf{a}^R$ also involve derivatives of the q^μ, we must treat problems involving a finite number of degrees of freedom separately from field theories when spacial derivatives occur in the latter situation. When only time derivatives occur, the methods of the last section are easily generalized and one is still led to Eqs. (13) and (14) when Eq. (1) obtains.

The methods of the last section may also be applied to field theories when the vectors $\mathbf{a}^R$ in Eqs. (1) are all linear differential operators with no dependence on field variables. However, the field variables must then belong to the domain of the $\mathbf{a}^R$ allowing inverse operators to be defined. Equations such as Eq. (2) will obtain which involve linear operators acting on the field variables. Such equations will also be scalars with respect to coordinate transformations. The gauge conditions in electrodynamics are examples of such equations.

When the vectors $\mathbf{a}^R$ are linear differential operators independent of the field variables, one can take Fourier transforms and deal with the problem in k space. The supplementary conditions are then of the algebraic type with the vectors $\mathbf{a}^R$ still independent of the field vairables.

C. Constraints in Phase Space

When phase-space constraint equations of the form

$$C^R(p^\mu, q^\mu) = 0$$

exist, one may use the ideas of Sec. A above to develop a transformation to independent parameters which may be used to specify the constraint hypersurface in phase space. Once that is done, however, one must then resort to the canonical formalism of Bergmann and Goldberg.[4]

III. ILLUSTRATIVE EXAMPLES

A. Particle on a Sphere

Using Cartesian coordinates x^1, x^2, x^3, the algebraic coordinate condition is

$$x_i x^i - \text{const} = 0, \tag{20}$$

where $x_i = x^i$ since they are Cartesian coordinates. Since $N = 1$, we have $\mathbf{a} = \mathbf{r}$, where $\mathbf{r}$ is the position vector with components (x^1, x^2, x^3). Introducing the unit radial vector $\hat{r} = \mathbf{r}/r$, we have

$$D_{\mu\nu} = \delta_{\mu\nu} - \hat{r}_\mu \hat{r}_\nu. \tag{21}$$

From Eq. (5) we get

$$\left.\begin{aligned} a_1^\mu &= \epsilon^{\mu\nu}\hat{r}_\nu, \\ a_2^\mu &= (\mathbf{a}_1 \times \hat{r})^\mu. \end{aligned}\right\} \tag{22}$$

However, because of spherical symmetry, it is simpler in the present case to introduce the unit vectors appropriate to spherical coordinates. Let θ and φ denote the spherical polar angles with φ being the polar angle in the $x^1 - x^2$ plane. Let $\hat{\theta}$ and $\hat{\varphi}$ denote the unit vectors tangent to the coordinate curves on the sphere such that $\hat{r}$, $\hat{\theta}$, and $\hat{\varphi}$ form a right-handed system. Then

$$D_{\mu\nu} = \hat{\theta}_\mu \hat{\theta}_\nu + \hat{\varphi}_\mu \hat{\varphi}_\nu. \tag{23}$$

Using s_θ and s_φ in place of η_1 and η_2, Eqs. (7)–(8) become

$$\mathbf{dr} = \hat{\theta} ds_\theta + \hat{\varphi} ds_\varphi, \tag{24}$$

$$\left.\begin{aligned} ds_\theta &= \hat{\theta}\cdot\mathbf{dr} = r\, d\theta \\ ds_\varphi &= \hat{\varphi}\cdot\mathbf{dr} = r \sin\theta\, d\varphi \end{aligned}\right\}, \tag{25}$$

where in Eqs. (25) we have utilized appropriate expressions in spherical coordinates. As is well known, Eq. (24) is not an exact differential when s_θ and s_φ are considered independent variables and are quasicoordinates. On the other hand, of course, θ and φ are true coordinates. For a Lagrangian of the form

$$L = \tfrac{1}{2} m\mathbf{v}^2 - V(\mathbf{r}), \tag{26}$$

Eqs. (12) become

$$\left.\begin{aligned} \hat{\theta}\cdot\frac{d}{dt}(m\mathbf{v}) &= -\frac{1}{r}\frac{\partial V}{\partial \theta}, \\ \hat{\varphi}\cdot\frac{d}{dt}(m\mathbf{v}) &= -\frac{1}{r\sin\theta}\frac{\partial V}{\partial \varphi}, \end{aligned}\right\} \tag{27}$$

which are the familiar equations for the problem in spherical coordinates. Finally, Eqs. (17) take the form

$$\left.\begin{aligned} \pi_\theta &= m\mathbf{v}\cdot\hat{\theta} \\ \pi_\varphi &= m\mathbf{v}\cdot\hat{\varphi} \end{aligned}\right\}, \tag{28}$$

while Eq. (18) yields

$$\mathbf{p} = m\mathbf{v}, \tag{29}$$

where use has been made of the supplementary condition, $\dot{r} = 0$.

B. Vacuum Electrodynamics in Configuration Space

Let the supplementary condition be the Coulomb gauge condition

$$\nabla\cdot\mathbf{A} = 0, \tag{30}$$

where $\mathbf{A}(\mathbf{x})$ represents the vector potential. It is seen from Eq. (30) that $\mathbf{a} = \nabla$ so that care must

be taken to maintain the proper sequence of factors throughout. If solutions for $\mathbf{A}$ are restricted to the domain of the inverse operator $(\nabla^2)^{-\frac{1}{2}}$, then we may write

$$D_{\mu\nu} = \delta_{\mu\nu} - \partial_\mu\partial_\nu(\nabla^2)^{-1} \tag{31}$$

where $\partial_\mu = \partial/\partial x^\mu$. Let $\hat{e}_1$ be a unit vector in the positive one direction. Let

$$a_1^\mu = \hat{e}_1^\nu D_{\mu\nu} = \hat{e}_1^\mu - \hat{e}_1\cdot\nabla\partial_\mu(\nabla^2)^{-1}. \tag{32}$$

$\mathbf{a}\cdot\mathbf{a}_1 = 0$ and we set

$$\mathbf{a}_2 = \mathbf{a}_1 \times \nabla. \tag{33}$$

The inverses of $\mathbf{a}_1^2$ and $\mathbf{a}_2^2$ are needed to form $\hat{a}_1$ and $\hat{a}_2$ so that we assume $\mathbf{A}$ is also restricted to the domain of such inverses.

Since Eq. (30) is linear in the field variables, Eqs. (7)–(8) may be directly integrated in the present case to yield forms linear in the field variables also. Replacing η^i by $\mathfrak{A}^i(\mathbf{x})$ and keeping the operators $\hat{a}_i$ to the left of field variables, we have

$$\mathbf{A}(\mathbf{x}) = \hat{a}_i\mathfrak{A}^i(\mathbf{x}), \tag{34}$$

$$\mathfrak{A}^i(\mathbf{x}) = \hat{a}_i\cdot\mathbf{A}(\mathbf{x}). \tag{35}$$

From the Lagrangian

$$L = \frac{1}{2}\int d^3x[(\partial_t\mathbf{A})^2 - (\nabla\times\mathbf{A})^2] \tag{36}$$

and Eqs. (17)–(18), we find

$$\pi_i = \hat{a}_i\cdot\partial_t\mathbf{A},$$

$$\mathbf{P}_A = \partial_t A = \hat{a}^i\pi_i.$$

Equation (13) becomes

$$(\nabla^2 - \partial_t^2)\mathbf{A} = 0. \tag{39}$$

Assuming $\mathfrak{A}^i$ and π_i are canonical conjugates so that

$$[\mathfrak{A}^i(\mathbf{x}, t), \pi_j(\mathbf{x}', t)] = \delta_{ij}\delta(\mathbf{x} - \mathbf{x}'), \tag{40}$$

Eq. (19) takes the well-known form

$$[A_\mu(\mathbf{x}, t), \partial_t A_\nu(\mathbf{x}, t)]$$

$$= \left[\delta_{\mu\nu}\delta(\mathbf{x} - \mathbf{x}') - \frac{1}{4\pi}\partial_\mu\partial_\nu'\left(\frac{1}{|\mathbf{x} - \mathbf{x}'|}\right)\right]. \tag{41}$$

C. Vacuum Electrodynamics in Momentum Space

We again assume the Coulomb gauge and take the Fourier transform of $\mathbf{A}(\mathbf{x})$. The basic equations in the present case may then be obtained by taking the Fourier transforms of Eqs. (30)–(41). In particular $\nabla \to ik$, the propagation vector multiplied by $(-1)^{\frac{1}{2}}$ and $\hat{a}_i \to \hat{e}_i$, the polarization vector.

Of course, the Fourier transforms of Eqs. (31) and (41) are well known. However, the use of a reduced Lagrangian to introduce a reduced Hamiltonian formalism is generally not presented in the standard literature.

IV. DISCUSSION

In the sentence below Eq. (18), reference is made to its quantum mechanical analog. The statement requires elaboration because $D^{\mu\nu}$ is generally a function of functions of the q^μ and inverses of such functions. Such inverses arise in connection with the transformation from $(q^\mu, p^\mu) \to (\eta^i, \pi_i)$. If the q^μ and p^μ are to be treated as quantum operators, the required inverses must exist in the domain of physical states. The latter situation also presumes that no undefined inverses are introduced in transforming from $\dot{\eta}^i \to \pi_i$ in going from the Lagrangian to the Hamiltonian formalism. In any case, the above problem does not arise when the supplementary conditions are algebraic and linear in the q^μ. In the latter circumstance, the vectors $\mathbf{a}^R$ and $\mathbf{a}_i$ are all c numbers.

Singular Bethe–Salpeter Scattering Amplitudes*

ARTHUR R. SWIFT† AND BENJAMIN W. LEE‡
Department of Physics, University of Pennsylvania, Philadelphia, Pennsylvania

The Bethe–Salpeter equation for scattering is investigated in configuration space for the class of singular "potentials" (i.e., a $\lambda\phi^4$ theory in the ladder approximation) which behave as r^{-4} near the light cone. The discussion relies on the similarity between solutions of the Bethe–Salpeter equation and the Schrödinger equation where the corresponding problem is scattering by a r^{-2} potential. Through a consideration of the asymptotic properties of the two-particle, free Green's function, the elastic scattering amplitude is shown to be the coefficient of the outgoing wave part of the wave-function, just as it is in the nonrelativistic case. At zero total energy, it is just the coefficient of $e^{-mr}/r^{\frac{1}{2}}$. The differential form of the Bethe–Salpeter equation is expanded in four-dimensional spherical harmonics, and the singular part of the potential is incorporated into the differential operator. The resulting equation is formally solved by converting it into what is now a Fredholm integral equation. Care is taken to choose the proper asymptotic behavior for the solutions of the new equation. The discussion of the singular potential is carried out at zero total energy in order to obtain spherical symmetry. The technique for handling the singular potential and extracting the T matrix at zero energy is demonstrated by application to two examples. The exact scattering amplitude is found for exchange of two massless mesons. A first-order solution is obtained for a phenomenological potential that approximate the exchange of two massive mesons. This solution exhibits many of the features expected from a truly physical potential.

I. INTRODUCTION

THE purpose of this paper is to discuss the properties of a class of singular Bethe–Salpeter[1] equations in the ladder approximation. The Bethe–Salpeter equation, originally developed more than ten years ago, has recently become the object of intensive study.[2–6] Primary among the reasons for this renewed interest is the fact that the Bethe–Salpeter equation promises to provide a nonperturbative, fully covariant, approach to two-body problems. In addition the equation includes the features of analyticity and elastic unitarity demanded of physically meaningful theories. However, before all these virtues can be fully exploited, much work must be done on the general nature of the equation. In the absence of any scheme to completely solve the Bethe–Salpeter equation there are two ways of extracting information. Either very general properties are elicited from the symmetry, asymptotic structure, and singularities of the equation; or particular values of the parameters are chosen to make the equation exactly soluble. In this paper we will apply both approaches quite extensively. For example, we will choose our total energy to be zero in order to obtain four-dimensional spherical symmetry; the most singular terms of the general equation have this symmetry so our results should be significant.

We restrict our discussion to a consideration of meson scattering by means of a class of singular potentials. By a singular potential we will mean one that behaves as r^{-4} at the origin of a four-dimensional Euclidean space; on account of that behavior the corresponding Bethe–Salpeter equation is no longer of the Fredholm type. It might be mentioned that the potential for a $\lambda\phi^4$ theory has this r^{-4} behavior in the ladder approximation. Several investigators have considered the problem from the bound-state point of view and have shown how to convert the singular equation into a Fredholm one.[5,6] We refer the reader to this work for a complete discussion of the classification of potentials and the problems in treating singular ones.

It turns out that the conversion into a Fredholm equation is best accomplished in configuration space, and it is in configuration space that the interpretation of the Bethe–Salpeter amplitude is most obscure. Therefore, we devote some effort to demonstrating how, in analogy with the Schrödinger wavefunction, the Bethe–Salpeter "scattering wavefunction" can be used to calculate the scattering amplitude. To do this we examine its large spacelike distance limit.

* This work is supported in part by the U. S. Atomic Energy Commission.

† National Science Foundation Predoctoral Fellow. This work is part of a Ph.D. thesis to be submitted to the University of Pennsylvania.

‡ Alfred P. Sloan Foundation Fellow.

[1] H. A. Bethe and E. E. Salpeter, Phys. Rev. **84**, 1232 (1951); M. Gell-Mann and F. E. Low, Phys. Rev. **84**, 350 (1951).

[2] R. F. Sawyer, Phys. Rev. **131**, 1384 (1963).

[3] A. R. Swift and B. W. Lee, Phys. Rev. **131**, 1857 (1963).

[4] M. Baker and I. J. Muzinich, Phys. Rev. **132**, 2201 (1963).

[5] A. Bastai, L. Bertocchi, S. Fubini, G. Furlan, and M. Tonin, Nuovo Cimento **30**, 1512 (1963).

[6] A. Bastai, L. Bertocchi, G. Furlan, and M. Tonin, Nuovo Cimento **30**, 1532 (1963).

For a center of mass energy greater than $2m$, where m is the mass of the scattered particles, the asymptotic form at a finite time is strongly reminiscent of that of the Schrödinger wavefunction. In particular the T matrix is given by the coefficient of e^{ipr}/Er (where E is the center-of-mass energy), just as it is in the nonrelativistic case. On the other hand, for a total energy less than or equal to $2m$, the asymptotic behavior is considerably more complicated. At zero energy the behavior of the "outgoing" wave is simply $e^{-mr}/r^{\frac{1}{2}}$ with the T matrix again as its coefficient. The discontinuous behavior of the asymptotic form at $2m$ is a reflection of the analyticity properties of the Bethe–Salpeter amplitude as a function of the total energy squared; the point $4m^2$ turns out to be a branch point of the two-particle, free Green's function. The behavior at zero total energy, explainable in terms of an accumulation phenomenon, is sufficiently different from that of finite energy to render suspect any efforts to place much emphasis on the exact form of the solutions. It is true, however, that techniques developed for solving the zero-energy equation should be applicable to future attacks on the finite-energy problem. It is with this expectation that we specialize our discussion of singular potentials to the zero-energy case. We should also mention that the zero-energy results do have some physical use; when used in crossing relations, they give the forward scattering amplitudes in the crossed channel.

Although Bastai, *et al.*[5] have given a complete discussion of the technique for removing the singular part of a potential, the method they develop leads to difficulties in the scattering problem. Therefore, we present an alternate approach to this question which is more useful in a treatment of scattering in as much as it does not generate a r^{-2} potential, which behaves badly at infinity. The equation we obtain is of the Fredholm type, but care has to be exercised in choosing the free solutions with the correct asymptotic behavior. We demonstrate that a knowledge of the asymptotic form of the exact solution, coupled with the requirement of regularity at the origin, is sufficient to uniquely determine the correct solution of the altered equation. The situation again is very analogous to that prevailing in potential theory, as we explicitly demonstrate in Appendix C.

In Sec. II of this paper we present a general discussion of the Bethe–Salpeter equation in order that this paper may be relatively self-contained. Section III is devoted to a treatment of the two-particle, free Green's function with particular emphasis on its asymptotic and analytic properties. That information is used to extract the T matrix from the zero-energy and physical energy amplitudes. Thus, Secs. II and III are intended to be preliminary to the discussion of the singular potential. Most of the mathematical details for Sec. III are worked out in Appendices A and B. In Sec. IV we discuss the singular potential. Then, in Sec. V as an illustration of technique we obtain the T matrix for elastic scattering of mesons by exchange of zero-mass mesons and for a more realistic, though nonphysical, potential. Appendix C is devoted to treating the Schrödinger equation in the same manner as we have the relativistic equation in order to demonstrate further the analogy between the two problems.

II. THE GENERAL EQUATIONS FOR NONZERO ENERGY

As the starting point of our discussion of the scalar Bethe–Salpeter equation, we write down the ladder approximation form of the equation in momentum space.[7]

$$\Psi(p, q, E) = \Psi_0(p, q, E) + \int \frac{d^4k}{(2\pi)^4} \frac{\Psi(p, k, E)V(k - q)}{[(k + \tfrac{1}{2}E)^2 - m^2][(k - \tfrac{1}{2}E)^2 - m^2]}, \quad (1)$$

where E is the total energy of the system in the center of mass, while p and q are the relative four-momenta of the initial and final particles, respectively. Here m is the mass of the particles being scattered; they are considered to be identical but distinguishable particles. When p and q are taken to have their physical values for elastic scattering ($p^2 = q^2$, $p^2 + m^2 = \frac{1}{4}E^2$), $\Psi(p, q, E)$ becomes the T matrix for meson–meson scattering. $\Psi_0(p, q, E)$ is given by

$$\Psi_0(p, q, E) = [1/(2\pi)^4]V(p - q)$$

and $V(p - q)$ is of the form of an integral over the products of propagators for the exchanged particles of the ladder diagrams; it can be taken to be the Fourier transform of a potential. For example, in the case of a $\lambda\varphi_1^2\varphi_2^2$ theory, $V(p - q)$ can be written as

$$\begin{aligned} V(p - q) &= \int d^4x\, e^{i(p-q)\cdot x} V(x) \\ &= \frac{(i\lambda)^2}{(2\pi)^4} \int \frac{d^4k}{[k^2 - \mu^2][(k - p + q)^2 - \mu^2]} \\ &= \frac{\lambda^2}{4} \int d^4x\, e^{i(p-q)\cdot x} \Delta_F^2(x), \end{aligned} \quad (2)$$

[7] Our metric is defined by $g_{00} = -g_{ii} = 1$.

where $\Delta_F(x)$ is the usual causal Green's function[8] for the Klein–Gordon equation and μ is the mass of the exchanged particles.

If $\Psi(x)$ is defined as the Fourier transform of $\Psi(p, q, E)$,

$$\Psi(x) = \frac{1}{(2\pi)^4}\int d^4q\, e^{i(q)\cdot x}\Psi(p, q, E),$$

it is found to satisfy an integral equation of the form

$$\Psi(x) = \frac{1}{(2\pi)^4}\, V(x)e^{ip\cdot x} + V(x)\int d^4x'\, G(x, x')\Psi(x') \qquad (3)$$

where

$$G(x, x') = \frac{1}{(2\pi)^4}\int \frac{d^4k\, e^{ik\cdot(x-x')}}{[(k+\frac{1}{2}E)^2 - m^2][(k-\frac{1}{2}E)^2 - m^2]}. \qquad (4)$$

If we let $\Psi(x) = V(x)\Phi(x)/(2\pi)^4$, we find that

$$\Phi(x) = e^{ip\cdot x} + \int d^4x'\, G(x, x')V(x')\Phi(x'). \qquad (5)$$

This equation is recognizable as the integral Bethe–Salpeter equation in configuration space for the scattering problem.[9] Equation (5), although not of the Fredholm type for singular potentials, is nevertheless meaningful. Using the known behavior of $\Phi(x)$ at zero and infinity, we can show that the integral in (5) is convergent. The differential equation and boundary condition coorresponding to (5) is

$$\left[\left(\frac{\partial}{\partial x} + i\frac{E}{2}\right)^2 + m^2\right]\left[\left(\frac{\partial}{\partial x} - i\frac{E}{2}\right)^2 + m^2\right] \times [\Phi(x) - \Phi_0(x)] = V(x)\Phi(x) \qquad (6)$$

with $\Phi_0(x) = e^{ip\cdot x}$. The first point to be noticed is that the above relations provide a direct connection between the T matrix and the Bethe–Salpeter amplitude in configuration space:

$$T(p, q, E) = \int d^4x\, e^{-iq\cdot x}\Psi(x) = \frac{1}{(2\pi)^4}\int d^4x\, e^{-iq\cdot x}V(x)\Phi(x). \qquad (7)$$

This is almost identical to the relation between the scattering amplitude and the Schrödinger wavefunction in nonrelativistic theory.

The Green's function $G(x, x')$ can be written in several forms, but the simplest one for taking the asymptotic and zero-energy limits is

$$G(x, x') = \frac{i}{4(2\pi)^2}\int_{-1}^{1} d\alpha\, e^{E\cdot(x-x')\alpha/2} \times K_0\left[\lambda\left(m^2 - \frac{E^2}{4}(1-\alpha^2)\right)^{\frac{1}{2}}\right] \qquad (8)$$

where $\lambda = |x - x'|$. $K_0(x)$ is the zero-order modified Bessel function of the second kind, and the relative time has been taken pure imaginary. We derive this form in Appendix B. We have assumed in deriving (8) that E is in a range in which we may perform Wick's rotation of the contour[10] and use a Euclidean metric; however, in this representation E can take all values. In the cases under consideration here, the potential is a function only of $|x| = r$ in the Euclidean metric. If $\Phi(x)$ is expanded in four-dimensional spherical harmonics[11] of the angles $\Omega \equiv (\beta, \theta, \varphi)$, the following equation is obtained:

$$\sum_{n,l,m}\left\{\left[\frac{\partial^2}{\partial r^2} + \frac{3}{r}\frac{\partial}{\partial r} - m^2 + \left(\frac{E}{2}\right)^2 - \frac{n(n+2)}{r^2}\right] - E^2\left(\cos\beta\frac{\partial}{\partial r} - \frac{\sin\beta}{r}\frac{\partial}{\partial\beta}\right)^2\right\}[\Phi_{nlm}(r) - \Phi_{0nlm}(r)]y_{ln}^m(\Omega) = V(r)\sum_{nlm}\Phi_{nlm}(r)y_{ln}^m(\Omega). \qquad (9)$$

III. PROPERTIES OF THE BETHE–SALPETER GREEN'S FUNCTION

The integral in (8) is very difficult to do exactly; however, as demonstrated in Appendix B, it can be evaluated in the limit as λ becomes infinite. Before doing that we look at another form of $G(x, x')$ which displays its analyticity properties as a function of $s = E^2$. From (A5) we see that

$$G(x, x') = \frac{i}{(2\pi)^2}\frac{1}{w}\int_{4m^2}^{\infty}\frac{ds'}{s' - s} \times \exp\left[-i\tfrac{1}{2}(s')^{\frac{1}{2}}t\right]\sin\left[\frac{(s'-4m^2)^{\frac{1}{2}}}{2}w\right] \times \left[\frac{1}{(s')^{\frac{1}{2}}}\cos\left(\tfrac{1}{2}t(s)^{\frac{1}{2}}\right) + i\frac{1}{(s)^{\frac{1}{2}}}\sin\left(\tfrac{1}{2}t(s)^{\frac{1}{2}}\right)\right], \qquad (10)$$

where $w = |\mathbf{x} - \mathbf{x}'|$ and $t = x_0 - x_0'$. We now see that $G(x, x')$ is an analytic function of s with a cut extending from $4m^2$ to infinity; as explained in Appendix A, the apparent cut from minus infinity

[8] For explicit forms of $\Delta_F(x)$ see N. N. Bogoliubov and D. V. Shirkov, *Introduction to Quantized Fields* (Interscience Publishers, Inc., New York, 1959), p. 147. For spacelike distances it is given by $\mu K_1(\mu r)/4\pi^2 r$.

[9] S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory*, (Row, Peterson, and Company, New York, 1961), p. 716.

[10] G. C. Wick, Phys. Rev. 94, 1124 (1954).

[11] M. Gourdin, Nuovo Cimento 8, 338 (1958).

to zero does not exist. Thus we see, as mentioned above, $G(x, x')$ can be continued out of the region $s < 4m^2$ to the physical region by supplying a small positive imaginary part to E^2 or equivalently a negative imaginary part to m^2. This justifies the procedures we use in Appendix B. Furthermore, since $4m^2$ is a branch point of $G(x, x')$, we can expect its behavior to be pathological there. Indeed, as described below, the effect of this branch point appears in a consideration of the asymptotic behavior of $G(x, x')$.

Next, we evaluate the asymptotic form of $G(x, x')$ starting from (8). In view of the above mentioned singularity of the Green's function we evaluate its limiting form for four separate cases. The mathematical details are worked out in Appendix B. For $E > 2m$, we have from (B29a)

$$\lim_{\lambda\to\infty} G(x, x') = \frac{i}{4\pi}\frac{e^{ip\lambda}}{E\lambda} + \frac{i}{4\pi^2}\frac{\cosh[(P/E)E\cdot(x-x')]}{pE\lambda^2} \tag{11}$$

where $p^2 + m^2 = \frac{1}{4}E^2$. If the potential $V(x')$ falls off sufficiently rapidly as $r' = |x'|$ becomes large, (11) leads to the limit

$$\lim_{r\to\infty} G(x, x')V(x') = \left[\frac{i}{4\pi}\frac{e^{ip(r-x\cdot x'/r)}}{Er} + \frac{ie^{(p/E)E\cdot(x-x')}}{pEr^2}\right]V(x'). \tag{12}$$

We have assumed that $E\cdot x$ is positive. Substitution of this relation into (5) shows us that

$$\Phi(x) \sim e^{ip\cdot x} + \frac{i}{4\pi}\frac{e^{ipr}}{Er}\int d^4x' e^{-ip(x\cdot x'/r)}V(x')\Phi(x') + \frac{i}{8\pi^2}\frac{e^{pE\cdot x/E}}{pEr^2}\int d^4x' e^{-pE\cdot x'/E}V(x')\Phi(x'). \tag{13}$$

This equation may be rewritten using a simpler notation.

$$\Phi(x) \sim e^{ip\cdot x} + \frac{e^{ipr}}{Er}f(\Omega) + \frac{e^{pr\cos\beta}}{pEr^2}g, \tag{14}$$

where $f(\Omega)$ and g may be identified by comparisons with (13). ($\cos\beta$ is defined by $x_4 = r\cos\beta$ in our four-dimensional Euclidean space).

The term proportional to g will dominate for $\cos\beta \neq 0$. If $\cos\beta$ is negative, we just replace it by $|\cos\beta|$ and use a slightly different g. If, however, $\cos\beta$ is zero, the term proportional to $f(\Omega)$ will dominate. This is just the situation we have if Ω is taken to define the direction of the outgoing relative momentum in the scattering process. In this case[12]

$$\Phi(r, \Omega_q) = e^{ipr\hat{p}\cdot\hat{q}} + \frac{e^{ipr}}{Er}f(\Omega_q)$$

and

$$f(\Omega_q) = \frac{i}{4\pi}\int d^4x' e^{-iq\cdot x'}V(x')\Phi(x') \tag{15}$$

since $|q| = |p|$. Comparison of (15) with (7) now gives us another method of extracting the T matrix from $\Phi(x)$,

$$T(p, q, E) = -(i/4\pi^3)f(\Omega_q). \tag{16}$$

The Bethe–Salpeter amplitude in configuration space, if taken at zero time (or any finite time), has the elastic scattering amplitude as the coefficient of an outgoing spherical wave. This behavior, as is well known, holds also for the scattering solutions of the Schrödinger equation. Nonrelativistically the T matrix, up to constant factors, is the coefficient of e^{ipr}/r in the asymptotic wavefunction. This is just the nonrelativistic limit of what we have found for the Bethe–Salpeter amplitude and encourages a wavefunction interpretation of $\Phi(x)$.

If we look next at what happens when $E = 2m$, we see from (B24b) that

$$\lim_{\lambda\to\infty} G(x, x') = \frac{i}{2(2\pi)^2}\frac{1}{m\lambda\sin\gamma}$$

where $\cos\gamma$ is defined by

$$\cos\gamma = \frac{E\cdot(x-x')}{E\lambda} = \frac{r\cos\beta - r'\cos\beta'}{\lambda}.$$

This limit implies in turn that the asymptotic form of $\Phi(x)$ is given by

$$\Phi(x) \sim 1 + \frac{i}{2(2\pi)^2 mr}\int d^4x' \frac{1}{\sin\gamma}V(x')\Phi(x') = 1 + \frac{2}{\pi}\frac{1}{mr\sin\beta}f. \tag{17}$$

The last step in (17) follows from the fact that in the limit of r becoming infinite $\cos\gamma$ become identical to $\cos\beta$. In this zero-momentum case we have no exponential behavior; in fact, we have a scattering length form of the Bethe–Salpeter wavefunction. At zero time ($\beta = \frac{1}{2}\pi$) f is the zero-momentum limit of (15) and gives the corresponding T matrix by (16). Again the Bethe–Salpeter amplitude exhibits a behavior extremely similar to that of the Schrödinger wavefunction.

[12] This result for the Bethe–Salpeter amplitude may be well known, [for example, S. Okubo and D. Feldman, Phys. Rev. 117, 279, 292 (1963)]; however, the main purpose of the section is to obtain (22), which may be a new result.

If we now go to the region where $E < 2m$, we find an entirely different asymptotic formula. From (B29c) $G(x, x')$ is seen to have the form

$$\lim_{\lambda\to\infty} G(x, x') = \frac{2i}{(2\pi)^2}\left(\frac{\pi}{2m\lambda}\right)^{\frac{1}{2}}\frac{m}{E\lambda}e^{-m\lambda} \times\left\{\frac{2m\cos\gamma\sinh(\frac{1}{2}E\lambda\cos\gamma)+E\cosh(\frac{1}{2}E\lambda\cos\gamma)}{(2m^2\cos^2\gamma-E^2)}\right\} \tag{18}$$

where $\cos\gamma$ is as defined previously. The asymptotic form of $\Phi(x)$, indicated by the above limiting form of the Green's function, is complicated and not very illuminating. When $\cos\beta$ is zero, the situation is simpler, and we have

$$\Phi(x)\sim e^{ip\cdot x}+\frac{2i}{(2\pi)^2}\left(\frac{\pi m}{2E^2}\right)^{\frac{1}{2}}\frac{e^{-mr}}{r^{\frac{3}{2}}}f'(\Omega), \quad E\neq 0. \tag{19}$$

Here, p is a pure imaginary vector of length $i(m^2-\frac{1}{4}E^2)^{\frac{1}{2}}$. We see that for zero time (or finite time, $\Phi - e^{ip\cdot x}$ behaves like $e^{-mr}/r^{\frac{3}{2}}$; however, its coefficient, $f'(\Omega)$, is not simply related to the T matrix.

Until now we have worked at finite energies and momenta. As can be seen from (8) and (9), the presence of finite-energy terms removes four-dimensional spherical symmetry and enormously complicates the problem. Therefore, we go to the case of zero energy and hope no essential physics is lost. In this limit the integral in (8) is trivial, and the integral equation for $\Phi_0(x)$ is

$$\Phi_0(x)=e^{ip\cdot x}+\frac{i}{2(2\pi)^2}\int d^4x'\,K_0(m\lambda)V(x')\Phi_0(x') \tag{20}$$

where $p = im(0, 0, 1, 0)$ and defines the z axis. Using the known asymptotic form of $K_0(y)$,[13] we see that

$$\lim_{r\to\infty} G_0(x, x')V(x') = \frac{i}{2(2\pi)^2}\left(\frac{\pi}{2mr}\right)^{\frac{1}{2}}e^{-mr}e^{mz\cdot z'/r}V(x') \tag{21}$$

where $G_0(x, x')$ is the zero-energy Green's function. The asymptotic form of $\Phi_0(x)$ is

$$\Phi_0(x)\sim e^{ip\cdot x}+[e^{-mr}/(mr)^{\frac{1}{2}}]f_0(\Omega), \tag{22}$$

where

$$f_0(\Omega)=[i/2(2\pi)^2](\tfrac{1}{2}\pi)^{\frac{1}{2}} \times\int d^4x'\,e^{mz\cdot z'/r}V(x')\Phi_0(x'). \tag{23}$$

When Ω is taken in the direction of q, which like p has length im, $f(\Omega)$ is the T matrix up to constant factors. More explicitly from (7) we see that

$$T(p, q, E=0)=\left(\frac{-i}{\pi^2}\right)(1/2\pi)^{\frac{1}{2}}f_0(\Omega). \tag{24}$$

Thus we see that in some respects the zero-energy case resembles the physical scattering problem with $E \geq 2m$ more than it does the $0 < E < 2m$ case. The zero-energy behavior is a continuation of the finite-energy behavior since (18) with $E = 0$ does give (21). This limit, however, depends on whether $\cos\gamma$ is zero or not while (21) is valid for all $\cos\gamma$. Since the method of extracting the T matrix is so similar in the zero-energy and physical situations and since problems of normalization of $\Phi(x)$ should be similar in both cases, we go on to a consideration of singular potentials with the hope that the techniques developed at zero energy will be useful for the physical problem.

IV. ZERO-ENERGY SCATTERING BY A SINGULAR POTENTIAL

In the zero-energy limit the coupled differential equations for $\Phi_{nlm}(r)$ represented by (9) decouple and can be written in the form

$$\left[\frac{d^2}{dr^2}+\frac{1}{r}\frac{d}{dr}-m^2-\frac{\nu^2}{r^2}\right]^2 r\Phi_\nu(r) \equiv (O_\nu-m^2)^2 r\Phi_\nu = V(r)r\Phi_\nu(r) \tag{25}$$

for each order $n = \nu - 1$. We have suppressed the (l, m) indices and rewritten $\Phi_n(r)$ as $\Phi_\nu(r)$. Now, as mentioned by Bastai, *et al.*,[5] $V(r)$ in several theories including $\lambda\varphi_1^2\varphi_2^2$ behaves like r^{-4} as r goes to zero. This means that the integral equation obtained by a Green's function solution of (25) is not of the Fredholm type. However, it is possible to remove this most singular term by incorporating it into a new differential operator. Specifically, if $V(r)$ can be written as

$$V(r)=a/r^4+\bar{V}(r)$$

where $\bar{V}(r)$ goes as r^{-3} at worst, there exists the identity[14]

$$(O_\nu-m^2)^2-a/r^4 = (1/r)(O_{\nu_1}-m^2)r^2(O_{\nu_2}-m^2)(1/r), \tag{26}$$

where

$$\nu_1^2+\nu_2^2=2(\nu^2+1), \qquad \nu_1^2\nu_2^2=(\nu^2-1)^2-a.$$

[13] H. B. Dwight, *Tables of Integrals and Other Mathematical Data* (The Macmillan Company, New York, 1957).

[14] This transformation corresponds to $\beta = 2$ in Ref. 5. According to their footnote 11, Banerjee, Kisslinger, and Levinson have also considered the case $\beta = 2$. Professor Kisslinger informs us that Levinson and Muzinich have considered much the same problem as we discuss here.

Solving these for ν_1 and ν_2, we find

$$\begin{matrix}\nu_1\\ \nu_2\end{matrix} = \bar{\nu} \mp \omega = \left[\frac{1+\nu^2+((1-\nu^2)^2-a)^{\frac{1}{2}}}{2}\right]^{\frac{1}{2}} \mp \left[\frac{1+\nu^2-((1-\nu^2)^2-a)^{\frac{1}{2}}}{2}\right]^{\frac{1}{2}}. \quad (27)$$

In order to keep our solutions real for all ν we will restrict a to lie in the range $-4 < a < 0$.[15] In the limit of vanishing coupling constant $\bar{\nu} = \nu$ and $\omega = 1$.

$\Phi_\nu(r)$ now satisfies the equation

$$(1/r^2)[O_{\nu_1} - m^2]r^2[O_{\nu_2} - m^2]\Phi_\nu(r) = \bar{V}(r)\Phi_\nu(r). \quad (28)$$

If we let $\rho = mr$ and set $\bar{V}(r)$ equal to zero, we find the four independent solutions, $I_{\nu_1}(\rho)$, $I_{\nu_2}(\rho)$, $K_{\nu_1}(\rho)$, and $K_{\nu_2}(\rho)$. I_ν and K_ν are the first and second kind of modified Bessel functions of order ν. I_ν is well behaved at the origin, but grows exponentially at infinity; K_ν goes as $\rho^{-\nu}$ at the origin and decays exponentially at infinity. These four solutions may be used to construct a Green's function for the differential operator. The result is

$$\bar{G}_\nu(r, r') = (1/4\omega\bar{\nu})[I_{\nu_1}(\rho_<)K_{\nu_1}(\rho_>) - I_{\nu_2}(\rho_<)K_{\nu_2}(\rho_>)]. \quad (29)$$

$\Phi_\nu(r)$ now satisfies the new integral equation

$$\Phi_\nu(r) = \Phi'_{0\nu}(r) + \int_0^\infty r'^3\bar{G}_\nu(r, r')\bar{V}(r')\Phi_\nu(r')\,dr'. \quad (30)$$

$\Phi'_{0\nu}$ is some linear combination of the $\bar{V} = 0$ solutions.

$\Phi'_{0\nu}$ can be determined by an appeal to (22) which shows that the scattering modifies only the coefficient of e^{-mr}. The term in e^{mr} must thereby be unaltered from that contained in the expansion of $e^{ip\cdot x}$. At zero energy it has the expansion[16]

$$e^{ip\cdot x} = \sum_{n,l,m} \frac{I_{n+1}(mr)}{mr} A_{nlm}y_{ln}^m(\Omega_x), \quad (31)$$

where

$$A_{nlm} = (2\pi)^2[y_{ln}^m(\Omega_p)]^*. \quad (32)$$

This implies that $\Phi_\nu(\rho)$ has the asymptotic form

$$\Phi_\nu(r) \sim \frac{A_\nu}{(2\pi)^{\frac{1}{2}}}\frac{e^\rho}{\rho^{\frac{1}{2}}}, \qquad A_\nu \equiv A_{nlm}. \quad (33)$$

At the origin the behavior of Φ_ν is distorted by the potential; in fact, Φ_ν must go as $\rho^{\pm\nu_1}$ or $\rho^{\pm\nu_2}$. If we insist that Φ_ν be regular at the origin, or, more stringently, that it go as ρ^{ν_1}, we have another normalizing condition on Φ_ν. Parenthetically we remark that this behavior at the origin is maintained for the physical energy problem. This in turn means that the full amplitude $\Phi(x)$ will go as r^α at the origin where α is given by the $\nu = 1$ term of $\Phi(x)$.

$$\alpha = [1 + \tfrac{1}{2}(-a)^{\frac{1}{2}}]^{\frac{1}{2}} - [1 - \tfrac{1}{2}(-a)^{\frac{1}{2}}]^{\frac{1}{2}}.$$

This behavior verifies our statement that the integral in (5) is convergent. Returning to the problem of normalizing $\Phi'_{0\nu}$, we see that choosing the behavior ρ^{ν_1} corresponds to insisting that as the coupling constant goes to zero $\Phi_\nu(r)$ must match $\Phi_{0\nu}(r)$. The most general form for $\Phi'_{0\nu}$ in (30) is

$$\Phi'_{0\nu} = A_1I_{\nu_1} + A_2I_{\nu_2} + A_3K_{\nu_1} + A_4K_{\nu_2}. \quad (34)$$

Asymptotically $\Phi'_{\nu 0}$ becomes

$$A_1\frac{e^\rho}{(2\pi\rho)^{\frac{1}{2}}}\left(1 - \frac{4\nu_1^2-1}{8\rho}\right) + A_2\frac{e^\rho}{(2\pi\rho)^{\frac{1}{2}}}\left(1 - \frac{4\nu_2^2-1}{8\rho}\right). \quad (35)$$

Using (33) and the fact that $V(\rho)$ asymptotically becomes $-a\rho^{-4}$, we find that $\int \bar{G}\bar{V}\Phi$ goes as $e^\rho\rho^{-7/2}$. This in turn means (35) will dominate $\int \bar{G}\bar{V}\Phi$; accordingly, A_1 and A_2 must be chosen to make (35) agree with (33). Thus, $A_2 = -A_1$ and $A_1 = A_\nu/2\omega\bar{\nu}$. It should be pointed out that this choice of A_1 and A_2 makes $\Phi_{0\nu}$ and $\Phi'_{0\nu}$ agree to order $\rho^{-5/2}$ as well as $\rho^{-3/2}$. A comparison of the above discussion with Appendix C shows how similar the Schrödinger wavefunction and the Bethe–Salpeter amplitude really are. The problem of choosing the correct free behavior, after the singular potential has been absorbed into the differential operator, is essentially identical in the two cases.

Next we consider the behavior at $\rho = 0$. If Φ_ν goes as ρ^{ν_1}, then $\int \bar{G}\bar{V}\Phi_\nu$ also goes as ρ^{ν_1}. In fact, if we had insisted that Φ_ν go as ρ^{ν_2}, we would have found such behavior inconsistent with the integral equation. In any case A_3 and A_4 must vanish to ensure proper behavior at the origin. Thus, Φ_ν satisfies the integral equation

$$\Phi_\nu(r) = \frac{A_\nu}{2\omega\bar{\nu}}[I_{\nu_1} - I_{\nu_2}] + \int_0^\infty r'^3\bar{G}_\nu(r, r')\bar{V}(r')\Phi_\nu(r')\,dr'. \quad (36)$$

The integrals in this equation converge for all values of the coupling constant. Moreover, the kernel is square integrable and has finite trace for all values of the coupling constant, both positive

[15] The restrictions on a are considered in more detail in Ref. 5, though their justification is no more rigorous than ours.

[16] This result is obtained from the expansion in Appendix II of Ref. 11 by letting ρ go to $e^{i\pi/2}\rho$ and using the prescription $J_n(e^{i\pi/2}Z) = e^{-in\pi}I_n(Z)$.

and negative. The proof of these statements is easy, so we content ourselves with demonstrating the existence of the trace. The other statements are proved in a similar fashion.

The trace of the kernel is given by the integral

$$\text{tr } K = \int_0^\infty r^3 \bar{G}_\nu(r, r)\bar{V}(r)\, dr \propto \int_0^\infty \rho^3 [K_{\nu_1}(\rho)I_{\nu_2}(\rho) - K_{\nu_2}(\rho)I_{\nu_2}(\rho)]\bar{V}(\rho/m)\, d\rho. \quad (37)$$

Under the assumption that the true potential $V(r)$ was exponentially damped $\bar{V}(\rho/m)$ will be dominated at infinity by the counterterm $-ar^{-4}$. Asymptotically the product $K_\nu I_\nu$ is proportional to ρ^{-1}. Hence at infinity the integrand behaves as ρ^{-5} which is sufficient to ensure convergence. At the origin we have removed the r^{-4} behavior from $V(r)$, so $\bar{V}(\rho)$ behaves at worst like ρ^{-3}. The product $K_\nu I_\nu$ becomes constant at the origin, so we see that tr K does exist. Equation (36) is thus a Fredholm equation, and all the usual theorems apply.

Equation (36) should be contrasted with the one derived from the work of Bastai, *et al.*,[5] by the same method. In that case, the resultant equation admits only bound-state solutions.

V. APPLICATIONS

An explicit solution of (36) with a physical potential runs into formidable computational difficulties. For example, if we had a $\lambda\phi^4$ theory where the exchanged mesons had the same mass as the scattered ones, $V(r)$ would be proportional to $[K_1(mr)]^2/r^2$. The integrals necessary to obtain even a first-order solution with this potential are not readily available. If, however, we consider scattering by exchange of massless mesons, the problem becomes completely soluble. From (2) the potential is given by

$$V(x) = \int \frac{d^4p}{(2\pi)^4} e^{-ip\cdot x} \int \frac{d^4k}{(2\pi)^4} \frac{(i\lambda)^2}{k^2(k-p)^2}$$
$$= \left(\frac{i\lambda}{(2\pi)^4}\right)^2 \int \frac{d^4k}{k^2} e^{ik\cdot x} \int \frac{d^4p}{p^2} e^{-ip\cdot x}$$
$$= \frac{\lambda^2}{(2\pi)^4}\frac{1}{r^2}\left[\int_0^\infty dk\, J_1(kr)\right]^2 = \frac{\lambda^2}{(2\pi)^4}\frac{1}{r^4} \equiv \frac{a}{r^4}. \quad (38)$$

Therefore, for the exchange of massless mesons, $\bar{V}(r)$ is identically zero and the complete solution for Φ_ν is given by $\Phi_{0\nu}'$. To get the T matrix, we extract the coefficient of $e^{-\rho}/\rho^{\frac{1}{2}}$ from $\Phi_{0\nu}'$ by use of the more exact form of the asymptotic behavior of $I_\nu(\rho)$[17]:

$$I_\nu(\rho) = \frac{1}{(2\pi\rho)^{\frac{1}{2}}}[e^\rho + e^{-(\nu+\frac{1}{2})\pi i}e^{-\rho}], \quad -\tfrac{3}{2}\pi < \arg\rho < \tfrac{1}{2}\pi. \quad (39)$$

The asymptotic behavior of $I_\nu(\rho)$ exhibits a Stoke's phenomenon in the sector $-\frac{1}{2}\pi < \arg\rho < \frac{1}{2}\pi$. In Appendix C we discuss the choice of the phase factor in the second term on the right-hand side of (39). The coefficient of $e^{-\rho}/\rho^{\frac{1}{2}}$ is now seen to be equal to

$$\frac{A_\nu}{2\omega\bar{\nu}} \frac{[e^{-i(\nu_1+\frac{1}{2})\pi} - e^{-i(\nu_2+\frac{1}{2})\pi}]}{(2\pi)^{\frac{1}{2}}} = \frac{A_\nu}{(2\pi)^{\frac{1}{2}}} \frac{e^{-i\bar{\nu}\pi}\sin\omega\pi}{\omega\bar{\nu}}. \quad (40)$$

Substitution of this result into (24) gives

$$T(p, q, E = 0) = \frac{-i}{2\pi^3}\sum_{n,l,m} A_{nlm}\frac{e^{-i\bar{\nu}\pi}\sin\pi\omega}{\omega\bar{\nu}} y_{ln}^m(\Omega_q)$$
$$= \frac{-2i}{\pi}\sum_{nlm}\frac{e^{-i\bar{\nu}\pi}\sin\pi\omega}{\omega\bar{\nu}}[y_{ln}^m(\Omega_p)]^* y_{ln}^m(\Omega_q)$$
$$= \frac{-i}{\pi^3}\sum_{\nu=1}^\infty \frac{\nu}{\omega\bar{\nu}} e^{-i\bar{\nu}\pi}\sin\pi\omega C_{\nu-1}^1(\hat{p}\cdot\hat{q}). \quad (41)$$

In the last step we have used the addition theorem for the four-dimensional spherical harmonics[11] to give the Gegenbauer function of $\hat{p}\cdot\hat{q} = p\cdot q/m^2$. The expansion should be symmetrized to take account of the identity of the scattered particles. This condition restricts the sum to odd values of ν. The presence of the phase factor in (40) seems surprising at first since it arises from an ostensibly real function. The nonreality exists only to the extent that we are using the asymptotic form of $I_\nu(x)$; the imaginary part is infinitesimal compared to the real part. An examination of Appendix C shows that exactly the same sort of phase factor occurs in the nonrelativistic problem. There the resultant T matrix is independent of energy; for positive energy the phase factor is expected, and our prescription for extracting the T matrix for negative energy works to provide the same phase factor, as it must. We conjecture that the same sort of behavior holds for the relativistic problem to the extent that (41) may be valid even for positive energies.

As a check on our methods (41) can be compared with the results of Baker and Muzinich.[4] They solved the problem of scattering of massless particles by exchange of massless particles. Comparison of their Eq. (III.9) with (41) shows that the results agree,

[17] Y. L. Luke, *Integrals of Bessel Functions*, (McGraw-Hill Book Company, Inc., New York, 1962), p. 32.

at least in form, in the weak coupling limit. In fact, their result for $T(p, q, E = 0)$ is in error if higher orders in the coupling constant are considered.

As a second, and perhaps more physical example, we consider scattering by the potential (setting $m = 1$, hereafter)

$$V(r) = \frac{2a}{\pi^{\frac{1}{2}}} \frac{K_{\frac{1}{2}}(2r)}{r^{7/2}} = \frac{ae^{-2r}}{r^4}. \tag{42}$$

This potential matches the behavior at the origin of the $\lambda\phi^4$ potential described at the beginning of this section. At infinity it differs only by a multiplicative factor proportional to r^{-1}. Therefore, it should be representative of the sort of behavior that might be found in a physical problem. More importantly, the integrals involved in a first-order approximation are possible.

By the application of Fredholm theory we see that the first-order solution of (36) is given by

$$\Phi_\nu(r) = \frac{A_\nu}{2\omega\bar{\nu}} \left\{ [I_{\nu_1} - I_{\nu_2}] \left[1 - \int r'^2\, dr' \tilde{G}_\nu(r', r') \bar{V}(r') \right] + \int r'^3\, dr' \tilde{G}_\nu(r, r') \bar{V}(r') [I_{\nu_1}(r') - I_{\nu_2}(r')] \right\} \times \left[1 - \int r^3\, dr \tilde{G}_\nu(r, r) \bar{V}(r) \right]^{-1}, \tag{43}$$

where

$$\bar{V}(r) = a \left[\frac{2}{\pi^{\frac{1}{2}}} \frac{K_{\frac{1}{2}}(2r)}{r^{7/2}} - \frac{1}{r^4} \right] = \frac{a}{r^4} [e^{-2r} - 1], \tag{44}$$

and $G_\nu(r, r')$ is given by (29). The denominator function may be calculated exactly. The required integral is the trace of $\tilde{G}_\nu \bar{V}$,

$$\operatorname{tr} \tilde{G}_\nu \bar{V} = \frac{a}{2\nu\bar{\omega}} \int_0^\infty \frac{dr}{r} [K_{\nu_1} I_{\nu_1} - K_{\nu_2} I_{\nu_2}] \times \left[\frac{2}{\pi^{\frac{1}{2}}} r^{\frac{1}{2}} K_{\frac{1}{2}}(2r) - 1 \right]. \tag{45}$$

Taking the difference of the formulas[18]

$$\frac{2}{\pi^{\frac{1}{2}}} \int_0^\infty \frac{dx}{x^{\frac{1}{2}}} I_\nu(x) K_\nu(x) K_{\frac{1}{2}-\epsilon}(2x) = \frac{\Gamma(\frac{1}{2} - \epsilon)\Gamma(\epsilon)\Gamma(\frac{1}{2} + \nu - \epsilon)\Gamma(\nu + \epsilon)}{2\pi^{\frac{1}{2}}\Gamma(1 + \nu - \epsilon)\Gamma(\frac{1}{2} + \nu + \epsilon)} \tag{46}$$

and[19]

$$\int_0^\infty x^{-1+\epsilon} K_\nu(x) I_\nu(x)\, dx = \frac{\Gamma(\nu + \frac{1}{2}\epsilon)\Gamma(\frac{1}{2}\epsilon)\Gamma(1 - \epsilon)}{4^{\epsilon}\Gamma(1 + \nu - \frac{1}{2}\epsilon)\Gamma(1 - \frac{1}{2}\epsilon)}, \tag{47}$$

we get the integral in (45) in the limit as ϵ goes to zero. (The divergent parts cancel nicely.)

$$\int_0^\infty \frac{dr}{r} K_\nu I_\nu \left[\frac{2}{\pi^{\frac{1}{2}}} r^{\frac{1}{2}} K_{\frac{1}{2}}(2r) - 1 \right] = \frac{[6 \ln 2 - 4\psi(\nu + \frac{1}{2}) + \psi(\nu) + \psi(\nu + 1)]}{4\nu}; \tag{48}$$

$\psi(\nu)$ is the logarithmic derivative of $\Gamma(\nu)$. Equation (48) holds for $\nu \neq 1$. For $\nu = 1$ our method breaks down, but we show in Appendix D that (48) is then equal to $-\frac{1}{4}\pi^2$. The trace of the kernel for nonzero ν is thus given by

$$\operatorname{tr} \tilde{G}_\nu \bar{V} = \frac{a}{16\omega\bar{\nu}} \left\{ \frac{6 \ln 2 - 4\psi(\nu_1 + \frac{1}{2}) + \psi(\nu_1) + \psi(\nu_1 + 1)}{\nu_1} - (\nu_1 \leftrightarrow \nu_2) \right\}. \tag{49}$$

To lowest order in a we can replace ω by unity and $\bar{\nu}$ by ν everywhere in this expression,

$$\begin{aligned} \operatorname{tr} \tilde{G}_\nu \bar{V} &= aD(\nu) \\ &\simeq \frac{a}{16\nu} \left\{ \frac{12 \ln 2}{\nu^2 - 1} + \frac{4\psi(\nu + \frac{3}{2})}{\nu + 1} - \frac{4\psi(\nu - \frac{1}{2})}{\nu - 1} - \frac{\psi(\nu + 1) + \psi(\nu + 2)}{\nu + 1} + \frac{\psi(\nu - 1) + \psi(\nu)}{\nu - 1} \right\}, \qquad \nu > 1 \\ &\simeq \frac{a}{4} \left\{ -\frac{\pi^2}{4} - \frac{6 \ln 2 - 4\psi(\frac{3}{2}) + \psi(2) + \psi(3)}{8} \right\}, \qquad \nu = 1 \\ &\simeq -0.9a. \end{aligned} \tag{50}$$

Now we turn our attention to the numerator of (43). Since $I_{\nu_1} - I_{\nu_2}$ is itself of order a, we drop the trace term in the numerator. There are then two sorts of integrals involved, neither of which we can do. However, if we ask for the coefficient of $e^{-r}/r^{\frac{1}{2}}$, our task is much simpler. Terms of the form

$$\begin{aligned} I_\nu(r) \int_r^\infty \frac{dr'}{r'} K_\lambda(r') I_\mu(r') [e^{-2r'} - 1] &\sim \frac{1}{2(2\pi r)^{\frac{1}{2}}} [e^r + \delta_\nu e^{-r}] \\ &\times \int_r^\infty \frac{dr'}{r'^{\frac{1}{2}}} e^{-r'} [e^{+r'} + \delta_\mu e^{-r'}] [e^{-2r'} - 1] \end{aligned} \tag{51}$$

do not contribute. The coefficient δ_ν is just the phase factor from (39). The parts of (51) which are

[18] *Tables of Integral Transforms*, edited by A. Erdélyi (McGraw-Hill Book Company, Inc., New York, 1954), Vol. II, p. 372.

[19] See Ref. 17, p. 325.

exponentially decreasing fall off at least as fast as $e^{-r}/r^{\frac{1}{2}}$. Thus we are left with

$$K_{\nu_1}\int_0^r \frac{dr'}{r'} I_{\nu_1}[I_{\nu_1} - I_{\nu_2}](e^{-2r'} - 1)$$

$$- K_{\nu_2}\int_0^r \frac{dr'}{r'} I_{\nu_2}[I_{\nu_1} - I_{\nu_2}](e^{-2r'} - 1) \tag{52}$$

$$\sim \left(\frac{\pi}{2r}\right)^{\frac{1}{2}} e^{-r}\left\{\int_0^r \frac{dr'}{r'}[I_{\nu_1} - I_{\nu_2}]^2[e^{-2r'} - 1]\right\} \tag{53}$$

$$\sim \left(\frac{\pi}{2r}\right)^{\frac{1}{2}} e^{-r}\left\{\int_0^\infty \frac{dr'}{r'} e^{-2r'}(I_{\nu_1} - I_{\nu_2})^2 - \int_0^r \frac{dr'}{r'}(I_{\nu_1} - I_{\nu_2})^2\right\}. \tag{54}$$

In going from (53) to (54) we extended the upper limits of one of the integrals to infinity, since it is convergent. In the second integral we have to extract the term that behaves like a constant. The second integral can be done exactly; however, it is easier to obtain its asymptotic behavior indirectly. We use the integral[20]

$$\int_0^s J_\nu(kt)J_\mu(kt)\frac{dt}{t} = -\frac{kz}{\mu^2 - \nu^2}\{J_{\mu+1}(kz)J_\nu(kz) - J_\mu(kz)J_{\nu+1}(kz)\} + \frac{J_\mu(kz)J_\nu(kz)}{\mu + \nu} \sim \frac{+2}{\pi(\mu^2 - \nu^2)}\sin\left(\frac{\mu - \nu}{2}\pi\right). \tag{55}$$

Letting μ approach ν in this integral gives us the additional result

$$\int_0^s \frac{dt}{t}[J_\nu(kt)]^2 \sim \frac{1}{2\nu}. \tag{56}$$

If we let $k = e^{\frac{1}{2}i\pi}$, we can extract immediately the terms that will behave as a constant in the second integral of (54). The result is

$$\int_0^r \frac{dr'}{r'}(I_{\nu_1} - I_{\nu_2})^2 \sim \frac{e^{-i\pi\nu_1}}{2\nu_1} + \frac{e^{-i\pi\nu_2}}{2\nu_2} - \frac{4e^{-i\frac{1}{2}[(\nu_1+\nu_2)\pi]}\sin\frac{1}{2}[(\nu_1 - \nu_2)\pi]}{\pi(\nu_1^2 - \nu_2^2)} = \frac{e^{-i\pi\nu_1}}{2\nu_1} + \frac{e^{-i\pi\nu_2}}{2\nu_2} - \frac{e^{-i\bar{\nu}\pi}\sin\omega\pi}{\pi\omega\bar{\nu}}. \tag{57}$$

Since this expression is to be multiplied by a, we set a equal to zero everywhere within it. The result is

$$\int_0^r \frac{dr'}{r'}[I_{\nu_1} - I_{\nu_2}]^2 \sim \frac{-\nu e^{-i\nu\pi}}{\nu^2 - 1}, \qquad \nu > 1. \tag{58}$$

The other integral in (54) can be done exactly by use of the formula[21]

$$\int_0^\infty x^{-\frac{1}{2}}I_\mu(x)I_\lambda(x)K_{\frac{1}{2}}(2x)\,dx = \frac{1}{2^{\mu+\lambda+2}}\frac{\Gamma\frac{1}{2}(\mu + \lambda + 1)\Gamma\frac{1}{2}(\mu + \lambda)}{\Gamma(\mu + 1)\Gamma(\lambda + 1)} \times {}_4F_3\left(\frac{\lambda + \mu + 1}{2}, \frac{\mu + \lambda}{2}, \frac{\mu + \lambda + 1}{2}, \frac{\mu + \lambda}{2} + 1; \mu + 1, \lambda + 1, \mu + \lambda + 1; 1\right), \tag{59}$$

where ${}_pF_q$ represents a generalized hypergeometric function. To zero order in a, the above integral leads to

$$\int_0^\infty \frac{dr}{r} e^{-2r}[I_{\nu_1} - I_{\nu_2}]^2 = \frac{2}{\pi^{\frac{1}{2}}}\left\{\frac{\Gamma(\nu - \frac{1}{2})\,{}_3F_2(\nu - \frac{1}{2}, \nu - \frac{1}{2}, \nu - 1; \nu, 2\nu - 1; 1)}{2^{2\nu}(\nu - 1)\Gamma(\nu)} + \frac{\Gamma(\nu + \frac{3}{2})\,{}_3F_2(\nu + \frac{3}{2}, \nu + \frac{3}{2}, \nu + 1; \nu + 2, 2\nu + 3; 1)}{2^{2\nu+4}(\nu + 1)\Gamma(\nu + 2)} - \frac{\Gamma(\nu + \frac{1}{2})\,{}_3F_2(\nu + \frac{1}{2}, \nu + \frac{1}{2}, \nu + 1; \nu + 2, 2\nu + 1; 1)}{2^{2\nu+1}\Gamma(\nu + 2)}\right\} \equiv N(\nu), \qquad \nu > 1. \tag{60}$$

Unfortunately we are unable to reduce the ${}_3F_2$ functions any further. For $\nu = 1$ we can use (59), only we must be more careful. The term involving $I_{\nu_1}^2$ will dominate. For $\nu = 1$, $\nu_1 = \frac{1}{2}(-a)^{\frac{1}{2}}$, and (59) gives

$$\int_0^\infty \frac{d\rho}{\rho} e^{-2\rho}[I_{\frac{1}{2}(-a)^{\frac{1}{2}}}(\rho)]^2 = \frac{1}{2}\frac{\pi^{\frac{1}{2}}}{(-a)^{\frac{1}{2}}} \equiv N(1). \tag{61}$$

We can now combine (50), (58), and (60) together with (40) to give the lowest-order coefficient of $e^{-\rho}/\rho^{\frac{1}{2}}$ in Φ_ν,

$$T_\nu(p, q, E = 0) = \frac{-iA_\nu}{\pi^2 2\nu} \times \frac{-\frac{a}{8(\nu^2 - 1)}e^{-i\nu\pi} + \frac{a\nu}{8\nu(\nu^2 - 1)}e^{-i\nu\pi} + \frac{aN(\nu)}{8\nu}}{(1 - aD(\nu))} = \frac{-aiA_\nu N(\nu)}{\pi^2 16\nu^2}, \qquad \nu > 1. \tag{62}$$

[20] See Ref. 17, p. 255.

[21] See Ref. 18, p. 140.

We have approximated $\sin \omega\pi$ by

$$\sin \omega\pi \simeq -a[\pi/8(\nu^2 - 1)].$$

Thus we see that the contribution from the counterterm cancels the $\bar{V} = 0$ term, as it must. If $\nu = 1$, we use instead of (62)

$$T_1(p, q, E = 0) = \frac{i(-a\pi)^{\frac{1}{2}} A_1}{2(2\pi)^2[1 - aD(1)]}. \quad (63)$$

The $\nu = 1$ term is then of order $(-a)^{\frac{1}{2}}$, while all the other terms are of order a. This fact makes our potential seem all the more physical, for in a true $\lambda\phi^4$ theory there would be a contact interaction term of order $(-a)^{\frac{1}{2}}$ which would contribute only to $\nu = 1$. Presumably, it would be introduced to cancel out the $(-a)^{\frac{1}{2}}$ behavior found here and leave a term proportional to a. More directly, the $(-a)^{\frac{1}{2}}$ dependence comes from the difference in behavior of (27) between ν identically one and any other integer.

It is interesting to ask where the denominator of (62) vanishes to give a bound state. As ν becomes infinite, $D(\nu)$ vanishes as $\nu^{-3} \ln \nu$, so there are no bound states in this limit for small a. At the other extreme there is a pole for $\nu = 1$ but not for any larger value due to our restriction $-4 < a < 0$. Numerically from (50) we see that the pole occurs at $a = -1.1$.

In this section we have treated two problems. One is exactly soluble but shows no possibility of bound states. The other model, using a more realistic potential, illustrates many of the phenomena expected in a physical theory. It also provides many of the problems encountered in handling a physical situation. The practical utility of extracting the T matrix as the coefficient of asymptotic behavior is clearly demonstrated.

In conclusion we remark that the singular nature of the potential in a Bethe–Salpeter equation is not an obstacle to solution if the singularity is of the form r^{-4}. We have shown how to remove the singularity for the scattering problem at zero energy. This same method will work for physical energies if the wavefunction is again expanded in four-dimensional spherical harmonics. The major stumbling block to effective utlization of the Bethe–Salpeter equation in physical problems remains the lack of spherical symmetry due to energy factors. We have discussed the two-particle, free Green's function for physical energies; hopefully, a knowledge of the asymptotic properties of the configuration-space amplitude and its relation to the T matrix will provide a starting point for an attack on physical problems.

ACKNOWLEDGMENTS

One of us (B.W.L.) wishes to thank S. Fubini and R. F. Sawyer for interesting conversations.

APPENDIX A. SPECTRAL REPRESENTATION OF THE GREEN'S FUNCTION

Starting from (4), we write

$$G(x, x') = \int \frac{d^4k}{(2\pi)^4} \times \frac{e^{ik\cdot(x-x')}}{[(k + \frac{1}{2}E)^2 - m^2 + i\epsilon][(k - \frac{1}{2}E)^2 - m^2 + i\epsilon]}. \quad (A1)$$

The integrand, considered as a function of k_0, has poles at

$$k_0 \pm \tfrac{1}{2}E = \pm[(k^2 + m^2)^{\frac{1}{2}} - i\epsilon]. \quad (A2)$$

For $x_0 - x_0' > 0$ the contour of integration may be closed in the upper half k_0 plane with the result

$$G(x, x') = \frac{-i}{(2\pi)^2}\frac{1}{Er} \times \int_0^\infty \frac{k\, dk \exp\{-i[\mathbf{k}\cdot\mathbf{z} - t(k^2 + m^2)^{\frac{1}{2}}]\}}{2E[(k^2 + m^2)]^{\frac{1}{2}}} \times \left\{\frac{e^{-i\frac{1}{2}Et}}{E + 2(k^2 + m^2)^{\frac{1}{2}}} + \frac{e^{i\frac{1}{2}Et}}{E - 2(k^2 + m^2)^{\frac{1}{2}}}\right\}, \quad (A3)$$

where $\mathbf{z} = \mathbf{x} - \mathbf{x}'$ and $t = x_0 - x_0'$. Next the angular integrations are done to give

$$G(x, x') = \frac{-i}{(2\pi)^2}\frac{1}{Er} \times \int_0^\infty \frac{k\, dk \sin kr}{(k^2 + m^2)^{\frac{1}{2}}} \frac{\exp[-it(k^2 + m^2)^{\frac{1}{2}}]}{E^2 - 4(k^2 + m^2)} \times \{2E\cos(\tfrac{1}{2}Et) + 4i(k^2 + m^2)^{\frac{1}{2}}\sin(\tfrac{1}{2}Et)\}. \quad (A4)$$

Here $|\mathbf{z}|$ has been written as r. We now change variables to $s = E^2$ and $s' = 4(k^2 + m^2)$ to get a form for $G(x, x')$ that better displays its analyticity properties,

$$G(x, x') = \frac{i}{(2\pi)^2}\frac{1}{r}\int_{4m^2}^\infty \frac{ds'}{s' - s}\sin\left[\frac{(s' - 4m^2)^{\frac{1}{2}}}{2}r\right] \times \exp[-i\tfrac{1}{2}(s'^{\frac{1}{2}}t)]\left[\frac{\cos(\frac{1}{2}s^{\frac{1}{2}}t)}{s'^{\frac{1}{2}}} + i\frac{\sin(\frac{1}{2}ts^{\frac{1}{2}})}{s^{\frac{1}{2}}}\right]. \quad (A5)$$

We see from this equation that $G(x, x')$ has a cut extending from $4m^2$ to infinity. There is also an apparent cut, due to square roots, that extends from minus infinity to zero. However, the cosine is an even function so that the square root in its argument does not lead to a cut. The presence of

$s^{-\frac{1}{2}}$ cancels the cut coming from the argument of the sine function.

APPENDIX B. CALCULATION OF THE ASYMPTOTIC PROPERTIES OF $G(x, x')$

We start by deriving (8) from (4). For simplicity, we redefine E by a factor of two and let $z = x - x'$ and $\lambda = |z|$. If the denominators in the integrand of (4) are combined by Feynman parametrization, we have

$$G(x, x') = \frac{1}{2(2\pi)^4} \int_{-1}^{1} d\alpha \int d^4k \times \frac{e^{ik\cdot z}}{[k^2 + 2k\cdot E\alpha + E^2 - m^2]^2}. \quad \text{(B1)}$$

If the origin of integration is shifted in the standard way to remove the $k\cdot E$ term in the denominator, the k_0 contour may be rotated to the imaginary axis provided $E^2 < m^2$. At the same time we rotate z_0 to give a Euclidean metric in configuration space. The result of these operations is

$$G(x, x') = \frac{i}{2(2\pi)^4} \int_{-1}^{1} d\alpha e^{E\cdot z\alpha} \times \int \frac{d^4k e^{ik\cdot z}}{[k^2 - E^2(1 - \alpha^2) + m^2]^2}. \quad \text{(B2)}$$

If $e^{ik\cdot z}$ is expanded in four-dimensional spherical harmonics,[11] only the first term will contribute, and it is proportional to $J_1(k\lambda)$,

$$G(x, x') = \frac{i}{2(2\pi)^2} \int_{-1}^{1} d\alpha e^{E\cdot z\alpha} \frac{1}{\lambda} \times \int_0^\infty \frac{k^2\, dk J_1(k\lambda)}{[k^2 + E^2\alpha^2 - E^2 + m^2]^2}. \quad \text{(B3)}$$

The final k integration may now be done by referring to any of a number of references on Bessel functions.[22] The result is

$$G(x, x') = [i/4(2\pi)^2]I, \quad \text{(B4)}$$

where

$$I = \int_{-1}^{1} d\alpha e^{E\cdot z\alpha} K_0[\lambda(m^2 - E^2(1 - \alpha^2))^{\frac{1}{2}}]. \quad \text{(B5)}$$

Equation (B4) is the form given in (8).

We now consider the integral I in detail. There are three distinct values of E for which this integral must be evaluated. For E greater than m, the argument of K_0 can become imaginary. For E less than m the argument is always real and positive; the asymptotic form of K_0 can then be used under the integral. Finally there is the special case of $E = m$ which must be evaluated separately.

[22] See Ref. 17, p. 330.

When the argument of K_0 is imaginary, we use a small positive imaginary part on E^2 to determine the correct continuation of K_0 to the Hankel function of the first kind. We write

$$K_0[\lambda(E^2\alpha^2 - p^2)]^{\frac{1}{2}} = K_0[-i\lambda(p^2 - E^2\alpha^2)^{\frac{1}{2}}] = \tfrac{1}{2}\pi i H_0^{(1)}[\lambda(p^2 - E^2\alpha^2)^{\frac{1}{2}}], \quad \text{(B6)}$$

where $p^2 = E^2 - m^2 > 0$ and $\alpha < p/E < 1$. The integral I can now be split into two parts,

$$I_{E>m} = I_1 + I_2 = \int_{p/E}^{1} d\alpha[e^{E\cdot z\alpha} + e^{-E\cdot z\alpha}]K_0[\lambda(E^2\alpha^2 - p^2)^{\frac{1}{2}}] + \frac{\pi i}{2} \int_{-p/E}^{p/E} d\alpha e^{E\cdot z\alpha} H_0^{(1)}[\lambda(p^2 - E^2\alpha^2)^{\frac{1}{2}}]. \quad \text{(B7)}$$

The asymptotic form of I_1 is obtained by making the change of variable from α to x where $\alpha = x/E\lambda + p/E$. Then I_1 becomes

$$I_1 = \frac{e^{E\cdot z(p/E)}}{E\lambda} \int_0^{(E-p)\lambda} dx e^{(E\cdot z/E\lambda)x} K_0[(x(x + 2p\lambda))^{\frac{1}{2}}] + \frac{e^{-E\cdot z(p/E)}}{E\lambda} \int_0^{(E-p)\lambda} dx e^{-(E\cdot z/E\lambda)x} K_0[(x(x + 2p\lambda))^{\frac{1}{2}}] = \frac{e^{E\cdot z(p/E)}}{E\lambda} I_a + \frac{e^{-E\cdot z(p/E)}}{E\lambda} I_b. \quad \text{(B8)}$$

I_b can again be divided into two parts,

$$I_b = \left(\int_0^\infty - \int_{(E-p)\lambda}^\infty\right) dx e^{-(E\cdot z/E\lambda)x} K_0[(x(x + 2p\lambda))^{\frac{1}{2}}].$$

The first part can be evaluated asymptotically by use of the formula[23]

$$I_k(\alpha, \beta) = \int_0^\infty x^{-k-\frac{1}{2}}(\alpha + x)^{k-\frac{1}{2}} e^{-\beta x} K_0[(x(x + \alpha))^{\frac{1}{2}}]\, dx = (1/\alpha)e^{\frac{1}{2}\alpha\beta}[\Gamma(\tfrac{1}{2} - k)]^2 W_{k,0}(z_1)W_{k,0}(z_2)$$

$$|\arg\alpha| < \pi, \quad \mathrm{Re}\,\beta > -1, \quad \mathrm{Re}\,k < \tfrac{1}{2},$$

$$z_2^1 = \tfrac{1}{2}\alpha[\beta \pm (\beta^2 - 1)^{\frac{1}{2}}]^{\frac{1}{2}}. \quad \text{(B9)}$$

For the first part of I_b we want

$$-\lim_{k\to\frac{1}{2}} \lim_{\alpha\to\infty} \frac{\partial}{\partial\beta} I_{k,0}(\alpha, \beta) = \frac{2}{\alpha}. \quad \text{(B10)}$$

This limit is easily found by using the simple asymptotic form of the Whitaker functions $W_{k,0}$.[24] The first part of I_b is then given by $2/2p\lambda$. The second term can be evaluated by using the asymp-

[23] See Ref. 18, p. 377.

[24] W. Magnus and F. Oberhettinger, *Formulas and Theorems for the Special Functions of Mathematical Physics* (Chelsea Publishing Company, New York, 1949), p. 89.

totic form of K_0 in the integrand and using the general formula

$$\int_a^b \frac{e^{g(x)}}{f(x)}\,dx = \left\{\frac{1}{f(x)g'(x)} + \frac{f'(x)g'(x) + f(x)g''(x)}{f^2(x)g'^3(x)} + \cdots\right\}e^{g(x)}\Big|_a^b, \quad \text{(B11)}$$

where we have simply integrated by parts. It is assumed that $f(x)$ has no zeros in the range of integration. This formula shows us immediately that the second part of I_b vanishes exponentially relative to the first and, hence, can be neglected.

In the limit in which we are working $I_b = I_a$. Therefore, we find the asymptotic form of I_1 to be

$$I_1 \sim (2/pE\lambda^2)\cosh(E\cdot zp/E). \quad \text{(B12)}$$

This form will be valid for all $E\cdot z = E\lambda\cos\beta$. Turning now to I_2, we set $\alpha = p\cos\theta/E$ and look at the integral

$$I_2 = \frac{\pi i}{2}\frac{p}{E}\int_0^\pi \sin\theta\, d\theta e^{E\cdot z(p/E)\cos\theta}H_0^{(1)}(p\lambda\sin\theta)$$
$$= \frac{\pi i p}{E}\int_0^\pi \sin\theta\, d\theta \times \cosh[E\cdot z(p/E)\cos\theta]H_0^{(1)}(p\lambda\sin\theta). \quad \text{(B13)}$$

We use the following identities[13]:

$$\cosh(y\cos\theta) = I_0(y) + 2\sum_{n=1} I_{2n}(y)\cos 2n\theta,$$

$$\sin\theta\cos 2n\theta = \tfrac{1}{2}[\sin(2n+1)\theta - \sin(2n-1)\theta], \quad \text{(B14)}$$

$$H_0^{(1)}(y) = J_0(y) + iY_0(y),$$

and the definite integrals[25]

$$\int_0^\pi J_0(z\sin t)\sin(2n+1)t\,dt = \pi(-1)^n J_{-\frac{1}{2}(2n+1)}\left(\frac{z}{2}\right)J_{\frac{1}{2}(2n+1)}\left(\frac{z}{2}\right), \quad \text{(B15)}$$

$$\int_0^\pi Y_0(z\sin t)\sin(2n+1)t\,dt = (-1)^n\left[J_{\frac{1}{2}(2n+1)}\left(\frac{z}{2}\right)\frac{\partial}{\partial\nu}J_\nu\left(\frac{z}{2}\right)\Big|_{\nu=-\frac{1}{2}(2n+1)} + J_{-\frac{1}{2}(2n+1)}\left(\frac{z}{2}\right)\frac{\partial}{\partial\nu}J_\nu\left(\frac{z}{2}\right)\Big|_{\nu=-\frac{1}{2}(2n+1)}\right] \quad \text{(B16)}$$

to evaluate I_2. If the asymptotic form of $J_n(z)$ is given by [13]

$$J_n(z) \sim (2/\pi z)^{\frac{1}{2}}\cos(z - \tfrac{1}{4}\pi - \tfrac{1}{2}n\pi), \quad \text{(B17)}$$

the limiting form of (B15) is given by

$$\int_0^\pi J_0(z\sin t)\sin(2n+1)t\,dt \sim (-1)^n\frac{2\sin z}{z}, \quad \text{(B18)}$$

and that of (B16) by

$$\int_0^\pi Y_0(z\sin t)\sin(2n+1)t\,dt \sim (-1)^n\frac{2\cos z}{z}. \quad \text{(B19)}$$

Using (B14), (B18), and (B19) in (B13), we now have

$$I_2 \sim \frac{2\pi p}{E}\left\{I_0\left(\frac{Ezp}{E}\right)\frac{e^{ip\lambda}}{p\lambda} + \sum_{n=1} I_{2n}\left(E\cdot z\left(\frac{p}{E}\right)\right)\frac{e^{ip\lambda}}{p\lambda}[(-1)^n - (-1)^{n-1}]\right.$$
$$= \frac{2\pi e^{ip\lambda}}{E\lambda}\left\{I_0\left(\frac{E\cdot zp}{E}\right) + 2\sum_{n=1}(-1)^n I_{2n}\left(\frac{E\cdot zp}{E}\right)\right\}. \quad \text{(B20)}$$

From (B14) we see that the series in (B20) is equal to unity. Therefore, we have the simple result

$$I_2 \sim 2\pi(e^{ip\lambda}/E\lambda). \quad \text{(B21)}$$

Adding I_1 and I_2, we have the final result

$$I_{E>m} = 2\pi\frac{e^{ip\lambda}}{E\lambda} + \frac{2}{pE\lambda^2}\cosh\left(\frac{E\cdot zp}{E}\right). \quad \text{(B22)}$$

Next we evaluate the integral I for $E = m$; this problem is considerably simpler than that above,

$$I_{E=m} = \int_{-1}^1 d\alpha\, e^{\alpha m\lambda\cos\beta}K_0(m\lambda\alpha)$$
$$= \int_0^\infty d\alpha\,[e^{\alpha m\lambda\cos\beta} + e^{-\alpha m\lambda\cos\beta}]K_0(m\lambda\alpha) \quad \text{(B23)}$$
$$- \int_1^\infty d\alpha 2\cosh(\alpha m\lambda\cos\beta)K_0(m\lambda\alpha)$$
$$\equiv I_1 - I_2. \quad \text{(B24)}$$

I_1 may be done exactly,[26]

$$I_1 = \frac{2i}{m\lambda(\cos^2\beta - 1)^{\frac{1}{2}}} = \frac{2}{m\lambda\sin\beta}. \quad \text{(B25)}$$

The integral I_2 can be evaluated by use of the asymptotic form of K_0 and (B11); it is seen to vanish exponentially for $\cos\beta < 1$, and as $\lambda^{-\frac{1}{2}}$ for $|\cos\beta| = 1$. Therefore, we have the result

$$I_{E=m} = 2/m\lambda\sin\beta. \quad \text{(B26)}$$

For $E < m$ the argument of K_0 never vanishes; accordingly, the asymptotic form may be used in the integral and we have[13]

[25] See Ref. 17, p. 294.

[26] See Ref. 18, Vol. I, p. 197.

$$I_{E<m} \sim \left(\frac{\pi}{2\lambda}\right)^{\frac{1}{2}} \int_{-1}^{1} \frac{d\alpha}{[m^2 - E^2(1-\alpha^2)]^{\frac{1}{4}}} \times e^{E\cdot z\alpha - \lambda(m^2 - E^2(1-\alpha^2))^{\frac{1}{2}}}. \quad \text{(B27)}$$

This integral can be evaluated by use of (B11), which provides a series of descending powers in λ,

$$I_{E<m} \sim \left(\frac{\pi}{2\lambda}\right)^{\frac{1}{2}} e^{-m\lambda}\left\{\frac{me^{E\cdot z}}{m^{\frac{1}{2}}E\lambda(m\cos\beta - E)} - \frac{me^{-E\cdot z}}{m^{\frac{1}{2}}E\lambda(m\cos\beta + E)}\right\}$$

$$= \left(\frac{\pi}{2m\lambda}\right)^{\frac{1}{2}} \frac{m}{E\lambda} e^{-m\lambda} \times \left\{\frac{m\cos\beta\sinh(E\lambda\cos\beta) + E\cosh(E\lambda\cos\beta)}{m^2\cos^2\beta - E^2}\right\}. \quad \text{(B28)}$$

This value of I, like the earlier ones, is valid for all values of $|\cos\beta| \leq 1$. Equation (B28) can be continued to $E = 0$ where it gives the proper behavior as read directly from (B5).

We summarize the results derived in this appendix:

$$I_{E>m} \sim \frac{2\pi e^{ip\lambda}}{E\lambda} + \frac{2}{pEr^2}\cosh(p\lambda\cos\beta), \quad \text{(B29a)}$$

$$I_{E=m} \sim 2/m\lambda\sin\beta, \quad \text{(B29b)}$$

$$I_{0<E<m} \sim \left(\frac{\pi}{2m\lambda}\right)^{\frac{1}{2}} e^{-m\lambda}\frac{m}{E\lambda} \times \frac{m\cos\beta\sinh(E\lambda\cos\beta) + E\cosh(E\lambda\cos\beta)}{m^2\cos^2\beta - E^2}, \quad \text{(B29c)}$$

$$I_{E=0} \sim 2\left(\frac{\pi}{2m\lambda}\right)^{\frac{1}{2}} e^{-m\lambda}. \quad \text{(B29d)}$$

APPENDIX C. NONRELATIVISTIC SCATTERING WITH A SINGULAR POTENTIAL

Let us consider the partial-wave radial Schrödinger equation with a singular (r^{-2}) potential,

$$\left[\frac{d^2}{dr^2} + \frac{2}{r}\frac{d}{dr} + k^2 - \frac{l(l+1)}{r^2}\right]\Psi_l(r) = \left[\frac{a}{r^2} + \bar{V}(r)\right]\Psi_l(r). \quad \text{(C1)}$$

$\bar{V}(r)$ is assumed to vanish faster than r^{-1} as r goes to infinity. This equation can be rewritten in the form

$$\left[\frac{d^2}{dr^2} + \frac{2}{r}\frac{d}{dr} + k^2 - \frac{\bar{l}(\bar{l}+1)}{r^2}\right]\Psi_l(r) = \bar{V}(r)\Psi_l(r), \quad \text{(C2)}$$

where

$$\bar{l}(\bar{l}+1) = l(l+1) + a,$$

and

$$\bar{l} = -\tfrac{1}{2} + \left[(l+\tfrac{1}{2})^2 - \frac{a}{4}\right]^{\frac{1}{2}}. \quad \text{(C3)}$$

The free solutions ($\bar{V} = 0$) of this equation are $j_{\bar{l}}(kr)$ and $h_{\bar{l}}^{(1)}(kr)$ where j_l and $h_l^{(1)}$ are the spherical Bessel and Hankel functions, respectively. The solution to (C2) can be written as

$$\Psi_l(r) = A_1 j_{\bar{l}}(kr) + A_2 h_{\bar{l}}^{(1)}(kr) + \int r^2\,dr G_{\bar{l}}(r, r')\bar{V}(r')\Psi_l(r'), \quad \text{(C4)}$$

where

$$G_{\bar{l}}(r, r') = -ikj_{\bar{l}}(kr_<)h_{\bar{l}}^{(1)}(kr_>).$$

To determine A_1 and A_2 we require that $\Psi_l(r)$ be regular at the origin. Since $h_{\bar{l}}^{(1)}(x) \to x^{-\bar{l}-\frac{1}{2}}$, A_2 must be zero. Then by using the very general relations

$$\Psi(x) \sim e^{i\mathbf{k}\cdot\mathbf{x}} + (e^{ikr}/kr)f(\theta, \varphi)$$

or

$$\Psi_l(r) \sim A_l j_l(kr) + (e^{ikr}/kr)f_l, \quad \text{(C5)}$$

we see that the incoming spherical wave part of $\Psi_l(r)$ is unmodified by the potential. From the well known asymptotic form of $j_l(x)$,[13]

$$j_l(x) \sim \frac{1}{2x}[e^{i[x-(l+1)\frac{1}{2}\pi]} + e^{-i[x-(l+1)\frac{1}{2}\pi]}],$$

we see that the coefficient of e^{-ikr}/kr in the asymptotic form of $\Psi_l(r)$ is

$$\tfrac{1}{2}A_l e^{+i(l+1)\frac{1}{2}\pi}. \quad \text{(C6)}$$

Comparison of (C6) with (C4) provides the relation

$$\tfrac{1}{2}A_l e^{i(\bar{l}+1)\frac{1}{2}\pi} = A_l e^{i(l+1)\frac{1}{2}\pi}. \quad \text{(C7)}$$

Therefore, the solution to (C2) is given by

$$\Psi_l(r) = A_l e^{i(l-\bar{l})\frac{1}{2}\pi} j_{\bar{l}}(kr) + \int r^2\,dr\,G_{\bar{l}}(r, r')\bar{V}(r')\Psi_l(r'). \quad \text{(C8)}$$

This equation is now of the Fredholm type.

We can now extract the exact scattering amplitude for a r^{-2} potential as the coefficient of e^{ikr}/kr in (C8) with $\bar{V}(r) = 0$. The result is just

$$f_l = -\tfrac{1}{2}(iA_l)e^{il\frac{1}{2}\pi}(e^{-\bar{l}\pi i} - e^{-il\pi}), \quad \text{(C9)}$$

where we have subtracted off the outgoing part of the plane wave. Now, if we had solved the problem for negative energy ($k = i\mu$), Eq. (C4) would have been replaced by

$$\Psi_l'(r) = A_1' i_{\bar{l}}(\mu r) + A_2' k_{\bar{l}}(\mu r) + \int r^2\,dr\,G_{\bar{l}}'(r, r')\bar{V}(r')\Psi_l'(r'), \quad \text{(C10)}$$

where $i_l(x)$ and $k_l(x)$ denote modified spherical

Bessel functions of the first and second kind,

$$G_l'(r, r') = -(2\mu/\pi) i_l(\mu r_<) k_l(\mu r_>). \quad \text{(C11)}$$

To choose A_1' and A_2' we insist on regularity at the origin and the matching of the growing exponential of $\Psi_l'(r)$ with that from $e^{-\mu z \cdot x}$,[27]

$$e^{-\mu x \cos\theta} = \sum_{l=0}^{\infty} A_l' i_l(\mu r) y_{lm}(\Omega_x).$$

The equation for negative energy corresponding to (C5) is

$$\Psi_l'(r) = A_l' i_l(\mu r) + (e^{-\mu r}/\mu r) f_l'. \quad \text{(C12)}$$

Since the growing behavior of i_l and i_λ is the same, we have immediately $A_1' = A_1'$ and $A_2' = 0$. When we try to determine the scattering amplitude in this case, we see there is an ambiguity since[28]

$$i_l \sim (1/x)[e^x - e^{\pm(l\pi i)} e^{-x}]. \quad \text{(C13)}$$

This asymptotic behavior gives an f_l' of the form

$$f_l' = -A_l'[e^{\pm\lambda\pi i} - e^{\pm l\pi i}]. \quad \text{(C14)}$$

By a comparison with (C9) we see that the minus sign is the correct choice for the phase factor. This fact can be understood by reference to the region of validity for the asymptotic forms in (C13),

$$I_\nu(z) \sim [1/(2\pi z)^{\frac{1}{2}}][e^z F(z) + e^{-z - E(\nu+\frac{1}{2})\pi i} F(-z)], \quad |z| \to \infty, \quad \text{(C15)}$$
$$-(2+\epsilon)\tfrac{1}{2}\pi < \arg z < (2-\epsilon)\tfrac{1}{2}\pi, \quad \epsilon = \pm 1.$$

Since in the nonrelativistic case we know that we wish to be able to continue our results to positive energy, we want $I_\nu(e^{-i\frac{1}{2}\pi} z)$ to be identified with $J_\nu(z)$. This means that we must choose that value of ϵ in (C15) which makes the asymptotic form valid for $\arg z = -\frac{1}{2}\pi$. Therefore $\epsilon = +1$, and the choice of the minus sign in (C14) is justified.

In the relativistic problem we make the same choice; we can either argue by analogy with the nonrelativistic result or state that the correct form in either case is given by picking the correct continuation in the asymptotic expansion of $e^{ip \cdot x}$. If we wish the asymptotic expansion for pure imaginary p to give that for real p on substitution of $e^{-i\frac{1}{2}\pi} p$ for p, we are required to choose the negative phase factor for I_ν.

APPENDIX D

In this appendix we calculate the integral in (48) for $\nu = 0$,

$$I = \int_0^\infty \frac{dx}{x} K_0(x) I_0(x)(e^{-2x} - 1). \quad \text{(D1)}$$

To this end we use the identity[29]

$$K_0(x) I_0(x) = \int_0^\infty \frac{J_0(xt)}{(t^2+4)^{\frac{1}{2}}}\, dt. \quad \text{(D2)}$$

Substituting this into (D1), we can now carry out the x integration[30]

$$\int_0^\infty dx \frac{J_0(xt)}{x}(e^{-2x} - 1) = \lim_{\nu\to 0} \int_0^\infty \frac{J_\nu(xt)}{x}(e^{-2x} - 1)$$
$$= \lim_{\nu\to 0}\left[\frac{1}{\nu}\left[\frac{t}{2 + (t^2+4)^{\frac{1}{2}}}\right]^\nu - \frac{1}{\nu}\right]$$
$$= \ln\left[\frac{t}{t + (t^2+4)^{\frac{1}{2}}}\right]. \quad \text{(D3)}$$

Putting (D3) into (D2), we find[13]

$$I = \int_0^\infty \frac{dt}{(t^2+4)^{\frac{1}{2}}} \ln\left[\frac{t}{2 + (t^2+4)^{\frac{1}{2}}}\right]$$
$$= \int_0^\infty dy \ln\left(\frac{\sinh y}{1 + \cosh y}\right) = 2\int_0^\infty dz \ln(\tanh z)$$
$$= 2\int_0^1 \frac{\ln w}{1 - w^2}\, dw = -\frac{\pi^2}{4}. \quad \text{(D4)}$$

[27] This result is obtained by continuing $j_l(x)$ in the usual expansion of $e^{ip \cdot x}$ by the prescription of Ref. 16.

[28] See for example, G. N. Watson, *Theory of Bessel Functions* (Cambridge University Press, London, 1948), 2nd ed., pp. 201–203. Historically, this appears to be the first Stokes' phenomenon discovered [Sir G. G. Stokes, *Memoirs and Scientific Correspondence. I* (Cambridge University Press, London, 1907), p. 62].

[29] See Ref. 17, p. 330.

[30] See Ref. 18, Vol. I, pp. 182, 326.

On the Canonical Relativistic Kinematics of N-Particle Systems

A. CHAKRABARTI
Centre de Physique Théorique de l'Ecole Polytechnique, Paris, France
(Received 17 December 1963)

The "canonical" relativistic kinematics is discussed for n-particle systems. Our aim is to derive the formulas in as simple and *symmetrical* a way as possible and thus to avoid the extra complications arising for $n > 2$ in the usual stepwise generalization of the two-particle method. At first we derive the canonical form of the infinitesimal operators $(\mathbf{N}, \mathbf{M})$ in a highly symmetrical form. It is then shown that though the restrictions imposed by the condition that our states be also energy eigenstates compel us to sacrifice a part of the symmetry and simplicity of the formulas, we can indeed include the effect of the spins of the component particle in a completely symmetric way. This reduces to a minimum the additional complications introduced when the component particles have nonzero spins. The corresponding, relatively simple, generalized (C–G) coefficients connecting the canonical and the direct-product states are calculated. It is shown that the use of "spinor" representation for the individual particles simplifies the deductions considerably. Explicit results are given usually for $n = 3$ only, since the generalization to $n > 3$ introduces no essentially new features.

I. INTRODUCTION

In two previous articles[1,2] we have discussed in detail certain aspects of "canonical" relativistic kinematics. In those articles we mostly discussed (with reference to the canonical representation) the form of the wave equation, the "position" and spin operators and transformation properties for the one-particle case. In what follows our aim is to extend our discussion to many-particle systems, considering, however, only noninteracting particles of nonzero rest mass. The notation used will closely follow that of Refs. 1 and 2.

Let there be a system of n particles (free). The infinitesimal operators of the Poincaré group are, in the direct-product representation,

$$P = \sum_{i=1}^{n} p_i,$$
$$M \equiv (\mathbf{N}, \mathbf{M}) = \sum_{i=1}^{n} M_i \equiv \sum_{i=1}^{n} (\mathbf{N}_i, \mathbf{M}_i), \tag{1.1}$$

where

$$\mathbf{N}_i = -ip_i^0(\partial/\partial \mathbf{p}_k) + \mathbf{n}_i,$$
$$\mathbf{M}_i = -i\mathbf{p}_i \times \partial/\partial \mathbf{p}_i + \boldsymbol{\zeta}_i. \tag{1.2}$$

The explicit form of the intrinsic part $(\mathbf{n}_i, \boldsymbol{\zeta}_i)$ will depend on the representation used for the individual particles (e.g., "spinor" or "canonical" representation). This point will be discussed in Sec. II B.

Our aim is to reduce M to the canonical form by a suitable transformation of variables, namely to express $(\mathbf{N}, \mathbf{M})$ as

$$\mathbf{N} = -P^0\mathbf{X} - [1/(P^0 + m)]\mathbf{P} \times \mathbf{S},$$
$$\mathbf{M} = -\mathbf{P} \times \mathbf{X} + \mathbf{S}. \tag{1.3}$$

When $\mathbf{S}$, $\mathbf{X}$ may be considered, respectively, as the canonical spin and center-of-mass (c.m.) operators of the composite system. This corresponds to a "canonical" separation of the motion of the c.m., i.e., the motion of the system as a whole from the internal motions about the c.m. (This in turn enables us, for example, to simplify the S-matrix elements by separating out the dependence on the quantum numbers relating to the c.m., i.e., those which are conserved in the scattering process). Moreover the canonical form of (1.3) ensures[1] that the states belonging to the irreducible representation $(m^2 = P^2, \mathbf{S}^2)$ will transform under the Poincaré group according to the Wigner formula

$$(U(a, \Lambda)\varphi)(P) = \exp iP\cdot a\mathfrak{D}^{(S)}(\Lambda_P \cdot \Lambda \cdot \Lambda_P^{-1})\varphi(P') \tag{1.4}$$

where

$$\Lambda \cdot P' = P, \qquad \Lambda_P \cdot P = (m, \mathbf{0}) = \Lambda_{P'} \cdot P'$$

[or equivalently,

$$U(a, \Lambda)\,|P\rangle = \exp i\Lambda P \cdot a\mathfrak{D}^{(S)}(\Lambda_{\Lambda P} \cdot \Lambda \cdot \Lambda_P^{-1})\,|\Lambda \cdot P\rangle].$$

It is to be noted that (1.3) implies[2]

$$\mathbf{X} = -\frac{1}{P^0}\left[\mathbf{N} + \frac{1}{m(P^0 + m)}\mathbf{P} \times \mathbf{W}\right], \tag{1.5}$$

where

$$W = -P \cdot M^* = (\mathbf{P}\cdot\mathbf{M}, P^0\mathbf{M} - \mathbf{P} \times \mathbf{N}).$$

In fact (1.3) and (1.5) are equivalent, and adopting

[1] A. Chakrabarti, J. Math. Phys. 4, 1215 (1963).
[2] A. Chakrabarti, J. Math. Phys. 4, 1223 (1963).

the above definition of the c.m. at once reduces $(\mathbf{N}, \mathbf{M})$ to the canonical form with

$$\mathbf{S} = \mathbf{M} + \mathbf{P}\times\mathbf{X} = \frac{1}{m}\left(\mathbf{W} - \frac{\mathbf{P}}{P^0 + m}W^0\right), \quad (1.6)$$

implying

$$\Lambda_P \cdot W = (0, \mathbf{S}).$$

Thus our real task reduces to the elucidation of the structure of $\mathbf{S}$. In order that we may have sufficient quantum numbers for a complete classification of the canonical states belonging to (m^2, s^2), we have to express $\mathbf{S}$ as a sum of separate (mutually commuting) angular momentum operators which, coupled together, give the total spin. These operators must, in order that such a description be possible, commute also with other operators necessary completing the description. Moreover, it is desirable that they each possess a clear physical significance and the maximum possible simplicity.

The case $n = 2$ has been treated by several authors.[3,4] For $n \geq 3$, one method is to proceed step by step, namely combine first two particles and then a third with the c.m. of the first two and so on.[3,5] But this makes the structure of $\mathbf{S}$ and the "internal operators" extremely complicated.

This is particularly the case when the component particles have nonzero spins. [The final result for the orbital or c.m. part must always be that implied by (1.5).] Our aim is to introduce the maximum amount of symmetry and simplicity in the structure of $\mathbf{S}$.

For this purpose we introduce a change of variables, at first in a rather general form.

II. CHANGE OF VARIABLES

A. Particles of Zero Spin

Let us consider a system of n spinless particles, so that

$$\mathbf{N} = -i\sum_{i=1}^{n} p_i^0(\partial/\partial\mathbf{p}_i),$$
$$\mathbf{M} = -i\sum_{i=1}^{n} \mathbf{p}_i \times \partial/\partial\mathbf{p}_i. \quad (2.1)$$

Let

$$\boldsymbol{\pi}_i = \mathbf{p}_i - \lambda_i(\mathbf{p}_1 + \cdots + \mathbf{p}_n), \quad (2.2)$$

where for the moment the structure of the λ's are left undetermined except for the constraint

$$\sum_{i=1}^{n} \lambda_i = 1, \quad (2.3)$$

which implies

$$\sum_{i=1}^{n} \boldsymbol{\pi}_i = 0, \quad (2.4)$$

so that, writing

$$\boldsymbol{\pi}_n = -\left(\sum_{i=1}^{n-1} \boldsymbol{\pi}_i\right), \quad (2.5)$$

we can take the set $(\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_{n-1})$ as mutually independent.

Let us now make the change of variables

$$(\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n) \rightarrow (\mathbf{P}, \boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_{n-1}). \quad (2.6)$$

We have

$$\frac{\partial}{\partial\mathbf{p}_i} = \left(\frac{\partial\mathbf{P}}{\partial\mathbf{p}_i}\right)\cdot\frac{\partial}{\partial\mathbf{P}} + \sum_{j=1}^{n-1}\left(\frac{\partial\boldsymbol{\pi}_j}{\partial\mathbf{p}_i}\right)\cdot\frac{\partial}{\partial\boldsymbol{\pi}_j}, \quad (2.7)$$

where symbolically

$$(\partial\mathbf{P}/\partial\mathbf{p}_i)^{lm} \equiv \partial P^m/\partial p_i^l, \quad \text{etc.} \quad (2.8)$$

Thus

$$\frac{\partial}{\partial\mathbf{p}_i} = \frac{\partial}{\partial\mathbf{P}} + \frac{\partial}{\partial\boldsymbol{\pi}_i} - \sum_{j=1}^{n-1}\left\{\lambda_j\frac{\partial}{\partial\boldsymbol{\pi}_j} + \left(\frac{\partial\lambda_j}{\partial\mathbf{p}_i}\right)\left(\mathbf{P}\cdot\frac{\partial}{\partial\boldsymbol{\pi}_j}\right)\right\} \quad (i = 1, \cdots, n-1) \quad (2.9)$$

$$\frac{\partial}{\partial\mathbf{p}_n} = \frac{\partial}{\partial\mathbf{P}} - \sum_{j=1}^{n-1}\left\{\lambda_j\frac{\partial}{\partial\boldsymbol{\pi}_j} + \left(\frac{\partial\lambda_j}{\partial\mathbf{p}_n}\right)\left(\mathbf{P}\cdot\frac{\partial}{\partial\boldsymbol{\pi}_j}\right)\right\}.$$

This gives

$$\sum_{i=1}^{n}\mathbf{p}_i\times\frac{\partial}{\partial\mathbf{p}_i} = \left\{\mathbf{P}\times\frac{\partial}{\partial\mathbf{P}} + \sum_{j=1}^{n-1}\boldsymbol{\pi}_j\times\frac{\partial}{\partial\boldsymbol{\pi}_j}\right\} - i\sum_{j=1}^{n-1}(\mathbf{M}\lambda_j)\left(\mathbf{P}\cdot\frac{\partial}{\partial\boldsymbol{\pi}_j}\right), \quad (2.10)$$

where

$$\mathbf{M}\lambda_j \equiv -i\sum_{i=1}^{n}\mathbf{p}_i\times\frac{\partial\lambda_j}{\partial\mathbf{p}_i}. \quad (2.11)$$

Exactly similarly,

$$\sum_{i=1}^{n} p_i^0\frac{\partial}{\partial\mathbf{p}_i} = P^0\frac{\partial}{\partial\mathbf{P}} + \sum_{j=1}^{n-1}(p_j^0 - \lambda_j P^0)\frac{\partial}{\partial\boldsymbol{\pi}_j} - i\sum_{j=1}^{n-1}(\mathbf{N}\lambda_j)\left(\mathbf{P}\cdot\frac{\partial}{\partial\boldsymbol{\pi}_j}\right), \quad (2.12)$$

where

$$\mathbf{N}\lambda_j \equiv -i\sum_{i=1}^{n} p_i^0\frac{\partial\lambda_j}{\partial\mathbf{p}_i}. \quad (2.13)$$

[3] B. Barsella and E. Fabri, Phys. Rev. **128**, 451 (1962).
[4] A. J. Macfarlane, J. Math. Phys. **4**, 490 (1963).
[5] A. J. Macfarlane, Rev. Mod. Phys. **34**, 41 (1962).

From (2.10) and (2.11) we see that we have only to take the λ's to be invariant under purely spatial rotations ($\mathbf{M}\lambda_i = 0$) to have

$$\sum_{i=1}^{n} \mathbf{p}_i \times \frac{\partial}{\partial \mathbf{p}_i} = \mathbf{P} \times \frac{\partial}{\partial \mathbf{P}} + \sum_{i=1}^{n-1} \boldsymbol{\pi}_i \times \frac{\partial}{\partial \boldsymbol{\pi}_i}. \tag{2.14}$$

For further restrictions on the λ's we examine (2.12). In the nonrelativistic case it suffices to choose the λ_i's to be constants. In the relativistic case we note that, if the λ's are supposed not only to be rotation- but Lorentz-invariant as well ($\mathbf{N}\lambda_i = 0$), then (using primes to distinguish this particular choice)

$$\sum_{i=1}^{n} p_i^0 \frac{\partial}{\partial \mathbf{p}_i} = P^0 \frac{\partial}{\partial \mathbf{P}} + \sum_{i=1}^{n-1} \pi_i'^0 \frac{\partial}{\partial \boldsymbol{\pi}_i'}, \tag{2.15}$$

as well as

$$\sum_{i=1}^{n} \mathbf{p}_i \times \frac{\partial}{\partial \mathbf{p}_i} = \mathbf{P} \times \frac{\partial}{\partial \mathbf{P}} + \sum_{i=1}^{n-1} \boldsymbol{\pi}_i' \times \frac{\partial}{\partial \boldsymbol{\pi}_i'},$$

where

$$\pi_i'^0 = p_i^0 - \lambda_i' P^0, \quad \text{or} \quad \boldsymbol{\pi}_i' = \mathbf{p}_i - \lambda_i' \mathbf{P}. \tag{2.16}$$

This leaves the λ's yet to some extent undetermined which enables us to impose the condition

$$P \cdot \pi_i' = 0, \tag{2.17}$$

giving

$$\lambda_i' = P \cdot p_i / P^2. \tag{2.18}$$

This mode of separation of the c.m. motion (reducing for $n = 2$ to that employed by Wightman[6]) has a certain simplicity. But, evidently, this does not give us the canonical form and hence does not lead to the associated desirable properties.

Comparing (1.3), (2.12), and (2.14), we find that the required condition for a canonical splitting is

$$-\frac{1}{P^0 + m} \mathbf{P} \times \left(\boldsymbol{\pi}_i \times \frac{\partial}{\partial \boldsymbol{\pi}_i} \right) = (p_i^0 - \lambda_i P^0) \frac{\partial}{\partial \boldsymbol{\pi}_i} - i(\mathbf{N}\lambda_i) \left(\mathbf{P} \cdot \frac{\partial}{\partial \boldsymbol{\pi}_i} \right). \tag{2.19}$$

The solution is, as may easily be verified,

$$\lambda_i = (P \cdot p_i + m p_i^0)/m(P^0 + m). \tag{2.20}$$

This means that, for the canonical separation,

$$\boldsymbol{\pi}_i = \mathbf{p}_i - [(P \cdot p_i + m p_i^0)/m(P^0 + m)]\mathbf{P}, \tag{2.21}$$

which is just the space part of $\Lambda_P \cdot p_i$ or p_i transformed to the c.m. system. This gives the clear physical significance, we might have expected. (In Appendix II we collect certain formulas relating to the $\pi''s$ and $\pi's$).

Thus for n spinless particles we can write in (1.3)

$$\mathbf{S} = -i \sum_{i=1}^{n-1} \boldsymbol{\pi}_i \times \frac{\partial}{\partial \boldsymbol{\pi}_i}, \tag{2.22}$$

and

$$\mathbf{M} = -i\mathbf{P} \times \partial/\partial \mathbf{P} + \mathbf{S},$$
$$\mathbf{N} = -iP^0(\partial/\partial \mathbf{P}) - [1/(P^0 + m)]\mathbf{P} \times \mathbf{S}, \tag{2.23}$$

where the general properties of the canonical representation allows us to write down directly

$$-i \frac{\partial}{\partial \mathbf{P}} = -\frac{1}{P^0} \left[\mathbf{N} + \frac{1}{m(P^0 + m)} \mathbf{P} \times \mathbf{W} \right]. \tag{2.24}$$

The actual expressions of the $\partial/\partial \boldsymbol{\pi}_i$'s are quite complicated but may be evaluated in a straightforward manner from (2.9) (see Appendix I).

The total operator $\mathbf{S}$ is of course the same as that obtained by proceeding stepwise. Our result only expresses the same in a more symmetric and simple way involving only one Lorentz transformation to the c.m. frame. The operator $\boldsymbol{\pi}_n$ corresponding to only one particle (arbitrary) has been allowed a special role by being eliminated. We have established that this somewhat unconventional choice does lead to a canonical form. Unfortunately, this attractive choice encounters a certain difficulty, which we discuss later (Sec. III).

B. Particles of Nonzero Spin

When the component particles have spin, some additional complications arise. In this case, since there are the additional terms $(\mathbf{n}_i, \boldsymbol{\zeta}_i)$ and, in order to assure the canonical form, we always define the c.m. operator as in (1.5), we have to add to $-i(\partial/\partial \mathbf{P})$, as evaluated in II A (and termed $\mathbf{X}_{0r}$ henceforth), the term

$$\mathbf{X}_\Sigma = \sum_{i=1}^{n} \mathbf{X}_{\Sigma_i} = -\sum_{i=1}^{n} \frac{1}{P^0} \left[\mathbf{n}_i + \frac{1}{m(P^0 + m)} \mathbf{P} \times (P^0 \boldsymbol{\zeta}_i - \mathbf{P} \times \mathbf{n}_i) \right] \tag{2.25}$$

$$= -\frac{1}{P^0} \left[\mathbf{n} + \frac{1}{m(P^0 + m)} \mathbf{P} \times (P^0 \boldsymbol{\zeta} - \mathbf{P} \times \mathbf{n}) \right], \tag{2.26}$$

where

$$(\mathbf{n}, \boldsymbol{\zeta}) = \sum_{i=1}^{n} (\mathbf{n}_i, \boldsymbol{\zeta}_i).$$

[6] A. S. Wightman, "L'invariance dans la mécanique relativiste," Ecole d'été de Physique Théorique, les Houches, (Hermann & Cie., Paris, 1960).

At this stage we have to decide what representation to employ for the component particles. If we use the canonical representation throughout, then

$$\mathbf{n}_i = -[1/(p_i^0 + m_i)]\mathbf{p}_i \times \boldsymbol{\zeta}_i. \tag{2.27}$$

But instead of doing so, let us at first suppose that we are using "spinor" representation[7,8] for the component particles. In this case $(\mathbf{n}_i, \boldsymbol{\zeta}_i)$ transforms as an antisymmetric tensor and commutes with $\partial/\partial\mathbf{p}_i$ (as well as, of course, with $\mathbf{p}_i$). *None of these properties are true for the canonical representation when* $\mathbf{n}_i$ *is given by* (2.27).

Thus the use of spinor representations for the particles $i(= 1, \cdots, n)$ enables us to write, according to the well-known transformation laws for antisymmetric tensors,[9] for $\boldsymbol{\zeta}_{i(0)}$ (or $\boldsymbol{\zeta}_i$ transformed to the c.m. frame),

$$\boldsymbol{\zeta}_{i(0)} = \boldsymbol{\zeta}_i - \frac{\mathbf{P}\times(\mathbf{P}\times\boldsymbol{\zeta}_i)}{m(P^0+m)} - \frac{\mathbf{P}\times\mathbf{n}_i}{m} \tag{2.28}$$

$$= \boldsymbol{\zeta}_i + \mathbf{P}\times\mathbf{X}_{\Sigma i}, \tag{2.29}$$

from (2.26). From (2.29) it follows that we can write

$$\mathbf{M} = -\mathbf{P}\times\mathbf{X} + \mathbf{S}_{0r} + \mathbf{S}_\Sigma, \tag{2.30}$$

where

$$\mathbf{X} = \mathbf{X}_{0r} + \mathbf{X}_\Sigma, \tag{2.31}$$

$$\mathbf{S}_{0r} = -i\sum_{j=1}^{n-1} \boldsymbol{\pi}_j \times \frac{\partial}{\partial\boldsymbol{\pi}_j}, \tag{2.32}$$

$$\mathbf{S}_\Sigma = \sum_{i=1}^{n} \boldsymbol{\zeta}_{i(0)} = \boldsymbol{\zeta}_{(0)}. \tag{2.33}$$

The choice of definition of $\mathbf{X}$ automatically gives us

$$\mathbf{N} = -P^0\mathbf{X} - [1/(P^0+m)]\mathbf{P}\times(\mathbf{S}_{0r} + \mathbf{S}_\Sigma). \tag{2.34}$$

If, on the other hand, we use canonical representation for the particles "i" [and hence (2.27)], the rhs of (2.28) and (2.29) no longer has the simple physical significance discussed above, the transformation law for the canonical $\boldsymbol{\zeta}_i$ being quite different.[1,3,4]

As has been discussed before,[1] the spinor representation can be transformed to the canonical one through a similarity transformation by the operator $Q(p, \Lambda_p)$, where $Q(p, \Lambda)$ is the matrix of transformation of the spinor wavefunctions under the operation Λ.

Now, suppose we transform both sides of (2.30) and (2.34) by the operator

$$\prod_{i=1}^{n} Q(p_i, \Lambda_{p_i}). \tag{2.35}$$

On the lhs, $\mathbf{N}$ and $\mathbf{M}$ will be converted, respectively, to the sums of the $\mathbf{N}_i$'s and $\mathbf{M}_i$'s in the canonical form. On the rhs, the ∂'s involved in the definitions of $\mathbf{X}$ and the $\partial/\partial\boldsymbol{\pi}_i$'s will now have their canonical significance, i.e., for nonzero spins they will have complicated nonlocal structures. And as for $\mathbf{S}_\Sigma$, we find that this is obtained in the form

$$(\mathbf{S}_\Sigma)_{tr} = \sum_{i=1}^{n} (\boldsymbol{\zeta}_{i(0)})_{tr} \tag{2.36}$$

where $(\boldsymbol{\zeta}_{i(0)})_{tr} = Q_i\boldsymbol{\zeta}_{i(0)}Q_i^{-1}$ is, in general, quite different from $(\boldsymbol{\zeta}_{i\,tr})_{(0)}$, namely the canonical $\boldsymbol{\zeta}$'s transformed according to appropriate law to the c.m. system. Barsella and Fabri[3] find (for $n = 2$) that in order to express $\mathbf{S}_\Sigma$ as the sum of the component canonical $\boldsymbol{\zeta}$'s reduced to the c.m. system, they have to modify the definition of the $\boldsymbol{\pi}_i$'s by introducing spin-dependent terms. However, even for this simplest case ($n = 2$), these terms are extremely complicated. (In fact, they do not attempt an explicit evaluation.)

If our starting point itself is the c.m. frame ($\mathbf{p} = 0$), then of course (2.36) becomes equivalent to

$$\sum_{i=1}^{n} (\boldsymbol{\zeta}_{i\,tr})_{(0)}.$$

III. CONSEQUENCES OF THE RESTRICTION OF ENERGY CONSERVATION

We have

$$P^0 = \sum_{i=1}^{n} (\mathbf{p}_i^2 + m_i^2)^{\frac{1}{2}} \tag{3.1}$$

$$= \left[\mathbf{P}^2 + \left\{\sum_{i=1}^{n} (\boldsymbol{\pi}_i^2 + m_i^2)^{\frac{1}{2}}\right\}^2\right]^{\frac{1}{2}}. \tag{3.2}$$

P^0 commutes both with $-\mathbf{P}\times\mathbf{X}$ and $\mathbf{S}$ as a consequence of the canonical definition (1.5) of $\mathbf{X}$ implying

$$-i[\mathbf{X}, P^0] = \mathbf{P}/P^0. \tag{3.3}$$

But the presence of the term $(\boldsymbol{\pi}_n^2 + m_n^2)$ in (3.2), where

$$\boldsymbol{\pi}_n = -\sum_{j=1}^{n-1} \boldsymbol{\pi}_j,$$

prevents the component terms $\boldsymbol{\pi}_i \times (\partial/\partial\boldsymbol{\pi}_i)$ from commuting separately with P^0. Thus if we want our canonical states to correspond to definite energy values, we are forced to abandon the simple sym-

[7] V. Bargmann and E. Wigner, Proc. Natl. Acad. Sci. U. S. 34, 211 (1948).
[8] H. Joos, Fortschr. Physik. 10, 65 (1962).
[9] C. Möller, *Theory of Relativity* (Clarendon Press, Oxford, England, 1955).

metric form of $\mathbf{S}_{0r}$ given in (2.32), since, for a complete enumeration of the states, the quantum number corresponding to the total spin is not sufficient, but we have to introduce subsidiary quantum numbers specifying the mode of coupling adopted to construct the total spin.

For $n = 2$, of course, there is no problem (since $\pi_2 = -\pi_1 = \pi$, say). For $n \geq 3$ one way to avoid the above-mentioned difficulty is to proceed by steps.[3,5]

But, in our method of the derivation of the canonical forms of $\mathbf{M}$ and $\mathbf{N}$, we have been able to separate out the contributions from the orbital and the spin parts of the component particles [as is indicated by the suffixes "0r" and "Σ," respectively, in (2.30), (2.34)]. Taking advantage of this fact, we can substitute a different equivalent expression for $\mathbf{S}_{0r}$ alone, leaving $\mathbf{S}_\Sigma$ in its simple symmetrical form.

Thus for $n = 3$ we may write

$$\mathbf{S}_{0r} = -i\left(\pi_1 \times \frac{\partial}{\partial \pi_1} + \pi_2 \times \frac{\partial}{\partial \pi_2}\right) \tag{3.4}$$

$$= -i\left(\pi_{12} \times \frac{\partial}{\partial \pi_{12}} + \pi_{(12)3} \times \frac{\partial}{\partial \pi_{(12)3}}\right), \tag{3.5}$$

where

$$\pi_{12} = \Lambda_{p_1+p_2} \cdot p_1, \qquad \pi_{(12)3} = \Lambda_{p_1+p_2+p_3} \cdot (p_1 + p_2).$$

[The equality of (3.4) and (3.5) is assured by the fact that the term $-\mathbf{P} \times \mathbf{X}_{0r}$ is the same for both cases which, in turn, follows from the fact that $\mathbf{N}$, $\mathbf{M}$ appear linearly in (1.5). This is of course the result one expects from a proper definition of the c.m.]

Since, in terms of these new variables,

$$P^0 = [\mathbf{P}^2 + \{(\pi_{12}^2 + m_{12}^2)^{\frac{1}{2}} + (\pi_{(12)3}^2 + m_3^2)^{\frac{1}{2}}\}^2]^{\frac{1}{2}} \tag{3.6}$$

where

$$m_{12} = (\pi_{12}^2 + m_2^2)^{\frac{1}{2}} + (\pi_{12}^2 + m_2^2)^{\frac{1}{2}}, \tag{3.7}$$

it may easily be seen both the terms on the rhs of (3.5) commute with P^0.

Thus, we may write, for $n = 3$,

$$\mathbf{M} = -\mathbf{P} \times \mathbf{X} - i\left(\pi_{12} \times \frac{\partial}{\partial \pi_{12}} + \pi_{(12)3} \times \frac{\partial}{\partial \pi_{(12)3}}\right) + (\zeta_{1(0)} + \zeta_{2(0)} + \zeta_{3(0)}). \tag{3.8}$$

Another alternative form for $\mathbf{S}_{0r}$ is obtained by starting the steps from the other end, so to say, i.e., making the substitution $\pi_1, \pi_2 \rightarrow \pi, \pi_0$, where $\pi = (\pi_1 + \pi_2)$ and $\pi_0 = \Lambda_{\pi_1+\pi_2} \cdot \pi_1$. In fact, the only difference between the two schemes is that π_0 differs, in general, from $\pi_{(12)}$ by a rotation due to the fact that the pure Lorentz transformations do not form a group.

It may be verified without difficulty that the terms $\zeta_{i(0)}$ do, indeed, have all the required commutation properties (see Appendix III), and hence the corresponding quantum numbers may be utilized, without any inconsistency, to classify the states.

For $n > 3$ the required generalization raises no problem.

The chief difference between our formula (3.8) (generalized for $n > 3$ if necessary) and that obtained by usual method of combining both the spins and momenta by steps is, of course, the symmetrical form of $\mathbf{S}_\Sigma$ in our case. But it should be noticed also that, though no longer expressed in a symmetrical form, the π's remain free of the spin matrices. In the other method the intermediate c.m. operator for each stage, in general, involves the corresponding spin operators [see the definition of X_Σ in (2.25)], which in turn make the definition of the subsequent internal coordinates (π's) extremely complicated.

IV. THE GENERALIZED C-G COEFFICIENTS

In this section we propose to derive the coefficients connecting the direct-product and canonical states.

Let us consider the case $n = 3$.

In order to express a state belonging to the irreducible representation (P^2, S^2) in terms of the direct-product states, we must do two things:

(a) transform over to the new variables involved [see (3.8)]; and

(b) couple together, according to the standard rules, the component angular momenta to obtain the total spin S.

Our procedure to this end will be analogous to that of Shirokov for the two-particle case.[10] For maximum simplicity we adopt the rest system of reference, namely $\mathbf{P} = 0$. In this frame, $\mathbf{M}$ becomes equivalent to

$$-i\left(\pi_{12} \times \frac{\partial}{\partial \pi_{12}} + \pi_{(12)3} \times \frac{\partial}{\partial \pi_{(12)3}}\right) + (\zeta_1 + \zeta_2 + \zeta_3), \tag{4.1}$$

where, now

$$\pi_{(12)3} = \mathbf{p}_1 + \mathbf{p}_2 = -\mathbf{p}_3.$$

Along with P^0, $\mathbf{P}(=0)$, the five components of (4.1)

[10] Iu. M. Shirokov, Zh. Eksperim. i Teor. Fiz. 35, 1005 (1958) [English transl.: Soviet Phys.—JETP 8, 703 (1959)].

will furnish us with sufficient quantum numbers to classify the states completely.

The expression (3.1) [and (4.1) as a special case] for $\mathbf{M}$ assures us two things:

(a) the total angular momentum obtained by coupling the above five contributions does, indeed, represent the total spin S of the composite system in the rest frame.

(b) the state thus constructed will transform under the Poincaré group according to the standard canonical rules[1] for the irreducible representation (P^2, S^2).

In this particular system of reference, no transformation is necessary to pass from ζ_i to $\zeta_{i(0)}$ $(i = 1, 2, 3)$.

As for momenta, we note that, since an eigenstate of $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$ is already an eigenstate of $\mathbf{P}, \pi_{12}, \pi_{(12)3}$, we have only to convert the momentum eigenfunctions into angular momentum ones. The required expansion coefficients are the spherical harmonics

$$Y_{l'm'}(\hat{\pi}_{12}), \quad Y_{l''m''}(\hat{\pi}_{(12)3}). \tag{4.2}$$

[The indices $\hat{\pi}_{12}$, $\hat{\pi}_{(12)3}$ denote that we have to take the angular coordinates corresponding to the directions of the unit vectors $\hat{\pi}_{12}$, $\hat{\pi}_{(12)3}$, respectively.] Apart from this, however, we have to introduce certain factors to ensure the orthonormality of canonical states. These factors arise from the Jacobian of transformation corresponding to the change of variables effected in the integrals over the momentum space and are related to the changes in the density of states. The evaluation of these factors is discussed in Appendix IV. For our case $(n = 3)$, the required expression is

$$\rho_{(12)3} = 2m_{12}^{\frac{1}{2}}\lambda^{-\frac{1}{2}}(m_{12}^2, m_1^2, m_2^2)\cdot 2m^{\frac{1}{2}} \times \lambda^{-\frac{1}{2}}(m^2, m_{12}^2, m_3^2), \tag{4.3}$$

where

$$m_{12}^2 = (p_1 + p_2)^2$$

and

$$\lambda(a, b, c) = a^2 + b^2 + c^2 - 2(bc + ca + ab).$$

It is shown (in Appendix IV) that this factor does indeed lead to correctly orthonormalized states. As a last step we now must couple together the five component angular momenta. The corresponding coupling coefficient may be written as

$$(\sigma_1\sigma_2\sigma_3 m'm'' \mid \zeta'\zeta l S\Sigma) = (\zeta_1\zeta_2\sigma_1\sigma_2 \mid \zeta_1\zeta_2\zeta'\sigma')(\zeta'\zeta_3\sigma'\sigma^3 \mid \zeta'\zeta_3\zeta\sigma) \times (l'l''m'm'' \mid l'l''lm)(\zeta l\sigma m \mid \zeta l S\Sigma), \tag{4.4}$$

where the rhs is expressed in terms of the usual C–G coefficients σ_i, Σ denoting the z components of ζ_i and S, respectively. Thus finally, we have [writing $K = (m, \mathbf{0})$]

$$|K, \Sigma\,[m, S]; \zeta'\zeta l\rangle = \sum \delta(K - \sum_i p_i)\rho_{(12)3} \times (\sigma_1\sigma_2\sigma_3 m'm'' \mid \zeta'\zeta l S\Sigma), \quad Y_{l'm'}(\hat{\pi}_{12})Y_{l''m''}(\hat{\pi}_{(12)3})\Big(\prod_{i=1}^{3} |p_i, \sigma_i[m_i, \zeta_i]\rangle\Big), \tag{4.5}$$

where the summation is to be taken over the indices $\sigma_1, \sigma_2, \sigma_3, m', m''$, and is supposed to include the integrations over the momenta as well. [On the lhs we have suppressed for the sake of brevity such additional quantum numbers as m_i, ζ_i $(i = 1, 2, 3)$, l', l''].

Had we chosen an arbitrary frame of reference (instead of $\mathbf{p} = 0$), the changes of variable $\zeta_i \to \zeta_{i(0)}$ would have involved the matrices $Q^{\zeta_i}(p_i, \Lambda_P)$ corresponding to the transformations of the component spinor wavefunctions[1] instead of the rotation matrices that arise[5,10] when the canonical representation is used for the particles i. But, for the rest frame, in light of the discussion following (2.36) [namely equivalence of $(\zeta_{i(0)})_{tr}$ and $(\zeta_{i\,tr})_{(0)}$ for $\mathbf{P} = 0$], we find that (4.1) holds for both the representations for the particles i. Thus, (4.5) also holds for both the cases. Moreover, in both cases, once we have classified the states in any one frame of reference (the rest frame, for example), the transition to any other frame presents no problem, since the canonical form of $(\mathbf{N}, \mathbf{M})$ guarantees the transformation law (1.4).

The generalization to cases $n > 3$ presents no difficulty.

V. CONCLUSION

To sum up we may say that though we have not been able to fully utilize, due to commutation difficulties, the maximum amount of simplicity and symmetry in the formulas obtained at first [(2.30)], we have shown that at least a part of the attractive features may be retained, thus reducing the additional complications due to the particle spins to a minimum.

Recently Werle[11] has given a rather simple method for the relativistic many-body problem, using helicity couplings and quantum numbers related to moving axes. Compared to this latter one, our method (based on l–s coupling) has the two advantages that it gives a simple (canonical) transformation

[11] J. Werle, Nucl. Phys. **44**, 579, 637 (1963).

law for the states constructed under Lorentz transformation also (and not only for rotation), and that in the classification of the states the "inner" quantum numbers are, from the beginning, discrete ones.

We may add that once we have the generalized C-G coefficients, the applications to the S-matrix theory may be carried out as usual,[5] and present no essentially new features.

APPENDIX I

In this section we briefly indicate the explicit evaluation of the operators $\partial/\partial\pi_i$ appearing in (2.22). Let us consider, for the sake of simplicity, the case $n = 3$. (For $n > 3$ we may proceed exactly similarly.)

Subtracting the last equation of (2.9) from the preceding ones, we have (for $n = 3$), writing $\partial_{ij} \equiv \partial/\partial\mathbf{p}_i - \partial/\partial\mathbf{p}_j$,

$$\begin{aligned}\partial_{13} &= \frac{\partial}{\partial\pi_1} - (\partial_{13}\lambda_1)\left(\mathbf{P}\cdot\frac{\partial}{\partial\pi_1}\right) \\ &\qquad - (\partial_{13}\lambda_2)\left(\mathbf{P}\cdot\frac{\partial}{\partial\pi_2}\right), \\ \partial_{23} &= \frac{\partial}{\partial\pi_2} - (\partial_{23}\lambda_1)\left(\mathbf{P}\cdot\frac{\partial}{\partial\pi_1}\right) \\ &\qquad - (\partial_{23}\lambda_2)\left(\mathbf{P}\cdot\frac{\partial}{\partial\pi_2}\right).\end{aligned} \tag{I 1}$$

Taking the scalar products of both sides of the above two equations with $\mathbf{P}$, we get two equations in the two unknowns $(\mathbf{P}\cdot\partial/\partial\pi_1)$, $(\mathbf{P}\cdot\partial/\partial\pi_2)$. These give directly

$$\begin{aligned}\left(\mathbf{P}\cdot\frac{\partial}{\partial\pi_1}\right) &= [(1 - \mathbf{P}\cdot\partial_{23}\lambda_2)(\mathbf{P}\cdot\partial_{13}) \\ &\quad + (\mathbf{P}\cdot\partial_{13}\lambda_2)(\mathbf{P}\cdot\partial_{23})] \\ &\quad \times [(1 - \mathbf{P}\cdot\partial_{23}\lambda_2)(1 - \mathbf{P}\cdot\partial_{13}\lambda_1) \\ &\quad - (\mathbf{P}\cdot\partial_{13}\lambda_2)(\mathbf{P}\cdot\partial_{23}\lambda_2)]^{-1},\end{aligned} \tag{I 2}$$

and a similar equation for $(\mathbf{P}\cdot\partial/\partial\pi_2)$ obtained by interchanging the indices 1 and 2. Now substituting these values of $(\mathbf{P}\cdot\partial/\partial\pi_1)$, $(\mathbf{P}\cdot\partial/\partial\pi_2)$ in (I 1) and transposing we get the desired result.

For the particularly simple case of $n = 2$, Barsella and Fabri[3] has given the explicit formula.

APPENDIX II

In this section we collect certain formulas concerning the variables introduced in (2.15) and (2.21).

First we note that

$$\Lambda_P\cdot\pi_i' = (0, \pi_i), \tag{II 1}$$

implying

$$\pi_i'^2 = -\pi_i^2. \tag{II 2}$$

Thus

$$\pi_i^2 = -(p_i - \lambda_i' P)^2 = (P\cdot p_i)^2/P^2 - p_i^2 \tag{II 3}$$

$$= \lambda(P^2, p_i^2, (P - p_i)^2)/4P^2, \tag{II 4}$$

where

$$\lambda(a, b, c) = a^2 + b^2 + c^2 - 2(bc + ca + ab).$$

For $n = 2$, (II 4) reduces to

$$\pi_1^2 = \lambda(m^2, m_1^2, m_2^2)/4m^2 = \pi_2^2. \tag{II 5}$$

These results, of course, follow directly from the fact that $\Lambda_P\cdot p_i = \pi_i$. Using the fact that $p_i = \Lambda_P^{-1}\cdot\pi_i$, we can express directly $\mathbf{p}_i$ as

$$\mathbf{p}_i = \pi_i + \left(\frac{(\pi_i^2 + m_i^2)^{\frac{1}{2}}}{m} + \frac{\pi\cdot\mathbf{P}}{m(m + P^0)}\right)\mathbf{P}. \tag{II 6}$$

We may add that the π_i's may be related to Λ_P in another way.

Considering the 4-vector W for n spin-zero particles (or the contribution of the orbital parts only to W in the general case) we have [writing $\mathbf{x}_i = i(\partial/\partial\mathbf{p}_i)$]

$$\mathbf{W} = -\sum_{i=1}^{n} p_i^0(\mathbf{p}_i/p_i^0 - \mathbf{P}/P^0)\times\mathbf{x}_i, \tag{II 7}$$

where we note that the expression $(\mathbf{p}_i/p_i^0 - \mathbf{P}/P_0^0)$ is just the vector difference of the relativistic 3-velocities corresponding to $\mathbf{p}_i$ and $\mathbf{P}$, respectively. The following relation may be verified:

$$\Lambda_P\cdot(\mathbf{P}\cdot\pi_i/P^0, \pi_i) = (0, p_i^0(\mathbf{p}_i/p_i^0 - \mathbf{P}/P^0)). \tag{II 8}$$

APPENDIX III

In this section we demonstrate that the operators $\zeta_{i(0)}$ have all the required commutation properties [see the remark following (3.8)].

Let us consider the case $n = 3$, and write as in (3.8)

$$\begin{aligned}\mathbf{M} &= -\mathbf{P}\times\mathbf{X} - i\left(\pi_{12}\times\frac{\partial}{\partial\pi_{12}} + \pi_{(12)3}\times\frac{\partial}{\partial\pi_{(12)3}}\right) \\ &\qquad + (\zeta_{1(0)} + \zeta_{2(0)} + \zeta_{3(0)}) \\ &= -\mathbf{P}\times\mathbf{X} + \mathbf{S}_{0r} + \mathbf{S}_\Sigma.\end{aligned} \tag{III 1}$$

Now as a consequence of our choice of spinor representation for the component particles, $\zeta_{i(0)}$ contains only $\mathbf{P}$, P^0 apart from $(\mathbf{n}_i, \zeta_i)$ which commute with the orbital operators $\mathbf{p}_i$, $\partial/\partial\mathbf{p}_i$. Thus evidently the $\zeta_{i(0)}$'s commute mutually and also separately with the $\pi_{12}\times\partial/\pi_{12}$ and $\pi_{(12)3}\times\partial/\partial\pi_{(12)3}$, since these

latter operators have been constructed to commute with P^0 as well as $\mathbf{P}$.

As a consequence of the general formulas for the canonical representation [(1.3), (1.5)] we have the result (easily verified) that $-\mathbf{P}\times\mathbf{X}$ commutes with $\mathbf{S}$. Moreover, since the use of the spinor representation for the component particles allows us to conclude by inspection that $\mathbf{S}_{0r}$ commutes with both $\mathbf{P}\times\mathbf{X}_{0r}$ (since nothing is changed in the expressions obtained for spin-zero particles) and $\mathbf{P}\times\mathbf{X}_{\Sigma}$ (since the latter involves only $\mathbf{P}$, P^0 and the individual spin matrices), we have the stronger result

$$[\mathbf{P}\times\mathbf{X},\ \mathbf{S}_{\Sigma}]_- = 0. \tag{III 2}$$

We may easily convince ourselves [from (2.26) and (2.33)] that replacing $(\mathfrak{n},\ \zeta)$ by $(\mathfrak{n}_i,\ \zeta_i)$ preserves the commutation relation. This gives

$$[\mathbf{P}\times\mathbf{X}_{0r} + \mathbf{P}\times\mathbf{X}_{\Sigma i},\ \zeta_{i(0)}]_- = 0. \tag{III 3}$$

But, evidently, $\zeta_{i(0)}$ commutes with $\mathbf{P}\times\mathbf{X}_{\Sigma j}$ $(j \neq i)$. Hence, adding these terms to $\mathbf{P}\times\mathbf{X}_{\Sigma i}$, we have

$$[\mathbf{P}\times\mathbf{X},\ \zeta_{i(0)}]_- = 0, \tag{III 4}$$

which is the required result.

It is to be noted that it is the use of spinor representation for the component particles that enables us to establish these results practically by inspection.

For $n > 3$ no new problems arise.

APPENDIX IV

In this section we discuss the effect of the change of variables leading to the canonical representation on the integrals over the momentum space.

The evaluation of the factor $\rho_{(12)3}$ in (4.3) (also its generalizations for $n > 3$) and the proof that it does lead to properly orthonormalized states may be discussed exactly as in Ref. 5. So actually we need no new discussion for our final result (4.5). However, for the sake of completeness we indicate how to evaluate the effect of the change of variables (2.6), which includes the $n = 2$ case as a particular one.

We may proceed in a fashion quite similar to that for $n = 2$,[5] by using (2.16) as an intermediate step and then utilizing (III1). We have

$$\int \prod_{i=1}^{n} \frac{d\mathbf{p}_i}{2p_i^0} = \int \prod_{i=1}^{n} d^4p_i\ \delta(p_i^2 - m_i^2)\theta(p_i^0) \tag{IV 1}$$

$$= \int \prod_{i=1}^{n} d^4p_i\ \delta(p_i^2 - m_i^2)$$

$$\times\ \theta(p_i^0)\ d^4P\ d\pi_1' \cdots d\pi_{n-1}'$$

$$\times\ \delta(P - \sum_i p_i)\delta(\pi_1' - p_1 + \lambda_1' P) \cdots$$

$$\times\ \delta(\pi_{n-1}' - p_{n-1} + \lambda_{n-1}' P), \tag{IV 2}$$

the δ functions added automatically implying also

$$\delta(\pi_n' - p_n + \lambda_n' p)] = \int \theta(P^0)\delta(P^2 - m^2)\ d^4P$$

$$\times \prod_{i=1}^{n-1} \theta(\pi_i'^0 + \lambda_i' P^0)\delta(\pi_i'^2 + \lambda_i'^2 P^2 - m_i^2). \tag{IV 3}$$

[The θ functions now serve to ensure the space-like nature of the π_i''s (since $P\cdot\pi_i' = 0$) by limiting in magnitude the negative value of $\pi_i'^0$'s which are zero only for $\mathbf{P} = 0$.]

Now, passing on to the variables π_i and noting that

$$\Lambda_P \cdot \pi_i' = (0, \pi_i), \tag{IV 4}$$

we have (exactly so for $n = 2$)

$$\int d\pi_i' \delta(\pi_i'^2 + \xi_i^2),$$

where

$$\xi_i^2 = \lambda_i'^2 P^2 - m_i^2$$

$$= \int d\pi_i \delta(-\pi_i^2 + \xi_i^2) \tag{IV 5}$$

$$= \int \frac{\xi_i}{2}\, d\Omega(\hat{\pi}_i). \tag{IV 6}$$

In (IV 6), $d\Omega(\hat{\pi}_i)$ is the solid-angle element about the direction of the unit vector $\hat{\pi}_i$ and

$$\xi_i = \left[\frac{(P\cdot p_i)^2}{P^2} - p_i^2\right]^{\frac{1}{2}}$$

$$= \left[\frac{\lambda(P^2,\ p_i^2,\ (P - p_i)^2)}{4P^2}\right]^{\frac{1}{2}}, \tag{IV 7}$$

where $\lambda(a, b, c)$ is defined as in (4.3).

For $n = 2$, $(P - p_i)^2 = p_2^2$, and we see that (IV 6) reduces to the corresponding formula in Ref. 5. Thus utilizing the orthogonality properties of the spherical harmonics as in (5), we see that the factor

$$2(\xi_1\xi_2)^{-\frac{1}{2}} y_{l_1 m_1}(\hat{\pi}_2) y_{l_2 m_2}(\hat{\pi}_2) \tag{IV 8}$$

leads to the proper orthonormal properties, just as the corresponding factor in (4.5). However, replacing the quantum numbers $(l',\ l'',\ m',\ m'')$ by $(l_1,\ l_2,\ m_1,\ m_2)$ leads to the commutation difficulties discussed in Sec. 3.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 5, NUMBER 7 JULY 1964

Momentum Distribution in the Ground State of the One-Dimensional System of Impenetrable Bosons*

A. LENARD
Plasma Physics Laboratory, Princeton University, Princeton, New Jersey
(Received 4 February 1964)

Girardeau has shown that an exact analytical formula may be given for the ground-state wavefunction of a system of one-dimensional impenetrable bosons. Starting with this formula, we give a mathematically rigorous analysis leading to the determination of major features of the momentum distribution in the limit of an infinitely large system.

1. INTRODUCTION

THE study of quantum mechanical many-body systems has great contemporary interest, and in recent years much knowledge has accumulated in this field. However, the results of most investigations, while plausible, are not established with standards of mathematical rigor. This state of affairs, unsatisfactory as it may be, is probably inherent in the subject. Therefore, any not completely trivial system whose reasonably detailed properties are subject to exact analysis has an intrinsic interest, even if it is not representative of any real physical situation.

The purpose of this paper is the rigorous discussion of a specific problem of this kind. We shall study the nature of the momentum distribution in the ground state of a system consisting of a great number of one-dimensional impenetrable particles satisfying Bose–Einstein statistics. In a recent paper Girardeau[1] has shown that for this system explicit analytic formulas may be given for all stationary-state wavefunctions in coordinate-space representation. The momentum distribution in any of these stationary states is defined by the well-known rules of quantum mechanical transformation theory. However, even the simplest momentum-space properties are given in terms of many-dimensional integrals of the type encountered in classical statistical mechanics, so that no effective information is immediately available. Of interest is the usual "thermodynamic limit," namely the limit in which both the number of particles and the size of the container tend to infinity, proportionally to each other. We shall show that for our system it is possible to establish a number of interesting specific facts in this limit.

The problem of the momentum distribution was already attacked by Girardeau[1] who applied an approximation method suggested by the work of Bogoliubov.[2] He found a singular behavior at the origin in momentum space reminiscent of, though not identical with, the behavior of an ideal Bose system (Bose–Einstein condensation). This claim was disproved by Schultz[3] who observed a mathematical correspondence between the Girardeau model and a certain asymptotic limit of an exactly soluble "spin system." Using this correspondence, Schultz was able to show that in the ground state of the Girardeau model no Bose–Einstein condensation exists, not even in the generalized sense proposed by Girardeau. Our work confirms this conclusion. We do not use the spin analogy of Schultz, but proceed in a straightforward manner from the ground-state wavefunction given by Girardeau and, by a step-by-step process, transform the quantities of interest into forms for which powerful techniques of analysis become applicable. It will be seen that an especially important role is played by certain results of Szegö in the theory of Toeplitz matrices.[4]

An interesting question is the relationship of the present paper to studies of more general and realistic models of interacting Bose particles, and in particular, to the three-dimensional model with hard-core repulsive interparticle forces. It has been suggested in the past that the singular behavior of Bose systems observed at low temperatures is due to Bose–Einstein condensation. Such a theory is tenable only if a strong short-range interparticle repulsion is insufficient to change the qualitative features of the momentum distribution from those of an ideal gas, though, of course, it must change them quantitatively. Whether this is or is not the case is not known with certainty, although approximate calculations suggest that it is. Simple qualitative comparison of the one- and the three-dimensional problems

* This work was supported by the U. S. Atomic Energy Commission.

[1] M. Girardeau, J. Math. Phys. 1, 516 (1960).

[2] N. N. Bogoliubov, J. Phys. USSR 11, 23 (1947).

[3] T. D. Schultz, J. Math. Phys. 4, 666 (1963).

[4] U. Grenander and G. Szegö, *Toeplitz Forms and their Applications* (University of California Press, Berkeley, California, 1958).

suggests that in the latter case the momentum distribution is "sharper" at the origin. Thus a discovery of Bose–Einstein condensation in the one-dimensional model studied here would have been taken as very strong evidence of such condensation in the three-dimensional model. Unfortunately the facts are otherwise and so our investigation throws no light on the three-dimensional problem. It only shows that very strong interparticle repulsion can (at least in one dimension) sufficiently alter the momentum distribution to remove its most characteristic feature, the singularity at the origin.

In the following section we give a precise description of the problem and define the quantities which form the subject of the investigation. In Sec. 3 we derive the representation of the one-particle density matrix as a determinant of a large Toeplitz matrix. In Sec. 4 we present the proof of an important inequality due to Szegö. In Sec. 5 we transform the determinant representing the density matrix into a form where—from the point of view of the thermodynamic limit—it appears as the well-known "discrete approximation" to the Fredholm determinant of a certain kernel given in simple explicit form. These results are used in Sec. 6 to prove theorems about the thermodynamic limit. It is shown that the limiting momentum distribution exists in a mathematically precise sense, and certain bounds for its behavior for large and small momenta are derived. In Sec. 7 we discuss the relevance of these theorems from the point of view of the criteria of Penrose and Onsager[5] as well as Girardeau[1] for the presence or absence of Bose–Einstein condensation. Finally, in Sec. 8 we compare our results with the previous work of Girardeau[1] and Schultz.[3] Some mathematical material used in the text is collected in three appendices.

2. PRELIMINARIES

The quantum mechanical problem we shall study is defined by the following conditions:

(i) The wavefunction satisfies the free-particle Schrödinger equation for the motion of N particles in one dimension. We assume $N \geq 2$.

(ii) The wavefunction is symmetrical with respect to interchange of particle coordinates (Bose–Einstein statistics).

(iii) The wavefunction satisfies periodic boundary conditions with period L.

(iv) The wavefunction vanishes whenever two particle coordinates coincide.

[5] O. Penrose and L. Onsager, Phys. Rev. **104**, 576 (1956).

Condition (iii) may be interpreted to mean that the particles are constrained to move on a circle of circumference L. Condition (iv) is the expression of the mutual impenetrability of the particles. Recently, Lieb[6] studied a more general problem where this condition was replaced by the milder one that the logarithmic derivative of the wavefunction with respect to any of the coordinates increase discontinuously by a specified amount γ when the chosen coordinate crosses another one. Our model corresponds to the limit $\gamma \to \infty$, the other limit $\gamma \to 0$ being the case of free particles.

The problem defined by the conditions (i)–(iv) has a close relationship to the problem of *free* particles satisfying the exclusion principle.[1] The stationary-state wavefunctions of the two problems differ only by a multiplicative function which assumes only the two values ± 1, and the energy eigenvalues are identical.[1] In the following we shall be concerned exclusively with the ground state. Its wavefunction is

$$\psi_{N,L}(x_1, x_2, \cdots, x_N) = (N!\, L^N)^{-\frac{1}{2}} \left| \det_{1 \leq n,m \leq N} e^{2\pi i n x_m / L} \right|. \quad (1)$$

This may also be written in the factored form

$$\psi_{N,L}(x_1, x_2, \cdots, x_N) = (N!\, L^N)^{-\frac{1}{2}} \prod_{1 \leq n < m \leq N} |e^{2\pi i x_n / L} - e^{2\pi i x_m / L}|. \quad (2)$$

It is real, positive, and translation-invariant, properties which depend only on the fact that the particles satisfy Bose–Einstein statistics, periodic boundary conditions, and that it is the wavefunction of the ground state.[7] Although Girardeau gave (1) only for odd N, it is equally correct for even N.[8]

We shall be interested in the one-particle reduced density matrix (referred to in the following simply as the density matrix) given by

$$\rho_{N,L}(x - x') = N \int_0^L dx_1 \int_0^L dx_2 \cdots \int_0^L dx_{N-1} \times \psi_{N,L}(x_1, \cdots, x_{N-1}, x) \psi^*_{N,L}(x_1, \cdots, x_{N-1}, x'). \quad (3)$$

It is so normalized that $\rho_{N,L}(0) = N/L$ is the particle number density. It is real and positive and also of positive type which means that its Fourier coefficients with respect to the variable $\xi = x - x'$ are nonnegative. We write

[6] E. H. Lieb and W. Liniger, Phys. Rev. **130**, 1605 (1963).
[7] E. H. Lieb, Phys. Rev. **130**, 2518 (1963), footnote 15.
[8] Reference 6, footnote 6.

$$\rho^{(n)}_{N,L} = \int_0^L \rho_{N,L}(\xi) e^{-2\pi i n\xi/L}\, d\xi = \frac{N}{L}\int_0^L dx_1 \cdots$$
$$\times \int_0^L dx_{N-1} \left| \int_0^L dx \psi_{N,L}(x_1, \cdots, x_{N-1}, x) e^{-2\pi i n x/L} \right|^2$$
$$(n = 0, \pm 1, \pm 2, \cdots). \quad (4)$$

The Fourier coefficients have an important physical significance. $\rho^{(n)}_{N,L}$ is the quantum mechanical mean (expectation value) of the number of particles having momentum[9]

$$k_n = 2\pi n/L. \quad (5)$$

We note that it is actually independent of L. For the sake of keeping its physical meaning clear we shall, however, not drop the subscript L.

We now lay down the following precise definition. Let

$$F_{N,L}(k) = \frac{1}{N} \sum_{-\infty < n < (kL/2\pi)} \rho^{(n)}_{N,L}. \quad (6)$$

This function will be referred to in the following as the *momentum distribution function*. It is the mean fraction of particles whose momenta are less than k, regarded as a function of k; the subscripts N and L calling attention to the total number of particles and the box size on which it depends parametrically. In terms of the momentum distribution function, we now have

$$\frac{L}{N}\rho_{N,L}(\xi) = \int_{-\infty}^{\infty} e^{ik\xi}\, dF_{N,L}(k). \quad (7)$$

The reason for introducing the definition (6) and for the somewhat fanciful way of writing what is basically the Fourier series of the density matrix is the following. The momentum distribution function belongs to a class known as *probability distribution functions*. Another class of functions is formed by the Fourier–Stieltjes transforms, in the manner of (7), of probability distribution functions, known as *characteristic functions*. There is a one-to-one correspondence between functions belonging to the two classes, and important properties of functions belonging to one class mirror themselves in corresponding properties of the other class. Of particular significance is the fact that limiting operations preserve this correspondence. We are here interested in the "thermodynamic limit"[10]

$$N \to \infty, \quad L = lN \to \infty \quad (l > 0 \text{ constant}). \quad (8)$$

While our interest lies in the momentum distribution, it turns out that it is simpler to investigate the density matrix and from its limiting properties draw the appropriate conclusions by means of general theorems. For the mathematical details on the subject of distribution functions and their characteristic functions, the reader must be referred to the literature,[11] but we have collected in Appendix 1 the statements of all needed facts in a form sufficiently complete for the task on hand.

We mention here one simple property of the momentum distribution which follows from the fact that the ground-state energy is identical to the ground-state energy of free Fermions with the same values of N and L. Since the ground-state energy is also expressible as

$$N\int_{-\infty}^{\infty} k^2\, dF_{N,L}(k) = \sum_{n=-\infty}^{\infty} \rho^{(n)}_{N,L}\left(\frac{2\pi n}{L}\right)^2, \quad (9)$$

we have for this quantity[12]

$$N\int_{-\infty}^{\infty} k^2\, dF_{N,L}(k)$$
$$= \left(\frac{2\pi}{L}\right)^2 \sum_{q=1}^{N}\left(q - \frac{N+1}{2}\right)^2 = N\frac{\pi^2}{3}\frac{N^2-1}{L^2}. \quad (10)$$

Thus, in the thermodynamic limit,[13]

$$\lim_{N,L\to\infty} \int_{-\infty}^{\infty} k^2\, dF_{N,L}(k) = \frac{\pi^2}{3l^2}. \quad (11)$$

3. DETERMINANTAL REPRESENTATION OF DENSITY MATRIX

We shall now derive an algebraic representation for the density matrix, obtained by carrying out the integrations in (3).

Since the box size L enters throughout only as a scale factor, we write

$$\rho_{N,L}(\xi) = (1/L)R_N(2\pi\xi/L). \quad (12)$$

The function $R_N(\alpha)$ will be shown to be equal the determinant of a matrix with $N-1$ rows and columns whose elements are functions of α alone.

Our starting point is the identity

$$\psi_{N,L}(x_1, \cdots, x_{N-1}, x_N)$$
$$= (NL)^{-\frac{1}{2}}\psi_{N-1,L}(x_1, \cdots, x_{N-1})$$
$$\times \prod_{1\le n\le N-1} |e^{2\pi i x_n/L} - e^{2\pi i x_N/L}|, \quad (13)$$

[9] We put $\hbar = 1$ throughout.

[10] Since we are dealing with a single quantum state for the whole system, namely the ground state, this terminology is actually a misnomer. But there is no point in avoiding it as long as its meaning is clearly understood.

[11] M. Loeve, *Probability Theory* (D. Van Nostrand Company, New York, 1955), Chap. IV.

[12] For even N the wave vectors in the fermion problem must be half-odd integral multiples of the basic unit $2\pi/L$, cf. Ref. 8.

[13] This notation will be employed for the limit (8).

obtained by comparing $\psi_{N,L}$ and $\psi_{N-1,L}$ in the form given in (2). If we then substitute (13) into (3) and make use of the determinantal form (1) for $\psi_{N-1,L}$, we obtain

$$R_N(\alpha) = [(N-1)!\,(2\pi)^{N-1}]^{-1} \times \int_0^{2\pi} d\theta_1 \cdots \int_0^{2\pi} d\theta_{N-1} \left| \det_{1\le n,m\le N-1} e^{in\theta_m} \right|^2 \times \prod_{1\le r\le N-1} |e^{i\theta_r} - e^{i\frac{1}{2}\alpha}|\,|e^{i\theta_r} - e^{-i\frac{1}{2}\alpha}|. \tag{14}$$

Now we expand the determinant in the integrand in the usual way into its $(N-1)!$ terms

$$\det_{1\le n,m\le N-1} e^{in\theta_m} = \sum_P \delta \prod_{1\le n\le N-1} e^{i\nu_n\theta_n}. \tag{15}$$

Here P is a permutation on the first $N-1$ integers, ν_n is the integer into which P sends n, δ is the parity of P, and the sum goes over all P. This results in the formula

$$R_N(\alpha) = [(N-1)!]^{-1} \sum_P \sum_{P'} \delta\delta' \times \prod_{1\le n\le N-1} \left\{ \frac{1}{2\pi} \int_0^{2\pi} d\theta\, e^{i(\nu_n - \nu_n')\theta} \times |e^{i\theta} - e^{i\frac{1}{2}\alpha}|\,|e^{i\theta} - e^{-i\frac{1}{2}\alpha}| \right\}. \tag{16}$$

Let

$$f(\theta,\alpha) = |e^{i\theta} - e^{i\frac{1}{2}\alpha}|\,|e^{i\theta} - e^{-i\frac{1}{2}\alpha}| = 2\,|\cos\theta - \cos\tfrac{1}{2}\alpha|, \tag{17}$$

and

$$c_n(\alpha) = \frac{1}{2\pi} \int_0^{2\pi} e^{-in\theta} f(\theta,\alpha). \tag{18}$$

In (16) we have then

$$R_N(\alpha) = [(N-1)!]^{-1} \times \sum_P \sum_{P'} \delta\delta' \prod_{1\le n\le N-1} c_{\nu_n - \nu_n'}(\alpha). \tag{19}$$

It is easy to see that the summation over one of the running permutation variables is trivial and merely serves to cancel the factorial in front. But the remaining sum is just the expansion of a determinant. Thus we have

Theorem 1.[14]

$$R_N(\alpha) = \det_{1\le n,m\le N-1} c_{n-m}(\alpha). \tag{20}$$

The calculations of this paper are based on this purely algebraic representation of the density matrix.

[14] This has been found independently by Professor F. J. Dyson of the Institute for Advanced Study, Princeton, New Jersey.

We evaluate the elements of the determinant explicitly. The integrals (18) are elementary, but some care has to be exercised on account of the occurrence of the absolute-value sign and the fact that the cases $n = 0$ and $n = \pm 1$ have to be treated separately. The result may be written in the following simple manner:

$$c_n(\alpha) = 2\delta_{n,0}\cos\tfrac{1}{2}\alpha - \delta_{n,1} - \delta_{n,-1} + \frac{2}{\pi}\left[\frac{\sin\frac{1}{2}(n+1)\,|\alpha|}{n+1} + \frac{\sin\frac{1}{2}(n-1)\,|\alpha|}{n-1} - \frac{2\cos\frac{1}{2}\alpha\sin\frac{1}{2}n\,|\alpha|}{n} \right]. \tag{21}$$

This is valid for

$$-\pi \le \alpha \le \pi \tag{22}$$

and all n, with the proviso that terms which appear with vanishing denominators are to be interpreted formally by continuity. For instance,

$$c_0(\alpha) = 2\cos\tfrac{1}{2}\alpha + (2/\pi)[2\sin\tfrac{1}{2}|\alpha| - |\alpha|\cos\tfrac{1}{2}\alpha], \tag{23}$$

and similarly for $n = \pm 1$.

For the special case $\alpha = 0$, we have

$$c_n(0) = \begin{cases} 2 & (n=0) \\ -1 & (n=\pm 1) \\ 0 & (\text{otherwise}). \end{cases} \tag{24}$$

The determinant (20) is easily evaluated by expanding according to the first two rows. This results in the recursion formula

$$R_N(0) = 2R_{N-1}(0) - R_{N-2}(0), \tag{25}$$

whose solution is

$$R_N(0) = N. \tag{26}$$

The significance of Theorem 1 lies in the following. We have succeeded in carrying out the integrations involved in the definition of the density matrix and reduced it to the purely algebraic representation of a determinant with elements which are simple elementary functions. That this is possible is actually not too surprising since multiple integrals of related type have been evaluated before.[15] However, our problem is as yet far from solved because we are interested in limiting behavior for large N and this

[15] See, for instance, the articles by F. J. Dyson, J. Math. Phys. 3, 140, 157, and 166 (1962). The perceptive reader will notice that our ground-state wavefunction has an interpretation in terms of Dyson's one-dimensional "Coulomb gas on a circle" which, in turn, is related to the eigenvalue distribution of his random matrices.

cannot be read off the determinantal representation any more than it can in the original representation as a multiple integral.[16] Fortunately, the actual limit of interest [cf. (8) and (12)] demands that $\alpha \to 0$ *simultaneously* with $N \to \infty$, and we shall see below that the determinant (20) is a good starting point for its calculation.

The representation given in Theorem 1 has another fortunate feature. The determinant is of a very special type. Its elements are constant along any of the lines parallel to the principal diagonal, and they are the Fourier coefficients of a real, non-negative function. Matrices having this property are called Toeplitz matrices and have a very detailed analytic theory.[4] We make use of this theory to derive an important inequality for $R_N(\alpha)$. This is done in the next section; then the thermodynamic limit is investigated.

4. AN IMPORTANT INEQUALITY

The purpose of this section is to establish the following result.

Theorem 2.

$$R_N(\alpha) < \min_{1<R<\infty} \{R^{2N}(R^2 - 1)^{-\frac{1}{2}} \times (R^4 - 2R^2 \cos\alpha + 1)^{-\frac{1}{2}}\}. \quad (27)$$

This theorem is due to Szegö.[17] Before presenting its proof we discuss the reasons for its importance in the present context.

Let us first inquire what qualitative information may be obtained from it. For $\alpha = 0$, the minimizing value of R is

$$R = [1 + 1/(N - 1)]^{\frac{1}{2}}, \quad (28)$$

and the inequality reads

$$R_N(0) < N[1 + 1/(N - 1)]^{N-1}. \quad (29)$$

In this case we know $R_N(0) = N$, so that the inequality yields no new information, but we see that it does not "deteriorate" as $N \to \infty$, since the bound overshoots the actual value by a factor no more than $e = 2.7182 \cdots$.

For $\alpha \neq 0$ the direct minimization leads to a cubic equation for R^2 which, while it can be solved, yields a result too complicated to be useful. Instead, we may proceed by a slight weakening of the inequality, replacing the factor $(R^4 - 2R^2 \cos\alpha + 1)^{-\frac{1}{2}}$ by the larger quantity $|2 \sin \frac{1}{2}\alpha|^{-\frac{1}{2}}$. Then the minimizing value of R is

$$R = [1 + 1/(2N - 1)]^{\frac{1}{2}}, \quad (30)$$

and the inequality reads

$$R_N(\alpha) < \left|\frac{N}{\sin\frac{1}{2}\alpha}\right|^{\frac{1}{2}} \left(1 + \frac{1}{2N - 1}\right)^{N-\frac{1}{2}} < \left|\frac{eN}{\sin\frac{1}{2}\alpha}\right|^{\frac{1}{2}}. \quad (31)$$

Observe the square-root dependence on N. Together with (26) this shows that, in the limit $N \to \infty$, $R_N(\alpha)$ develops a large peak in the neighborhood of the origin $\alpha = 0$. The width of this peak can, of course, not be determined from the results so far obtained, since (31) only gives an upper bound. Nevertheless, the form of this inequality strongly suggests that in the thermodynamic limit when αN remains fixed as $N \to \infty$ one remains, in effect, on top of the peak and $R_N(\alpha)$ remains of order N. To the extent that we anticipate the existence of the thermodynamic limit of $\rho_{N,L}(\xi)$ with fixed ξ, this is to be expected, but the preceding consideration does show that the inequality of Theorem 2 is in some sense "uniformly good" for all α in spite of the fast variation of $R_N(\alpha)$ near $\alpha = 0$ in the case of large N.

Again, anticipating the existence of this limit, we now write[18]

$$\lim_{N,L\to\infty} \rho_{N,L}(\xi) = \lim_{N\to\infty} \frac{1}{N} R_N\left(\frac{2\pi\xi}{N}\right) = \rho(\xi). \quad (32)$$

Applying the inequality (31) and passing to the limit we then obtain

$$\rho(\xi) \leq |e/\pi\xi|^{\frac{1}{2}}. \quad (33)$$

This inequality is useful for the following reason. In the course of proving the existence of the limit (32), an infinite series representation for the function $\rho(\xi)$ emerges [cf. (58) below]. While not a power series in ξ, nevertheless the nature of this series is such that, from its first few terms, only the behavior of $\rho(\xi)$ for small $|\xi|$ may be deduced. The inequality (33) is the only information we have on its behavior for *large* $|\xi|$.

The proof[17] of Theorem 2 is based on Szegö's investigations of Toeplitz matrices.[4] In Appendix 2 we have collected the information from this theory needed for our purpose.

[16] In fact, for some purposes the multiple integral seems to offer advantages. Dyson (in an unpublished lecture at the Eastern Theoretical Physics Conference of October 1963) has shown how a not rigorous but very suggestive argument may be based on it, indicating an asymptotic property for large N.

[17] Communicated privately to the author by Professor G. Szegö.

[18] Here and in the following we set $L = N$ which corresponds merely to a choice for the unit of length, but for conceptual clarity we occasionally refer to L in the notation.

In the notation of Toeplitz theory,

$$R_N(\alpha) = D_{N-2}(f), \tag{34}$$

where[19] the function $f(\theta)$ is given by (17). Let R be an arbitrary positive number larger than one, and let

$$f_R(\theta) = |Re^{i\theta} - e^{i\frac{1}{2}\alpha}| \cdot |Re^{i\theta} - e^{-i\frac{1}{2}\alpha}|. \tag{35}$$

Then $f(\theta) < f_R(\theta)$ and, therefore, by Corollary 1 of Szegö's Theorem A [cf. Appendix 2],

$$D_{N-2}(f) < D_{N-2}(f_R). \tag{36}$$

Corollary 2 shows that for $n = N-2, N-1, N, \cdots$

$$D_{N-2}(f_R) \leq G(f_R)^{N-1}[D_n(f_R)/G(f_R)^{n+1}], \tag{37}$$

and the right-hand side of this inequality does not decrease with n. Now, the function $f_R(\theta)$ satisfies the conditions of Theorem B and so the right-hand side of (37) has a finite limit as $n \to \infty$, and this limit provides an upper bound for $R_N(\alpha)$. It turns out, luckily, that the limit, as specified by Theorem B can be given in elementary terms. The analytic function assigned to $f_R(\theta)$ by Theorem B is simply

$$g_R(z) = (z - Re^{i\frac{1}{2}\alpha})^{\frac{1}{2}}(z - Re^{-i\frac{1}{2}\alpha})^{\frac{1}{2}}, \tag{38}$$

as one verifies by checking the relevent properties which define it uniquely. Thus

$$h_n = -(1/nR^n)\cos\tfrac{1}{2}(\alpha n), \tag{39}$$

and

$$\lim_{n\to\infty} \frac{D_n(f_R)}{G(f_R)^{n+1}} = \exp \sum_{n=1}^{\infty} n\,|h_n|^2$$

$$= \left(1 - \frac{1}{R^2}\right)^{-\frac{1}{2}}\left(1 - \frac{2}{R^2}\cos\alpha + \frac{1}{R^4}\right)^{-\frac{1}{2}}. \tag{40}$$

This, together with

$$G(f_R) = |g_R(0)|^2 = R^2, \tag{41}$$

shows that

$$D_{N-2}(f) < R^{2N}(R^2 - 1)^{-\frac{1}{2}} \times (R^4 - 2R^2\cos\alpha + 1)^{-\frac{1}{2}}. \tag{42}$$

But $R > 1$ was arbitrary and the right-hand side of (42) is a continuous function of R which tends to ∞ both as $R \to 1$ and as $R \to \infty$, so that there is a minimum on the interval $1 < R < \infty$. This completes the proof of Theorem 2.

5. THE THERMODYNAMIC LIMIT

We now transform the determinant (20) into a form where the limit (32) can be carried out.

[19] In the remainder of this section we suppress the dependence on α which is regarded as a fixed parameter. For the notation used here the reader is advised to consult Appendix 2 or Ref. 4.

We begin by writing

$$c_n(\alpha) = c_n(0) + b_n(\alpha), \tag{43}$$

where $c_n(0)$ is given by (24), and $b_n(\alpha)$ is obtained from (21),

$$b_n(\alpha) = -4\delta_{n,0}\sin^2\frac{\alpha}{4} + \frac{8}{n\pi}\sin^2\frac{\alpha}{4}\sin\frac{n}{2}|\alpha| + \frac{2}{\pi}\left[\frac{\sin\frac{1}{2}(n+1)|\alpha|}{n+1} + \frac{\sin\frac{1}{2}(n-1)|\alpha|}{n-1} - 2\frac{\sin\frac{1}{2}n|\alpha|}{n}\right]. \tag{44}$$

We now write[20]

$$R_N(\alpha) = \det c_{n-m}(\alpha) = [\det c_{n-m}(0)] \times [\det(\delta_{n,m} + \sum_r M^{(N)}_{n,r} b_{r-m}(\alpha))], \tag{45}$$

where $M^{(N)}_{n,r}$ is the $N-1$ by $N-1$ matrix inverse to $c_{n-m}(0)$. The first factor is just $R_N(0) = N$, and therefore we have[18]

$$\rho_{N,L}(\xi) = \det\left(\delta_{n,m} + \sum_r M^{(N)}_{n,r} b_{r-m}\left(\frac{2\pi\xi}{N}\right)\right). \tag{46}$$

One may verify explicitly that

$$M^{(N)}_{n,m} = \begin{cases} \dfrac{m(N-n)}{N} & (1 \leq m \leq n \leq N-1), \\ \dfrac{n(N-m)}{N} & (1 \leq n \leq m \leq N-1). \end{cases} \tag{47}$$

It will become evident in a moment that it is convenient to rewrite this as follows:

$$M^{(N)}_{n,m} = N\omega\left(\frac{n}{N}, \frac{m}{N}\right), \tag{48}$$

where

$$\omega(s,t) = \begin{cases} s(1-t) & (0 \leq s \leq t \leq 1), \\ t(1-s) & (0 \leq t \leq s \leq 1) \end{cases} \tag{49}$$

is a continuous function on the unit square $0 \leq s, t \leq 1$. Thus

$$\rho_{N,L}(\xi) = \det\left(\delta_{n,m} + N\sum_r \omega\left(\frac{n}{N}, \frac{r}{N}\right) b_{r-m}\left(\frac{2\pi\xi}{N}\right)\right). \tag{50}$$

We now inspect closely the three terms of the matrix product $N\sum_r$ in (50) obtained from the three terms of $b_n(\alpha)$ as given in (44).

[20] To keep the notation uncluttered we suppress the range of indices in the sums and determinants. It is the first $N-1$ positive integers.

The first term gives

$$-4N \sin^2 \frac{\pi\xi}{2N}\, \omega\left(\frac{n}{N}, \frac{m}{N}\right) = -\frac{1}{N}\pi^2\xi^2\omega\left(\frac{n}{N}, \frac{m}{N}\right) + O\left(\frac{1}{N^3}\right). \quad (51)$$

Here and in the following we mean by the O-symbol a quantity which is bounded by a constant times the indicated argument, where the "constant" may depend on ξ but not on n and m. This ensures that we can investigate limiting behavior as $N \to \infty$ and fixed ξ *uniformly* in n and m.

Next we consider the second term in (44). Its contribution to the matrix product in (50) is

$$\frac{8}{\pi}\sin^2 \frac{\pi\xi}{2N} \sum_r \omega\left(\frac{n}{N}, \frac{r}{N}\right) \frac{\sin \pi |\xi| (r/N - m/N)}{r/N - m/N}. \quad (52)$$

In the limit of large N, the sum over r when divided by N appears as a Riemann approximating sum to an integral and the error is easily estimated by the customary methods. Thus (52) becomes

$$\frac{1}{N} 2\pi\xi^2 \int_0^1 d\sigma\, \omega\left(\frac{n}{N}, \sigma\right) \frac{\sin \pi |\xi| (\sigma - m/N)}{\sigma - m/N} + O\left(\frac{1}{N^2}\right). \quad (53)$$

Finally, we come to the third term in (44) and its contribution to the matrix product. The computation is greatly facilitated by the observation that the square bracket in (44) contains a *second difference* with respect to the subscript n. The sum over r in (50) may then be treated by the "summation by parts" procedure, the discrete analog of integration by parts. Great simplification arises because—apart from boundary terms which, of course, have to be duly taken into account—we have to deal with the second difference of the other matrix factor, but $M^{(N)}_{n,r+1} + M^{(N)}_{n,r-1} - 2M^{(N)}_{n,r} = \delta_{n,r}$ by definition of the matrix M [cf. (24)]. A careful manipulation reduces this contribution to

$$\frac{2}{\pi N}\left[\frac{(n/N)\sin(\pi|\xi|(1 - m/N))}{1 - m/N} + \frac{(1 - n/N)\sin(\pi|\xi| m/N)}{m/N} - \frac{\sin(\pi|\xi|(n/N - m/N))}{n/N - m/N}\right]. \quad (54)$$

Summing up (51), (53), and (54), we may write then

$$\rho_{N,L}(\xi) = \det\left[\delta_{n,m} + \frac{1}{N} K\left(\frac{n}{N}, \frac{m}{N}; \xi\right) + O\left(\frac{1}{N^2}\right)\right]. \quad (55)$$

The function K is defined as follows:

$$K(s, t; \xi) = -\pi^2\xi^2\omega(s, t) + 2\pi\xi^2 \int_0^1 d\sigma \frac{\omega(s, \sigma)\sin \pi|\xi|(\sigma - t)}{\sigma - t} + \frac{2}{\pi}\left[\frac{s \sin(\pi|\xi|(1 - t))}{1 - t} + \frac{(1 - s)\sin(\pi|\xi| t)}{t} - \frac{\sin(\pi|\xi|(s - t))}{s - t}\right]. \quad (56)$$

It is a continuous function of s and t on the unit square $0 \le s, t \le 1$, and depends on ξ parametrically. Since the left-hand side of Eq. (55) is a periodic function of ξ we must, strictly speaking, note that the equation as written is valid only in the interval $-\frac{1}{2}L \le \xi \le \frac{1}{2}L$; but this limitation becomes irrelevant as $N = L \to \infty$ and so we may think of ξ as arbitrary.

Now we observe that the determinant (55) is—apart from the $O(N^{-2})$ error terms in the elements—precisely the discrete approximation to the Fredholm determinant belonging to the integral kernel $K(s, t; \xi)$.[21] Thus we have shown the following.

Theorem 3. *For all fixed ξ, the thermodynamic limit of the density matrix exists and is equal to*

$$\rho(\xi) = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} \int_0^1 ds_1 \cdots \int_0^1 ds_n \det_{1 \le i,j \le n} K(s_i, s_j; \xi) \quad (57)$$

where the kernel $K(s, t; \xi)$ is given by (56).

In the next section we shall derive a number of specific conclusions from this fact.

6. THE LIMITING MOMENTUM DISTRIBUTION

We would now like to conclude that a limiting momentum distribution $F(k)$ exists satisfying

$$\rho(\xi) = \int_{-\infty}^{\infty} e^{ik\xi}\, dF(k), \quad (58)$$

which is the analog of (7) for a finite system, and that—in some sense—

$$\lim_{N,L\to\infty} F_{N,L}(k) = F(k). \quad (59)$$

Actually there are two separate questions involved. First, is there a probability distribution function $F(k)$ such that (58) is true? Second, if so, does (59)

[21] G. Kowalewski, *Einführung in die Determinantentheorie* (Chelsea Publishing Company, New York, 1948), 3rd ed., Paragraph 115. We prove in Appendix 3 that the error terms may be ignored.

hold for this function and in what precise sense? The answer to these questions is contained in the continuity theorem for distribution functions and characteristic functions (cf. Appendix 1). It asserts that the answer to the first question is affirmative and that (59) holds in the sense of a pointwise limit *at all continuity points of* $F(k)$ under the single additional assumption that $\rho(\xi)$ is continuous at the origin.

It is not difficult to verify that this is so, but we actually prove a much stronger property of which the continuity at the origin is a trivial consequence and which gives important additional information.

Theorem 4. *$\rho(\xi)$ is an entire function of the variable $|\xi|$.*

To be quite precise, this means that there is an entire function of the complex variable $f(z)$ such that $\rho(\xi) = f(|\xi|)$. To prove this, we consider the kernel (56) as a function of z, that is to say a function $Q(s, t; z)$ which arises when $|\xi|$ is everywhere formally replaced by z. This function is manifestly an entire function. Its power-series expansion is

$$Q(s, t; z) = -\pi^2 z^2 \omega(s, t) + \sum_{n=1}^{\infty} A_n(s, t) \frac{z^{2n+1}}{(2n+1)!}, \tag{60}$$

where the coefficients $A_n(s, t)$ are certain uniformly bounded functions of s and t on the unit square $0 \leq s, t \leq 1$. It follows that each term

$$\frac{1}{n!} \int_0^1 ds_1 \cdots \int_0^1 ds_n \det_{1 \leq i,j \leq n} Q(s_i, s_j; z) \tag{61}$$

in the expansion (57) of $f(z)$ is an entire function $f_n(z)$, say, of the complex variable z. Let r be an arbitrary positive number and let $M = M(r)$ denote the maximum absolute value of $Q(s, t; z)$ for $|z| = r$ and s, t on the unit square. According to the determinant inequality of Hadamard[22] we have then

$$|f_n(z)| < n^{\frac{1}{2}n} M^n / n!. \tag{62}$$

From this and the Weierstrass "M-test"[23] it follows that the series

$$f(z) = 1 + \sum_{n=1}^{\infty} f_n(z) \tag{63}$$

converges uniformly for $|z| = r$, and this in turn implies[24] that its sum represents an analytic function of z for $|z| \leq r$. But r was arbitrary, hence $f(z)$ is an entire function. This proves Theorem 4.

The first few terms of the convergent power-series expansion guaranteed by this theorem are easily calculated. We have[25]

$$K(s, t; \xi) = -\pi^2 \xi^2 \omega(s, t) + \tfrac{2}{3}\pi^2 |\xi|^3 s(1 - s) + O(|\xi|^5), \tag{64}$$

and therefore

$$\rho(\xi) = 1 - \frac{\pi^2}{6}\xi^2 + \frac{\pi^2}{9}|\xi|^3 + \frac{\pi^4}{120}\xi^4 + O(|\xi|^5). \tag{65}$$

The calculation of higher terms is straightforward in principle but becomes prohibitive in practice on account of the rapidly increasing algebraic complexity.

Naturally, (65) implies the continuity of $\rho(\xi)$ at the origin. Therefore, $\rho(\xi)$ is a characteristic function, and the corresponding distribution function $F(k)$ is given by the inversion formula of (58), namely,

$$F(k_1) - F(k_2) = \lim_{\lambda \to \infty} \int_{-\lambda}^{\lambda} \frac{e^{-ik_2\xi} - e^{-ik_1\xi}}{2\pi i \xi} \rho(\xi)\, d\xi. \tag{66}$$

This formula holds for all continuity points k_1 and k_2 of $F(k)$ and defines it uniquely. Moreover, (59) holds in the sense of a pointwise limit, again at all continuity points of $F(k)$. This is the most that can be said to follow from the uniqueness and continuity theorems (cf. Appendix 1) and our Theorems 3 and 4.

We shall now show that more can be said upon making use of the inequality (33). Substitute this inequality on the right-hand side of (66) and make use of

$$|\sin \tfrac{1}{2}(k_1 - k_2)\xi| \leq \text{Min}\,\{1, |\tfrac{1}{2}(k_1 - k_2)\xi|\}. \tag{67}$$

Thus

$$\begin{aligned} |F(k_1) - F(k_2)| &< e^{\frac{1}{4}}\pi^{-\frac{1}{2}} \int_{-\infty}^{\infty} |\xi|^{-\frac{1}{2}} \text{Min}\,\{1, |\tfrac{1}{2}(k_1 - k_2)\xi|\}\, d\xi \\ &= 4e^{\frac{1}{4}}\pi^{-\frac{1}{2}} |k_1 - k_2|^{\frac{1}{2}}. \end{aligned} \tag{68}$$

This inequality is a Hölder condition with index $\frac{1}{2}$. It implies, in particular, the continuity of $F(k)$ as seen by letting $k_1 \to k_2$. Now, if a sequence of probability distribution functions tends to a *con-*

[22] E. T. Whittaker and C. N. Watson, *A Course of Modern Analysis* (Cambridge University Press, Cambridge, England, 1958), 4th ed., Paragraph 11.11. The statement and proof must be slightly modified to make it valid for determinants with complex elements.

[23] Reference 22, Paragraph 3.4.

[24] Reference 21, Paragraph 5.3.

[25] It will be observed that $K(s, t; \xi)$ is not a symmetrical kernel. The author has tried in vain to modify the argument of Sec. 5 so that a representation in terms of a symmetrical kernel should emerge. Such a representation is, however, implicit in the work of Schultz, Ref. 3, and is discussed in Sec. 8 of this paper.

tinuous probability distribution function, then the convergence is uniform over the whole real axis.[26] We summarize our results as follows.

Theorem 5. *In the thermodynamic limit, the momentum distribution function $F_{N,L}(k)$ converges uniformly to a continuous limiting distribution $F(k)$ which satisfies the Hölder condition (68).*

The physically interesting information is the continuity of the limiting distribution. Its proof depends crucially on the Szegö inequality, Theorem 2, of which it forms the principal application.

The discontinuities of $F_{N,L}(k)$ which have the physical interpretation of the mean fractional occupation numbers of the various momentum states shrink to zero in the thermodynamic limit. How fast? Our next theorem gives an estimate.

Theorem 6. *The discontinuities of $F_{N,L}(k)$ are less than*

$$N^{-\frac{1}{2}}\left(1 + \frac{1}{2N-1}\right)^{N-\frac{1}{2}} \frac{1}{\pi} B(\tfrac{1}{4}, \tfrac{1}{2}). \tag{69}$$

Proof: Because of the positiveness of $\rho_{N,L}(\xi)$ the largest discontinuity of $F_{N,L}(k)$ occurs at the origin [cf. (6) and (4)], and that is

$$\frac{\rho^{(0)}_{N,L}}{N} = \frac{1}{2\pi N} \int_{-\pi}^{\pi} R_N(\alpha)\, d\alpha. \tag{70}$$

Substitution of the inequality (31) and evaluation of the integral proves the theorem. The numerical coefficient $\pi^{-1}B(\frac{1}{4}, \frac{1}{2})$ is less than 2.

Finally, we give a result on the behavior of the limiting momentum distribution for $|k| \to \infty$. In effect, it gives a bound to the rate at which $F(k)$ approaches its asymptotic values.

Theorem 7. *No absolute moments of $F(k)$ of order $\delta \geq 3$ exist. The second moment is finite and equal to $\frac{1}{3}\pi^2$.*

The proof depends on the differentiability theorem for characteristic functions (cf. Appendix 1). For, if any absolute moment of order $\delta \geq 3$ would exist (i.e., would be finite), then $\rho(\xi)$ would be three times continuously differentiable; but it is not, on account of the appearance of the $|\xi|^3$ term in (65). It is, however, twice differentiable, and

$$\int_{-\infty}^{\infty} k^2\, dF(k) = -\rho''(0) = \tfrac{1}{3}\pi^2. \tag{71}$$

This quantity is just the mean kinetic energy per particle in the thermodynamic limit [cf. also (11)].

[26] A proof of this fact is given in Appendix 1.

7. BOSE-EINSTEIN CONDENSATION

The phenomenon of Bose–Einstein condensation (or condensation in momentum space) for an *ideal* gas of Bose particles was discovered long ago,[27] but it was only relatively recently that proposals were put forward for mathematically precise criteria to distinguish the occurrence from the nonoccurrence of such condensation in the physically more interesting but much more difficult problem of a Bose gas with interparticle forces. We shall consider the application of two such criteria to the present one-dimensional model.

Penrose and Onsager[28] formulate the criterion in terms of the largest eigenvalue of the density matrix. According to this criterion, there is no condensation when the largest eigenvalue divided by the total number of particles tends to zero in the thermodynamic limit. Since

$$\int_0^L \rho_{N,L}(x - x')e^{2\pi i n x'/L}\, dx' = \rho^{(n)}_{N,L} e^{2\pi i n x/L}, \tag{72}$$

the eigenvalues in question are just the quantities $\rho^{(n)}_{N,L}$ ($n = 0, \pm 1, \pm 2, \cdots$) and the largest of them is $\rho^{(0)}_{N,L}$. This divided by N is just the largest discontinuity of the momentum distribution function. Thus we see that *Theorem 6 signifies the fulfillment of the Penrose–Onsager criterion for the absence of Bose–Einstein condensation.* It actually does more: it provides the specific bound $O(N^{-\frac{1}{2}})$ for the mean fraction of zero-momentum particles in the limit $N = L \to \infty$.

In his paper on the one-dimensional gas of Bose particles, Girardeau[1] suggested a more stringent criterion. This may be formulated in our notation as follows[29]:

$$\lim_{k \to +0} \limsup_{N,L \to \infty} [F_{N,L}(k) - F_{N,L}(-k)] = 0. \tag{73}$$

It is fulfilled because of the existence and continuity of the limiting momentum distribution (Theorem 5). The essence of Girardeau's criterion lies in the order of the two limiting processes. If the order is reversed one obtains the Penrose–Onsager criterion. The latter is always fulfilled when the Girardeau criterion is, because

$$\lim_{k \to +0} \limsup_{N,L \to \infty} [F_{N,L}(k) - F_{N,L}(-k)]$$

$$\geq \lim_{k \to +0} \limsup_{N,L \to \infty} [\rho^{(0)}_{N,L}/N] = \limsup_{N,L \to \infty} [\rho^{(0)}_{N,L}/N]. \tag{74}$$

[27] A. Einstein, Sitzber. Preuss. Akad. Wiss. Physik-Math. Kl., 261 (1924); 3, 18 (1925).

[28] O. Penrose and L. Onsager, Phys. Rev. 104, 576 (1956).

[29] Where we write "lim sup" Girardeau has "lim". This change was made to render the criterion logically independent of the existence of a limiting momentum distribution.

The implication does not hold the other way, however, as can be demonstrated by counter examples.

The $O(N^{-\frac{1}{2}})$ bound of Theorem 6 and the $O(|k_1 - k_2|^{\frac{1}{2}})$ bound of Theorem 5 are not unconnected. Both are rooted in the Szegö inequality (31) or ultimately (27). It is possible to write a single inequality which displays this connection. It is derived by writing the difference $F_{N,L}(k_1) - F_{N,L}(k_2)$ in the form of a finite sum from (6) and replacing the $\rho^{(n)}_{N,L}$ by their definition (4) as Fourier coefficients. This gives

$$F_{N,L}(k_1) - F_{N,L}(k_2) = \frac{1}{2\pi N}\int_0^{2\pi} R_N(\alpha) \times e^{-i(\alpha/2)(n_1+n_2+1)}\frac{\sin\frac{1}{2}\alpha(n_1-n_2)}{\sin\frac{1}{2}\alpha}\,d\alpha, \quad (75)$$

where n_1 is the largest and n_2 the smallest integer satisfying

$$k_2L/2\pi \leq n < k_1L/2\pi. \quad (76)$$

Making use of the inequalities (31) and (67), one obtains then

$$F_{N,L}(k_1) - F_{N,L}(k_2) < \text{const}\,[(L/N)(k_1 - k_2)]^{\frac{1}{2}}. \quad (77)$$

This inequality implies both the Hölder condition (68) of Theorem 5, and Theorem 6; the former by letting $N = L \to \infty$, the latter by putting $L(k_1 - k_2)$ equal to some positive number less than 2π.

It should be stressed that our conclusions must not be interpreted to mean that the momentum distribution shows no distinguished features at the origin in the thermodynamic limit. It is characteristic of our results that they provide bounds; bounds which are strong enough to satisfy physically motivated criteria for the absence of Bose–Einstein condensation, but not precise enough to reveal the qualitative properties of the momentum distribution. It seems quite likely that $F(k)$ is actually a quite smooth function of k (perhaps even analytic) outside the origin, and that near the origin its behavior is

$$F(k) = \tfrac{1}{2} + \text{const}\,\frac{k}{|k|^{\frac{1}{2}}} + o(|k|^{\frac{1}{2}}) \quad (k \to 0). \quad (78)$$

Whether there really is a "singularity" of this, or perhaps some even milder, type at the origin in momentum space remains an unsolved problem.[30]

8. COMPARISON WITH PREVIOUS WORK

The problem of the momentum distribution has been already considered in the past.[1,3] Here we shall consider some aspects of this previous work in the light of our own results.

Girardeau treats the problem of the momentum distribution with an approximation method due to Bogoliubov.[2] This method is not well adapted to the investigation of the question of Bose–Einstein condensation because in a sense it assumes an affirmative answer. This is so because the zero momentum state is distinguished at the outset by the assumption that it has a very much larger average occupation than the other momentum states. In addition to this difficulty of principle, Girardeau was forced to make certain approximations in his calculation whose validity is open to question. His conclusion was that the Penrose–Onsager criterion for the absence of Bose–Einstein condensation is satisfied, but not the more strict criterion (73), indeed this was his reason for proposing the latter in the first place. The suggested dependence of the average zero-momentum occupation number on the total number of particles[31]

$$\rho^{(0)}_{N,L} \sim N/\ln N \quad (N \to \infty) \quad (79)$$

is clearly contradicted by our Theorem 6. The incorrectness of Girardeau's result was already proved by Schultz.[3]

Schultz started by posing the quantum mechanical problem in the formalism of second quantized fields. Then he replaced the continuum of space points in the box $0 \leq x \leq L$ by a large but finite number of equidistant points in this interval so that the commutation rules became a finite algebraic system. He then made the crucial observation that this "discretized" problem is from the mathematical point of view completely identical to a certain problem of interacting spins, a problem whose solution had already been worked out.[32] In particular, the density matrix, or what is the same, the function $R_N(\alpha)$ was given in the following form[33]:

$$\frac{|\alpha|}{\pi}R_N(\alpha) = \lim_{\nu\to\infty}\nu \det_{1\leq q,p\leq\nu}\left[-\delta_{q,p+1} + \frac{1}{\nu}G_N\left(\frac{p+1-q}{N};\alpha\right)\right], \quad (80)$$

where

$$G_N(s;\alpha) = \frac{|\alpha|}{\pi}\frac{\sin\frac{1}{2}(N\alpha s)}{\sin\frac{1}{2}(\alpha s)}. \quad (81)$$

The essence of the method is the extra limiting

[30] The difficulty is that we have not been able to do better than the inequality (33) as regards the $|\xi| \to \infty$ behavior of $\rho(\xi)$. A sufficient sharpening of it would allow the derivation of (78).

[31] Reference 1, Eq. (A50).

[32] E. H. Lieb, T. Schultz, and D. Mathis, Ann. Phys. (N. Y.) **16**, 407 (1961).

[33] Reference 3, Eqs. (2), (11), (14), and (15).

process, visible in (80), which corresponds to going from the discrete to the continuum. It must be carried out independently of, and prior to, the thermodynamic limit. Leaving ν fixed and finite in (80) one gets an approximation to $R_N(\alpha)$ equal to replacing the integrals in (14) by their Riemann approximating sums with a subdivision of the interval $(0, 2\pi)$ into ν equal parts.[34]

By analyzing the nature of the ν by ν matrix which occurs in (80), Schultz was able to derive an inequality which corresponds to our (31) but is somewhat weaker. With its help he showed that (73) is fulfilled, thus controverting the claim of Girardeau.

Regarding the determinant in (80) we observe the following. It is precisely equal to the algebraic complement of the element in the lower left-hand corner ($q = 0$, $p = \nu$) of the larger determinant with $\nu + 1$ rows and colums,

$$\det_{0\leq q,p\leq \nu}\left[\delta_{q,p} - \frac{1}{\nu} G_N\left(\frac{q-p}{\nu}\,;\alpha\right)\right]. \tag{82}$$

From the point of view of the limiting process $\nu \to \infty$, this is of the Fredholm type; and not only its limiting behavior but also that of its *minors* can be found exactly.[35] One obtains[36]

$$\frac{|\alpha|}{\pi} R_N(\alpha) = G_N\binom{0}{1}\,;\alpha\Big) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \int_0^1 ds_1 \cdots \times \int_0^1 ds_n G_N\binom{0\ s_1 \cdots s_n}{1\ s_1 \cdots s_n}\,;\alpha\Big), \tag{83}$$

where

$$G_N\binom{s_0 s_1 \cdots s_n}{t_0 t_1 \cdots t_n}\,;\alpha\Big) = \det_{0\leq i,j\leq n} G_N(s_i - t_j;\alpha) \tag{84}$$

is the Fredholm symbol belonging to the kernel G_N defined by (81). We shall now show that the series in (83) is actually finite because

$$G_N\binom{s_0 s_1 \cdots s_n}{t_0 t_1 \cdots t_n}\,;\alpha\Big) = 0 \tag{85}$$

identically in the variables whenever $n \geq N$. This fact has its origin in the identity

$$\frac{\sin \frac{1}{2}(N\alpha s)}{\sin \frac{1}{2}(\alpha s)} = \sum_r e^{ir\alpha s}, \tag{86}$$

the summation ranging over $r = -\frac{1}{2}(N-1), \cdots, \frac{1}{2}(N-3), \frac{1}{2}(N-1)$. Thus the matrix of $n+1$ rows and columns whose determinant is (84) may be regarded as a matrix product of two rectangular but in general not square matrices, one having $n+1$ rows and N columns, and the second one N rows and $n+1$ columns. It is a general fact of the algebra of determinants[37] that the determinant of such a product vanishes when $N < n+1$.

It is rather remarkable that the expression given by (83) for $R_N(\alpha)$ represents the same function as the determinant (20) derived in the present paper, different though—to all appearances—it may seem. An explanation lies, perhaps, in the circumstance that the *elements* of a determinant may be altered in manifold ways without changing its *value*. Nevertheless, it would be interesting to see a short proof of the identity of these two expressions. We may remark in this connection that their general character *is* the same, namely a finite combination of trigonometric functions of $|\alpha|$ and powers of $|\alpha|$. This is obvious[38] in the case of (20), and in the case of (83) it follows from the identity (86), together with the fact that integration of functions belonging to the class indicated always produces functions again belonging to that class. A simple proof of the identity would seem to require an ingenious exercise in formal manipulation.

Our own representation of $R_N(\alpha)$ seems simpler because no integrations occur in it, but it is conceivable that (83) is more useful for certain purposes. For one, it is easier to carry out for it the thermodynamic limit (32). Instead of the elaborate calculations of Sec. 5, we now only have to notice that

$$\lim_{N\to\infty} G_N\Big(s;\frac{2\pi\xi}{N}\Big) = \frac{2\sin \pi s\,|\xi|}{\pi s} = g(s;\xi), \tag{87}$$

and in (83) we can go to the limit term by term.[39] This gives

$$2\,|\xi|\,\rho(\xi) = g\binom{0}{1}\,;\xi\Big) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \times \int_0^1 ds_1 \cdots \int_0^1 ds_n\, g\binom{0\ s_1 \cdots s_n}{1\ s_1 \cdots s_n}\,;\xi\Big), \tag{88}$$

with

$$g\binom{s_0 s_1 \cdots s_n}{t_0 t_1 \cdots t_n}\,;\xi\Big) = \det_{0\leq i,j\leq n} g(s_i - t_j;\xi). \tag{89}$$

Again, (88) looks very different from the representation given in Theorem 3, but it is easily verified, for instance, that the behavior for small $|\xi|$ is given

[34] This was pointed out to the author in a correspondence by Dr. Schultz.
[35] Reference 21, Paragraph 117.
[36] The possibility of the transformation from (80) to (83) seems not to have been noticed by Schultz.
[37] Reference 21, Sec. 34.
[38] Powers of $|\alpha|$ arise from $c_n(\alpha)$ with $n = 0, \pm 1$.
[39] The rigorous justification of the legitimacy of this step should not be difficult.

correctly by (65). Whether any properties of $\rho(\xi)$ can be better revealed in the representation (88) is not known.

ACKNOWLEDGMENT

The author wishes to express his indebtedness to Professor C. N. Yang whose remarks on the problem of Bose–Einstein condensation at the spring meeting (1963) on statistical mechanics at Yeshiva University induced him to undertake this investigation. He wishes to thank Professor F. J. Dyson for several fruitful conversations and much encouragement. And, above all, he is grateful to Professor G. Szegö for the correspondence which resulted in Theorem 2, and for his consent to let his contribution become an essential part of this paper.

APPENDICES

A1. Distribution Functions and Characteristic Functions

A (cumulative) probability distribution function $F(k)$ is a real nondecreasing function such that its total variation is from 0 to 1. Such a function has at most a denumerable number of ordinary jump discontinuities and is continuous at all other points. Its value at the discontinuity points may be conventionally fixed by continuity from the left, for example.

The characteristic function $\phi(\xi)$ belonging to the probability distribution function $F(k)$ is defined by the Fourier–Stieltjes transformation

$$\phi(\xi) = \int_{-\infty}^{\infty} e^{ik\xi}\, dF(k). \tag{90}$$

This relationship is unique, i.e., two probability distribution functions to which belongs the same characteristic function can differ only (inessentially) at discontinuity points. There is an explicit inversion formula

$$F(k_1) - F(k_2) = \lim_{\lambda\to\infty} \int_{-\lambda}^{\lambda} \frac{e^{-i\xi k_2} - e^{-i\xi k_1}}{2\pi i\xi}\, \phi(\xi)\, d\xi, \tag{91}$$

the implication being that the limit exists and is equal to the difference on the left, provided k_1 and k_2 are continuity points. (Uniqueness Theorem.)

One reason for the importance of this one-to-one relationship between these two classes of functions is that limiting operations preserve this correspondence. More specifically, let $\phi_n(\xi)$ be a sequence of characteristic functions belonging to the probability distribution functions $F_n(k)$, such that

$$\lim_{n\to\infty} \phi_n(\xi) = \phi(\xi) \tag{92}$$

exits (pointwise) and is continuous at $\xi = 0$. Then there is a probability distribution function $F(k)$ such that (i) its characteristic function is $\phi(\xi)$, and (ii) the limiting relation

$$\lim_{n\to\infty} F_n(k) = F(k) \tag{93}$$

holds at every continuity point of it. (Continuity Theorem.)

The mode of convergence in (93) is considerably sharper when $F(k)$ is continuous. In this case the convergence is *uniform for all* k. To prove this, let $\epsilon > 0$ be arbitrary and let k_i $(i = 1, 2, \cdots)$ be a finite number of points at which $F(k)$ assumes values which are integral multiples of $\frac{1}{2}\epsilon$. Let then $N = N(\epsilon)$ be so large that all differences $F_n(k_i) - F(k_i)$ are less in magnitude than $\frac{1}{2}\epsilon$ for all $n \geq N$; this can be done because i runs through a finite number of values only. If now k is arbitrary, it is easy to see that from the nondecreasing nature of the distribution functions it follows that $F_n(k)$ and $F(k)$ differ by less than ϵ for all $n \geq N$.

There is a simple relationship between the moments of the probability distribution function $F(k)$ and the derivatives of $\phi(x)$ at the origin. If the absolute moment

$$\int_{-\infty}^{\infty} |k|^{\delta}\, dF(k) \tag{94}$$

exists (i.e., is finite) then the characteristic function corresponding to $F(k)$ has everywhere continuous derivatives of all orders not exceeding δ. If δ is an even integer the value of the integral (94) is obtained from (90) by formal differentiation. (Differentiability Theorem.)

We have quoted these facts only to the extent needed in the text (Sec. 6). For more complete statements and proofs the reader is referred to the literature.[11]

A2. Szegö's Theory of Toeplitz Determinants[4]

A Toeplitz matrix is a finite matrix whose elements are given by

$$c_{n-m} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) e^{-i(n-m)\theta}\, d\theta \qquad (0 \leq n, m \leq N), \tag{95}$$

where $f(\theta)$ is a real, nonnegative function. Its determinant is denoted by $D_N(f)$. An important role

is played by the geometric mean

$$\exp \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln f(\theta)\, d\theta, \tag{96}$$

which is denoted by $G(f)$.

Szegö has shown that the ratio of two succeeding Toeplitz determinants may be characterized as the solution of a certain minimum problem.

Theorem A. *The ratio $D_N(f)/D_{N-1}(f)$ is equal to the minimum with respect to the N complex numbers $u_1, u_2, \cdots, u_N$ of the integral*

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) \times |1 + u_1 e^{i\theta} + u_2 e^{2i\theta} + \cdots + u_N e^{Ni\theta}|^2\, d\theta. \tag{97}$$

This theorem is proved[40] by consideration of certain orthogonal polynomials associated with the function $f(\theta)$. It has two important consequences.

Corollary 1. *If $\tilde{f}(\theta) > f(\theta)$ then $D_N(\tilde{f}) > D_N(f)$.*

Corollary 2. *The sequence $D_N(f)/G(f)^{N+1}$ ($N = 0, 1, 2, \cdots$) is nondecreasing.*

The first follows from Theorem A by the general principle that if one real function is always larger than another then the minimum of the first one is larger than that of the second. Corollary 2 is shown as follows. Write the minimizing trigonometric polynomial of (97) in the factored form

$$1 + u_1 e^{i\theta} + \cdots + u_N e^{Ni\theta} = (1 - A_1 e^{i\theta}) \cdots (1 - A_N e^{i\theta}), \tag{98}$$

where the A's are certain complex constants depending on f and N. Now, the integral (97) is an arithmetical mean of a function of θ and therefore is never less[41] than the geometrical mean of the same function. On the other hand, the geometrical mean of a product is the product of the geometrical means. Now, for each of the factors in (98) one shows by an elementary integration that

$$G(|1 - Ae^{i\theta}|^2) = \text{Max}\,\{1, |A|^2\} \geq 1, \tag{99}$$

and hence

$$D_N(f)/D_{N-1}(f) \geq G(f), \tag{100}$$

from which the assertion follows.

In view of Corollary 2 there are two posibilities. As $N \to \infty$, either the sequence tends to infinity or else it has a finite limit. Both possibilities actually occur,[42] but for a wide class of functions $f(\theta)$the limit is finite. This is the content of the second of Szegö's theorems which we make use of.

Theorem B. *Let $f(\theta)$ be strictly positive and let its derivative $f'(\theta)$ satisfy a Hölder condition with some index less than 1. Then the sequence $D_N(f)/G(f)^{N+1}$ tends to the finite limit*

$$\exp \sum_{n=1}^{\infty} n\, |h_n|^2. \tag{101}$$

The numbers h_n are coefficients in a power-series expansion

$$\ln g(z) = \sum_{n=0}^{\infty} h_n z^n, \tag{102}$$

$g(z)$ being defined uniquely (up to an irrelevant constant phase factor) by the properties (i) $g(z)$ is analytic, regular and not zero in the circle $|z| < 1$, and (ii) the squared absolute value of $g(re^{i\theta})$ tends to $f(\theta)$ as $r \to 1$.

Szegö's proof[43] is a subtle construction of analysis based on approximation of functions by trigonometric polynomials. An alternative proof based on different ideas has been given by Kac.[44]

We note that in the application contemplated (Sec. 4) it was impossible to apply Theorem B directly to the Toeplitz determinant $R_N(\alpha)$ because the $f(\theta)$ belonging to it is $2\,|\cos\theta - \cos\frac{1}{2}\alpha|$, and this does not satisfy the requisite conditions.

A3. A Theorem on Determinants

In order to complete the proof of Theorem 3 it is necessary to show that in the limit $N \to \infty$ the "error terms" of (55) may be ignored. We do this by proving a proposition which shows that a sufficient condition for this is that they be $o(N^{-\frac{1}{2}})$.

Let A_N and B_N be given N by N matrices for $N = 1, 2, 3, \cdots$ such that the elements of A_N are uniformly $O(N^{-1})$ and the elements of B_N are uniformly $o(N^{-\frac{1}{2}})$ as $N \to \infty$. Let I_N be the N by N unit matrix. Then the difference between the determinant of $I_N + A_N + B_N$ and the determinant of $I_N + A_N$ tends to zero as $N \to \infty$.

We prove this by making use of the celebrated inequality of Hadamard,[22] according to which the magnitude of a determinant is not more than the product of the Euclidean norms of its row vectors. The difference of our two determinants may be written with the help of the mean-value theorem

[40] Reference 4, p. 38.

[41] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities* (Cambridge University Press, Cambridge, England, 1934), Theorem 184.

[42] An example of the first possibility is furnished by our $R_N(0) = D_{N-2}(f) = N$ with $f(\theta) = 2(1 - \cos\theta)$.

[43] Reference 4, p. 76.

[44] M. Kac, Duke Math. J. 21, 501 (1954).

of differential calculus as follows:

$$\Delta_N = [(d/d\theta) \det (I_N + A_N + \theta B_N)]_{\theta=\theta_N}, \quad (103)$$

where $0 \leq \theta_N \leq 1$. According to the rule of differentiating determinants, this is

$$\Delta_N = \sum_{1\leq n,m\leq N} \sum (B_N)_{n,m}(-1)^{n+m} \det Q_N^{n,m}, \quad (104)$$

where $Q_N^{n,m}$ is that $N-1$ by $N-1$ matrix which arises from the matrix $I_N + A_N + \theta_N B_N$ by striking the nth row and the mth column. We consider first the N terms $n = m$ in the sum (104). All rows of $Q_N^{n,n}$ consist of one "large" element $1 + O(N^{-1})$ and $N-2$ "small" elements $O(N^{-1})$. The Euclidean norms of these rows are therefore

$$\left\{\left[1 + O\left(\frac{1}{N}\right)\right]^2 + (N-2)\left[O\left(\frac{1}{N}\right)\right]^2\right\}^{\frac{1}{2}} = 1 + O\left(\frac{1}{N}\right). \quad (105)$$

Hence

$$|\det Q_N^{n,n}| = \left[1 + O\left(\frac{1}{N}\right)\right]^{N-1} = O(1), \quad (106)$$

and the contribution of N such terms to Δ_N is $No(N^{-\frac{1}{2}})O(1) = o(N^{-\frac{1}{2}})$. On the other hand, consider the terms in the sum $n \neq m$. The matrix $Q_N^{n,m}$ now has one row, namely the mth, with no large element, so that the norm of this row is

$$\left\{(N-1)\left[O\left(\frac{1}{N}\right)\right]^2\right\}^{\frac{1}{2}} = O\left(\frac{1}{N^{\frac{1}{2}}}\right), \quad (107)$$

while its other rows have norms (105). Thus, for $n \neq m$,

$$|\det Q_N^{n,m}| = O\left(\frac{1}{N^{\frac{1}{2}}}\right)\left[1 + O\left(\frac{1}{N}\right)\right]^{N-2} = O\left(\frac{1}{N^{\frac{1}{2}}}\right). \quad (108)$$

There are $N^2 - N$ such terms in the sum (104), so their total contribution is $(N^2 - N)o(N^{-\frac{1}{2}})O(N^{-\frac{1}{2}}) = o(1)$. Thus $\lim \Delta_N = 0$, as was to be shown.

Correlation Functions and the Critical Region of Simple Fluids

MICHAEL E. FISHER*
The Rockefeller Institute, New York, New York
(Received 4 February 1964)

The "classical" (e.g. van der Waals) theories of the gas–liquid critical point are reviewed briefly and the predictions concerning the nature of the singularities of the coexistence curve, the specific heat, and the compressibilities are compared critically with experiment and with the analytical and numerical results for lattice gas models.

The critical singularities are related to the behavior of the pair correlation function $G(r) = g(r) - 1$ and the Ornstein–Zernike theory of critical scattering is reviewed. Alternative derivations of the theory are discussed and its validity is assessed in relation to experiment and to more detailed theoretical calculations. The nature and magnitude of the expected deviations from the "classical" theory are described. The analogies with critical magnetic phenomena are mentioned briefly.

THIS is a review article about theories of the critical point and their experimental and theoretical validity. Recent work has revealed the shortcomings of the well-known approximate theories and it is hoped that this review, although in the main nonmathematical in character, will provide some stimulus to further theoretical (and experimental) work on this old, but still imperfectly understood problem.

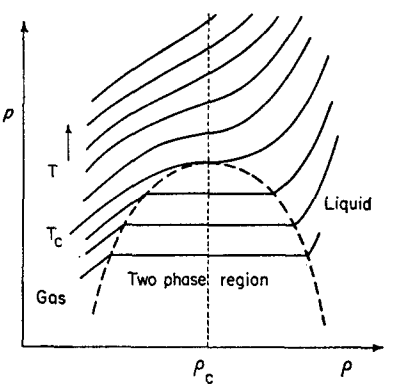


FIG. 1. Schematic isotherms for a simple fluid in the critical region.

1. THE CLASSICAL THEORY OF THE CRITICAL POINT

At temperatures below its critical temperature T_c a gas can be condensed by isothermal compression. At temperatures above T_c the transition from dense gas to liquid takes place without discontinuity of density or, as far as has been determined experimentally, without any higher-order singularities in the density or other variables. As the temperature increases to T_c the difference between the densities ρ_L and ρ_G of coexisting liquid and gas tends continuously to zero.[1] The limiting density ρ_c, and corresponding pressure p_c, define the *critical point* (see Fig. 1). In this article we shall be concerned with the properties of a simple fluid in the region of its critical point and in particular on the critical isochore $\rho = \rho_c$.

In many respects the behavior of binary fluid mixtures which undergo phase separation is closely analogous to the condensation of simple fluids and most of our remarks can be translated directly into such terms.[2] For simplicity, however, we refer in the main only to single-component systems.

As is well known, a qualitative account of condensation phenomena and the critical point is given by the classical equation of van der Waals

$$p/kT = \rho/(1 - b\rho) - a\rho^2/kT, \qquad (1.1)$$

provided this is supplemented by the "equal area rule" of Maxwell which ensures that the density ρ is a single-valued function of the pressure p.[3] The general appearance of the isotherms of a van der Waals gas in the critical region is similar to that of a real gas as shown in Fig. 1. In particular the p, ρ isotherms become flatter and flatter as T approaches T_c from above at the critical density and correspondingly the isothermal compressibility

$$K_T = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial p}\right)_T \qquad (1.2)$$

* On leave of absence from The Wheatstone Physics Laboratory, King's College, London, England.

[1] It now seems quite well established experimentally that for simple fluids, such as the noble gases, the coexistence curve does *not* have a significant "flat top." For a discussion of this suggestion see O. K. Rice, J. Phys. & Colloid Chem. 54, 1293 (1950); and Ref. 60 below. See also D. R. Thompson and O. K. Rice, quoted in Ref. 11 below.

[2] J. S. Rowlinson, *Liquids and Liquid Mixtures* (Butterworths Scientific Publications, Ltd., London, 1959), Chap. 5.

[3] Maxwell's thermodynamic derivation of the rule is unsatisfactory in that it is necessary to give thermodynamic significance to "unstable" (and "metastable") states on the original van der Waals isotherm. A formally equivalent but theoretically somewhat more convincing argument is to use the minimal properties of the Gibbs free energy (or chemical potential) to eliminate the unwanted parts of the van der Waals isotherm.

which is essentially the reciprocal of this slope, diverges to infinity at the critical point.

The derivation of the van der Waals equation given by Ornstein[4] suggests that it should be a reasonably good approximation when the pair interaction potential $\phi(r)$ of the molecules of the fluid has a short-range strongly repulsive core and a very long-range weakly attractive tail. Indeed Kac, Uhlenbeck, and Hemmer[5] have recently shown that the van der Waals isotherm (*with* flat part) follows rigorously for a one-dimensional gas of hard rods interacting with an attractive exponential potential *in the limit* that the exponential becomes infinitely long-ranged *and* infinitely weak [holding $\int_0^\infty \phi(r)\,dr$ constant]. It seems probable, however, that the behavior of a real gas *in its critical region* is crucially dependent on the finite or relatively *short* range of the attractive parts of more realistic potentials. This is supported by the comparison with real systems and exactly soluble models we present in the next section.

Three principal predictions concerning the critical region which follow from the van der Waals equation are:

(a) that the *coexistence curve* follows a square-root law, i.e., the difference between liquid and gaseous densities vanishes as

$$\rho_L - \rho_G \approx A(T_c - T)^{\frac{1}{2}} \qquad (T \to T_c-); \tag{1.3}$$

(b) that the *compressibility* along the critical isochore diverges as a simple pole,

$$K_T \approx B/|T - T_c| \qquad (\rho = \rho_c, T \to T_c+); \tag{1.4}$$

and

(c) that the *specific heat* (at constant volume) along the critical isochore rises to a maximum and then falls discontinuously as T increases through T_c, i.e.,

$$C_V(T) \approx C_c^{\pm} - D^{\pm}\,|T - T_c|, \qquad T \gtrless T_c, \tag{1.5}$$

with $C_c^- - C_c^+ = \Delta C > 0$.

The compressibility of the gas and of the liquid along the coexistence curve (i.e. at condensation) also diverges as a simple pole as $T \to T_c-$ according to the van der Waals equation but the amplitude corresponding to B in (1.4) is smaller. (The constants A, B, $C_c^{\pm}$ and $D^{\pm}$ can of course be written explicitly in terms of the van der Waals parameters a and b.)

It is important to note that these predictions are not peculiar to the van der Waals equation but follow from almost *all* approximate equations of state. Indeed they are essentially a direct consequence of the implicit or explicit assumption that the free energy and the pressure can be expanded in a Taylor series in density and temperature *at* the critical point: in other words that the critical point is not a singular point of the free energy expressed as a function of ρ and T (except in as far as Maxwell's rule is utilized *below* $T = T_c$).[6]

To demonstrate this put

$$\Delta p = p - p_c, \quad \Delta\rho = \rho - \rho_c, \quad \Delta T = T - T_c \tag{1.6}$$

and assume that

$$\Delta p = a(T) + b(T)\Delta\rho + c(T)\Delta\rho^{\,2} + d(T)\Delta\rho^3 + \cdots. \tag{1.7}$$

By definition $a(T_c) = 0$ so we assume similarly that

$$a(T) = a_1\Delta T + a_2\Delta T^2 + \cdots. \tag{1.8}$$

Since the compressibility is infinite at the critical point, $b(T_c)$ must vanish so again we assume

$$b(T) = b_1\Delta T + b_2\Delta T^2 + \cdots. \tag{1.9}$$

Finally since $b(T_c) = 0$ and the pressure above T_c must be a monotonic increasing function of ρ we have $C(T_c) = 0$ and, presumably, $C(T)$ is small in the critical region. We thus obtain for small $\Delta\rho$ and ΔT the isotherms

$$\Delta p = a_1\Delta T + b_1\Delta T\Delta\rho + d_0\Delta\rho^3 + \cdots. \tag{1.10}$$

Near the critical point the compressibility is hence given by

$$K_T \simeq \frac{(b_1\rho c)^{-1}}{\Delta T + (3d_0/b_1)\Delta\rho^2}, \tag{1.11}$$

from which the prediction (b) follows. Application of the equal area rule to the isotherms (1.10) with negative ΔT yields the coexistence curve

$$\Delta\rho^2 \simeq (b_1/d_0)\,|\Delta T| \qquad (T \le T_c), \tag{1.12}$$

which implies the square-root law (a). The divergence

[4] L. S. Ornstein, Dissertation, Leiden 1908; see also Ref. 5.

[5] M. Kac, G. E. Uhlenbeck, and P. C. Hemmer, J. Math. Phys. 4, 216 (1963). It should be noted that the correction to ideal gas behavior, arising from the hard core and represented by the parameter b, is exact in one dimension but only approximate in two or three dimensions. On the other hand the accuracy of the correction represented by the parameter a depends mainly on the long-range nature of the attractive tail of the potential and not so directly on dimensionality. Consequently, although the behavior of a three-dimensional model in the corresponding long-range limit should be van der Waals-*like* [in that Eqs. (1.3) to (1.5) should hold], one should still not expect van der Waals' equation (1.1) to hold precisely.

[6] See, for example, Landau's theory of the critical point and second-order phase transitions [L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon Press, Ltd., London, 1958), pp. 259–268 and pp. 434–439].

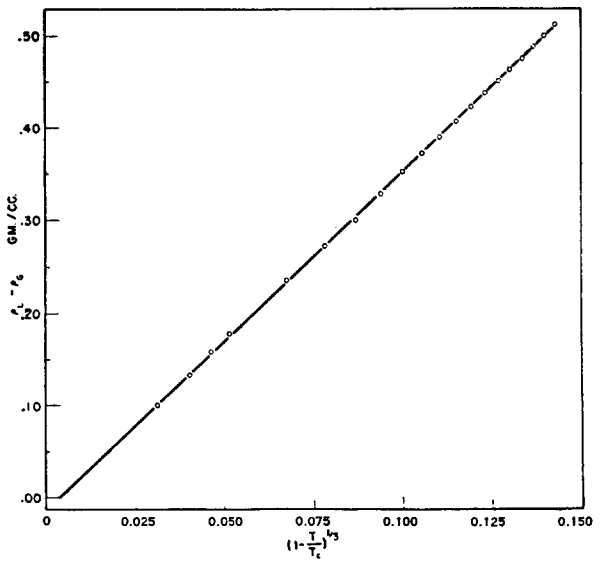


FIG. 2. Coexistence curve for xenon: plot of $(\rho_L - \rho_G)$ vs $[1 - (T/T_c)]^{\frac{1}{3}}$ (after Weinberger and Schneider[9]).

of the compressibility along the coexistence curve as $(T_c - T)^{-1}$ then follows from (1.10).

By integrating (1.9) with respect to Δp one finds the free energy $F(T, \rho)$. If one assumes that the additive constant of integration is also a nonsingular function of T and imposes the continuity of F on the coexistence curve (1.12) one finally derives the prediction (c) of the discontinuity in $C_V(T)$.

2. THE NATURE OF THE CRITICAL SINGULARITIES

Perhaps the most striking test of the predictions of the classical theories is provided by the data on the coexistence curves of simple gases. Some time ago Guggenheim[7] showed that the gases Ne, Ar, Kr, Xe, N_2, and O_2 obey closely a law of corresponding states of the form

$$(\rho_L - \rho_G)/2\rho_c = A(1 - T/T_c)^{\beta}, \quad (T \to T_c), \qquad (2.1)$$

with $\beta = \frac{1}{3}$. This relation was observed to hold with an accuracy of 0.5% (or better) in $\Delta\rho/\rho_c$ and in $\Delta T/T_c$ and over a range of temperatures from $T/T_c \simeq 0.6$ up to within $\frac{1}{2}$% of the critical temperature by when $(\rho_L - \rho_G)/2\rho_c$ had fallen to about 0.30. In earlier work on CO_2, Michels, Blaisse, and Michels[8] found that the coexistence curve could be fitted over a similar range but with somewhat greater accuracy by (2.1) with the index $\beta = 0.357$.

These results seem to be in clear disagreement with the classical prediction $\beta = \frac{1}{2}$ which should be applicable in range of $(T - T_c)$ and $(\rho - \rho_c)$ observed. They certainly show that the experimental coexistence curve is much flatter than the van der Waals curve. However, experiments near the critical region are very difficult to perform; a long time is needed to establish equilibrium and hysteresis phenomena are difficult to avoid; the system is very susceptible to minute amounts of impurities and, due to the large compressibility, highly sensitive to gravitational fields. Indeed as shown by Weinberger and Schneider[9] it is important to take special precautions to reduce the effects of gravity if the true shape of the coexistence curve is to be measured close to T_c. In a very careful study of xenon (see Fig. 2) they extended the density measurements down to temperatures differing from T_c by only 1 part in 30 000 (i.e., $\Delta T/T_c \simeq 0.003$%) and down to corresponding density differences of $(\rho - \rho_c)/2\rho_c \simeq 0.04$. [The temperature was controlled to within ± 0.001 °C.] Their data accurately obey the relation (2.1) with a nonclassical value of the index β over about three decades in $(T - T_c)$. Analysis of their measurements indicates[10] that

$$\beta = 0.345 \pm 0.015 \qquad (2.2)$$

which is not inconsistent with a value of exactly $\frac{1}{3}$ as may be seen in Fig. 2.

While it is always possible that measurements taken much closer still to the critical point might yet yield the value $\beta = \frac{1}{2}$ it seems reasonable to conclude that the classical theory does *not* provide the correct description of reality. Furthermore measurements of the phase boundaries of binary fluid mixtures near both their upper and lower critical points are also fitted well by the same cube-root law.[11] This suggests that the behavior close to a critical point is insensitive to the detailed nature of the intermolecular forces. To check how far β is really independent of the interaction potentials it would be desirable to have measurements on other systems of an accuracy matching the experiments on xenon. One should note, however,

[7] E. A. Guggenheim, J. Chem. Phys. **13**, 253 (1945).

[8] A. Michels, B. Blaisse, and C. Michels, Proc. Roy. Soc. (London) **A160**, 358 (1937). More accurate measurements on CO_2 have since been made by H. L. Lorentzen, Acta Chem. Scand. **7**, 1335 (1953) and later work.

[9] M. A. Weinberger and W. G. Schneider, Can. J. Chem. **30**, 422 (1952).

[10] Note that the plot of $\rho_L - \rho_G$ versus $[1 - (T/T_c)]^{1/3}$ in Fig. 2 is a good straight line down to $(\rho_L - \rho_G)/2\rho_c = 0.04$ but does not extrapolate exactly to the origin [$\rho_L = \rho_G$ at $T = T_c$] as it should. This suggests that the index β is not precisely $\frac{1}{3}$ and a log-log plot leads to the value quoted. The uncertainty reflects the spread of the experimental points about the best straight line.

[11] See Ref. 2, pp. 165–166 and especially the work of O. K. Rice referred to therein. See, also, D. R. Thompson and O. K. Rice, "Shape of the coexistence curve in the perfluoromethylcyclohexane-carbon tetrachloride system, II." J. Am. Chem. Soc. (1964) in press.

TABLE I. Critical indices.

Index	below T_c			above T_c ($\rho = \rho_c$)			
	α'	β	γ'	α	γ	ν	η
Defined in equations	(2.7, 2.12)	(2.1)	(2.8)	(2.7, 2.12)	(2.8)	(3.10, 5.6)	(4.7, 5.3)
Classical theory	$0_{\text{discon.}}$	$\frac{1}{2}$	1	$0_{\text{discon.}}$	1	$\frac{1}{2}$	0
Lattice gases $d = 2$	$0_{\log}$	$\frac{1}{8}$	$1\frac{3}{4}$	$0_{\log}$	$1\frac{3}{4}$	1	$\frac{1}{4}$
Lattice gases $d = 3$	≥ 0	$\simeq\frac{5}{16}$	$\geq 1\frac{1}{4}$	≥ 0, ≤ 0.2	$1\frac{1}{4}$	$\simeq 0.64$	$\simeq\frac{1}{16}$
Experiment	$\gtrsim 0_{\log}$	0.33–0.36	≥ 1.27(?)	≥ 0.1 ?	>1.1 ?	>0.55 ?	>0(?)

Note: α', β, and γ' are related by (2.20) and γ, ν, and η by (5.7). The queries, ? and (?), indicate greater and lesser degrees of experimental doubt.

that any value of β between $\frac{1}{2}$ and $\frac{1}{4}$ is inconsistent with the assumption that the critical point is a nonsingular point of the free energy in the sense discussed in the previous section.

The shape of the coexistence curve may also be studied theoretically for more-or-less idealized models of a fluid. The only model so far sufficiently tractable to yield significant predictions in the critical region is the very simplest lattice gas in which each molecule occupies a site of a lattice to the exclusion of other molecules and interacts, attractively, only with nearest-neighboring molecules. This model is equivalent to the well known Ising model of ferromagnetism which has been studied intensively.[12,13]

The exact calculation of the free energy of the plane square lattice gas along its critical isochore (corresponding to zero magnetic field) was first achieved by Onsager.[14] In addition, however, Onsager[15] and Yang[16] were able to find the coexistence curve (corresponding to the spontaneous magnetization). They found the relation (2.1) but with the index

$$\beta = \tfrac{1}{8}. \tag{2.3}$$

This result, which implies a very flat coexistence curve, is, of course, quite inconsistent with classical theory.[17] It is important, furthermore, to note that the index β is *independent of lattice structure* for all soluble plane Ising lattices (including the triangular, honeycomb, kagomé, and checkerboard lattices[12,13]).

For three-dimensional lattice gases no rigorous theoretical results are available. Nonetheless on the basis of sufficiently long power-series expansions[18] it has proved possible to draw quite accurate conclusions concerning the shape of the corresponding coexistence curves. [The coefficients are analyzed numerically with the aid of the recently introduced technique of Padé approximants.[19]]

The behavior again appears to be independent of lattice structure. [The simple, body-centered and face-centered cubic lattices[20,21] and the tetrahedral (diamond) lattice[22] have been studied; the latter on the basis of the more sensitive ratio method.[23]] Dimensionality, however, is important since, in contrast to (2.3), the index β is found to lie in the range[21,22]

$$0.303 \leq \beta \leq 0.318 \tag{2.4}$$

which is consistent with the conjecture $\beta = \frac{5}{16} = 0.31250$.

It is remarkable, and perhaps unexpected, that a model as simple as a lattice gas with only nearest-neighbor interactions should yield a result for the shape of the coexistence curve so close to the experimental results (2.1) and (2.2). The agreement suggests that in the critical region the lattice gas represents rather adequately the pertinent features of a real gas. It appears that only the grosser features of the model—in particular the dimensionality and the short range of the forces—are really essential for obtaining a good description of critical behavior.

It seems probable, nonetheless, that the difference of about 0.025 between the experimental and theoretical values of β is a real discrepancy due, presumably, to the more artificial aspects of the Ising Hamiltonian which, in particular, restricts the molecules to the lattice positions. There remains

[12] A comprehensive review of the Ising model is C. Domb, Advan. Phys. 9, Nos. 34, 35 (1960), while Ref. 13 is a brief review of more recent results.

[13] M. E. Fisher, J. Math. Phys. 4, 278 (1963).

[14] L. Onsager, Phys. Rev. 65, 117 (1944).

[15] L. Onsager, Nuovo Cimento Suppl. 6, 261 (1949); see also E. W. Montroll, R. B. Potts, and J. C. Ward, J. Math. Phys. 4, 308 (1963).

[16] C. N. Yang, Phys. Rev. 85, 808 (1952); T. D. Lee and C. N. Yang, Phys. Rev. 87, 410 (1952).

[17] Since β is the inverse of an integer it could still be possible for the free energy to be an analytic function of ρ at the critical point. In view of the logarithmic singularity in the specific heat, however (Ref. 14 and the discussion below), this possibility seems rather remote.

[18] The expansion variable is $\exp[-V_0/kT]$ where V_0 is the depth of the well in the pair interaction potential. Details of the expansions are given in Ref. 11.

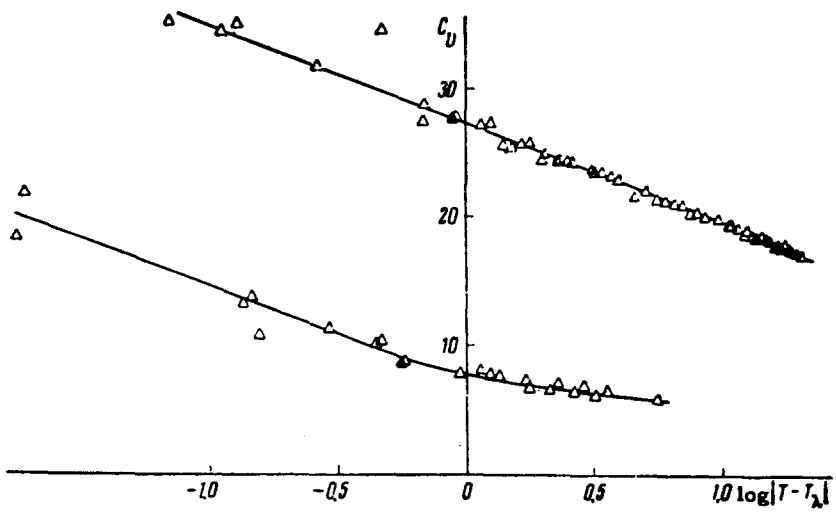
[19] G. A. Baker, Jr., J. L. Gammel, and J. G. Wills, J. Math. Anal. Appl. 2, 405 (1961).

[20] G. A. Baker, Jr., Phys. Rev. 124, 768 (1961).

[21] J. W. Essam and M. E. Fisher, J. Chem. Phys. 38, 802 (1963).

[22] J. W. Essam and M. F. Sykes, Physica 29, 378 (1963).

[23] C. Domb and M. F. Sykes, J. Math. Phys. 2, 63 (1961).



Variation of C_V of argon with $\log|T - T_c|$. ($T_c = 150.5°$ K).

FIG. 3. Variation of the constant-volume specific heat of argon along the critical isochore (after Bagatskii *et al.*[25]). (Note that the logarithm to the base ten of $|T - T_c|$ in °K is plotted.)

the theoretical problem of calculating β for more realistic continuum models. (For convenience the various results for β are collected in Table I.)

Turning now to the question of the specific heats, it has long been known that real gases exhibit a large "anomalous" specific-heat maximum above T_c which lies near the critical isochore and which is not expected on classical theory.[24] Similarly in the two-phase region below the critical point specific heats rise much more rapidly than expected as T approaches T_c. From the earlier measurements one could not conclude with certainty that the specific heat actually became infinite at T_c but recent measurements by Bagatskii, Voronel', and Gusak[25] of $C_V(T)$ for argon along the critical isochore (see Fig. 3) suggest strongly that

$$C_V(T) \to \infty \quad \text{as} \quad T \to T_c\pm. \tag{2.5}$$

Such a result is again inconsistent with classical theory.

The measurements covered a range from 15% below to 5% above T_c at temperature intervals of 0.04 to 0.05°C (corresponding to $\Delta T/T_c \simeq 0.03\%$). Over a range of one or two decades in $|T - T_c|$ the specific heat could be fitted quite well[25] by a logarithmic singularity of the form

$$C_V(T) \simeq -A^{\pm} \log|1 - (T/T_c)| + B^{\pm} \qquad (T \gtrless T_c), \tag{2.6}$$

where the marked asymmetry of the curve (see Fig. 3) indicates that $B^+ \ll B^-$ and possibly that $A^+ < A^-$. The specific-heat curve in fact resembles quite closely the famous lambda anomaly displayed by liquid helium at its transition to the superfluid state.[26] [In this case the formula (2.6) is followed very accurately over four or more decades.[26] It should be noted, however, that the lambda point of helium has a quantum-mechanical origin and is not a critical point in the usual sense.]

The data for argon are not at present, accurate enough to confirm (2.6) as closely as might be wished.[26a] To avoid prejudicing the conclusions one should preferably consider a singularity of a form such as

$$C_V(T) \simeq (A^{\pm}/\alpha)\{|1 - (T/T_c)|^{-\alpha} - 1\} + B^{\pm} \qquad (T \gtrless T_c), \tag{2.7}$$

and ask for the experimental value (and uncertainty) of the index α. [When $\alpha \to 0$ this expression reduces to the logarithmic singularity (2.6).] In particular the definite curvature of the plot of $C_V(T)$ versus $\log|T - T_c|$ for $T > T_c$ (see Fig. 3) suggests that the true value of α might be greater than say 0.1. If this is so it seems probable that below T_c the index has a somewhat different value, α' which is probably less than 0.1. It would be valuable to have similar and more extensive measurements on the other noble gases in order to test the relation

[24] See Ref. 2, pp. 100–101; Ref. 8; and A. Michels, J. M. H. Levelt, and W. de Graaff, Physica 24, 769 (1958).

[25] M. I. Bagatskii, A. V. Voronel', and B. G. Gusak, Zh. Eksperim. i Teor. Fiz. 43, 728 (1962) [English transl.: Soviet Phys.—JETP 16, 517 (1963)].

[26] M. J. Buckingham and W. M. Fairbank, in *Progress in Low Temperature Physics III*, edited by C. J. Gorter (North-Holland Publishing Company, Amsterdam, 1961), Chap. 3.

[26a] More recent experiments on oxygen have given very similar results: A. V. Voronel', Yu. R. Chasshkin, V. A. Popov, and V. G. Simkin, Zh. Eksperim. i Teor. Fiz. 45, 828 (1963) [English transl.: Soviet Phys.—JETP 18, 568 (1964)].

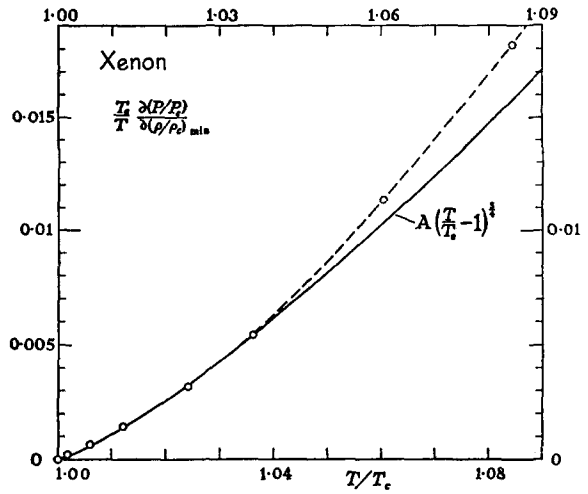


FIG. 4. Variation of the inverse of the maximum isothermal compressibility of xenon (based on the experiments of Habgood and Schneider[30]).

(2.7) more critically and to decide how far a logarithmic singularity might be truly "universal."[26a]

Significant theoretical predictions are again available only for the simple lattice gas models. Onsager's rigorous solution[14] for the plane square lattice (and subsequent results for *all* other soluble plane lattices[12,13]) yielded a *symmetric* logarithmic specific heat singularity, i.e., $\alpha = \alpha' = 0$ and $A^+ = A^-$, $B^+ = B^-$. Although this famous result demonstrates conclusively the weakness of the classical theory and is very suggestive in view of the experimental results it is, unfortunately, restricted to two-dimensional systems.

For three-dimensional Ising lattices the specific-heat series expansions have been calculated both above and below T_c. Numerical analysis of these series indicates that $C_V(T)$ is almost certainly infinite at T_c but the precise nature of the divergence is more difficult to ascertain. Below the critical point the series can be fitted well by a logarithmic singularity (i.e., $\alpha' = 0$).[13,22,27,28] As with the experimental results, however, it is not easy to exclude the possibility of a slightly sharper singularity corresponding in (2.7) to, say, $\alpha' = 0.06$. On the high-temperature side of the transition the series may be fitted moderately well by a logarithmic singularity if $A^+/A^- \simeq \frac{1}{3}$.[27,13] However, careful analysis of the ratios of coefficients definitely suggests a sharper singularity of the form (2.7) with $\alpha \simeq 0.2$.[29,13] This would lead to a shallow parabolic curve for $C_V(T)$ versus $\log|T - T_c|$ not inconsistent with the corresponding (lower) experimental curve in Fig. 3.

[27] M. E. Fisher and M. F. Sykes, Physica **28**, 939 (1962).
[28] G. A. Baker, Jr., Phys. Rev. **129**, 99 (1963).
[29] C. Domb and M. F. Sykes, Phys. Rev. **108**, 1415 (1957).

One may hope that with further work based on longer series or on more rigorous arguments the theoretical conclusions will be drawn more firmly. The present results are summarized in Table I.

The accurate experimental measurement of the isothermal compressibility of a gas near its critical point is not easy and the classical prediction (1.4) that K_T should diverge as $(T - T_c)^{-1}$ along the isochore does not seem to have been properly tested. In practice plots of $1/K_T$ versus temperature are distinctly concave upwards in the critical region which suggests that the compressibility might diverge more sharply than a simple pole, i.e., as

$$K_T(T) \approx \frac{B}{|(T/T_c) - 1|^{\gamma}} \quad (\rho = \rho_c, T \to T_c) \qquad (2.8)$$

with $\gamma > 1$. By way of illustration a plot of $1/K_T$ (suitably normalized) versus T/T_c for xenon is shown in Fig. 4.[30] The experimental results indicated by circles and the broken line, were obtained by differentiation of the experimental isotherms, a procedure which is necessarily subject to appreciable uncertainty when K_T is large. The solid curve in Fig. 4 corresponds to the nonclassical prediction (2.8) with $\gamma = 1.25$ (see below) and is evidently quite consistent with the experimental points near T_c. This fit cannot be considered very significant, however, since estimates of γ based on the data alone are rather indefinite, although they do seem to indicate that γ is greater than 1.1.

More extensive and accurate experimental data would be extremely valuable since the compressibility is an important theoretical parameter and one which is usually somewhat easier to calculate than the specific heat or the coexistence curve. Indeed the theoretical situation for the simple lattice gases is quite unequivocal (see Table I). On the basis of Onsager and Kaufman's exact calculations[31] of the correlation functions it can be shown[32] that the compressibility of the plane square lattice gas (which is isomorphic to the magnetic susceptibility) should diverge as (2.8) with

$$\gamma = 1\tfrac{3}{4}. \qquad (2.9)$$

This represents a very large deviation from the classical prediction $\gamma = 1$. Numerical examination of the corresponding series expansions confirms (2.9) in two dimensions for all other lattices and leads,

[30] Figure 4 is based on the measurements of H. W. Habgood and W. G. Schneider, Can. J. Chem. **32**, 98 (1954) as presented in their Fig. 4.
[31] B. Kaufman and L. Onsager, Phys. Rev. **76**, 1244 (1949).
[32] See Sec. 5 and M. E. Fisher, Physica **25**, 521 (1959).

in three dimensions, to the estimate

$$\gamma = 1.250, \tag{2.10}$$

which is again independent of lattice structure and accurate to ± 0.001 or better.[12,20,23]

The compressibility of the simple lattice gases can also be studied on the coexistence curve below T_c.[21] In two dimensions it is found that (2.8) holds with the index $\gamma' = \frac{7}{4}$ so that $\gamma' = \gamma$ although the amplitude B^- is much smaller than B^+. In three dimensions the analysis yields $\gamma' \simeq 1.25 = \gamma$ but cannot at present exclude the possibility that γ' exceeds γ (by perhaps 0.05).

On the basis of heuristic arguments related to the Frenkel–Bijl–Band picture of condensation[33] Essam and Fisher[21,34] conjectured that the indices for the specific heat, coexistence curve and compressibility below T_c were in general related by

$$\alpha' + 2\beta + \gamma' = 2, \tag{2.11}$$

where the index α' is defined, more precisely, by

$$\alpha' = \lim_{T \to T_c^-} [\log C_V(T)/|\log (T_c - T)|], \tag{2.12}$$

and similarly for β and γ'. A logarithmic specific heat still corresponds to $\alpha' = 0$ so that the relation (2.11) is certainly verified for the two-dimensional lattice gases (see Table I). The formula remains true even for a van der Waals gas since a discontinuity in specific heat is now also equivalent to $\alpha' = 0$ [assuming only $C_V(T_c-) > 0$].

Rushbrooke[35] has shown that the relation (2.11) can be proved as an inequality (with $\geq$ replacing $=$) by purely thermodynamic reasoning. His argument as presented applies only to a ferromagnetic system but it may be adapted for a fluid as follows.

Firstly recall that in the two-phase region the specific heat at constant *total* volume is related to the properties of the system in its liquid and gaseous phases separately by[36]

$$C_V = x_L C_\sigma^L + x_G C_\sigma^G - T\left(\frac{\partial p}{\partial T}\right)_\sigma \left[x_L \left(\frac{\partial V_L}{\partial T}\right)_\sigma + x_G \left(\frac{\partial V_G}{\partial T}\right)_\sigma\right], \tag{2.13}$$

where the subscript σ denotes properties along the coexistence curve and where the mole fractions are given by

$$x_L = \frac{V_G - V}{V_G - V_L}, \qquad x_G = \frac{V - V_L}{V_G - V_L}. \tag{2.14}$$

Now by the standard argument used to relate C_p and C_V one may show that in a single phase

$$C_\sigma = C_V - T\left(\frac{\partial p}{\partial V}\right)_T \left(\frac{\partial V}{\partial T}\right)_p \left(\frac{\partial V}{\partial T}\right)_\sigma. \tag{2.15}$$

On eliminating the factor $(\partial V/\partial T)_p$ through

$$\left(\frac{\partial V}{\partial T}\right)_\sigma = \left(\frac{\partial V}{\partial T}\right)_p + \left(\frac{\partial V}{\partial p}\right)_T \left(\frac{\partial p}{\partial T}\right)_\sigma, \tag{2.16}$$

and substituting for C_σ^L and C_σ^G separately in (2.12) one expresses C_V in terms of C_V^L and C_V^G. Finally on introducing the coexisting densities ρ_L and ρ_G and the corresponding isothermal compressibilities K_T^L and K_T^G, one gets

$$C_V(T) = x_L C_V^L + x_G C_V^G + \frac{x_L T}{\rho_L^3 K_T^L}\left(\frac{\partial \rho_L}{\partial T}\right)^2 + \frac{x_G T}{\rho_G^3 K_T^G}\left(\frac{\partial \rho_G}{\partial T}\right)^2. \tag{2.17}$$

Now C_V^L and C_V^G are necessarily positive since they are essentially mean square energy fluctuations. Consequently all terms on the right of (2.17) are positive and by dropping the first three we obtain the inequality

$$C_V(T) \geq \frac{x_G T}{\rho_G^3 K_T^G}\left(\frac{\partial \rho_G}{\partial T}\right)^2. \tag{2.18}$$

As the critical point is approached at constant density x_G (and x_L) approaches the value $\frac{1}{2}$, ρ_G (and ρ_L) tends to ρ_c and $(\partial \rho_G/\partial T)$ diverges as $(T_c - T)^{-1+\beta}$.[37] If K_T^G, the compressibility at condensation, diverges as $(T_c - T)^{-\gamma'}$ we obtain

$$\log C_V(T) \geq (2 - 2\beta - \gamma')\,|\log (T_c - T)| + \cdots. \tag{2.19}$$

The higher-order terms vanish on dividing by $|\log (T_c - T)|$ and taking the limit $T \to T_c-$ which, by (2.12), yields the index α'. We have thus proved quite generally

$$\alpha' + 2\beta + \gamma' \geq 2. \tag{2.20}$$

Various consequences follow from this inequality. For the two-dimensional lattice gases the rigorous results[14–16] $\alpha' = 0$ and $\beta = \frac{1}{8}$ show that $\gamma' \geq \frac{7}{4}$ thereby confirming the numerical estimates. [As before, the van der Waals gas corresponds to the case of equality.] If, for a three-dimensional lattice gas, the values $\beta = 0.3125$ and $\gamma' = 1.25$ are

[33] J. Frenkel, *Kinetic Theory of Liquids* (Oxford University Press, London, 1946), Chap. VII; J. Chem. Phys. **7**, 200, 538 (1939); A. Bijl, Doctoral Dissertation, Leiden, 1938; W. Band, J. Chem. Phys. **7**, 324, 927 (1939).

[34] Other conjectures relating the indices γ and β have been made by B. Widom [J. Chem. Phys. **37**, 2703 (1962)].

[35] G. S. Rushbrooke, J. Chem. Phys. **39**, 842 (1963).

[36] See, for example, Ref. 2, p. 41.

[37] We assume (as is true in reality and for the models considered) that as $T \to T_c$ $[\frac{1}{2}(\rho_L + \rho_G) - \rho_c]$ does not vanish as rapidly as does $(\rho_L - \rho_G)$.

accepted one must conclude that $\alpha' \simeq 0.125$. Conversely for a logarithmic specific heat singularity the compressibility index γ' would have to exceed 1.25 by about 0.1. At present it is difficult to judge between these alternatives.[21,35]

For a real gas the evidence of Fig. 3 suggests $0.1 > \alpha' \geq 0$ and the coexistence data indicate $\beta < 0.36$. Consequently we should certainly have

$$\gamma' \geq 1.27, \tag{2.21}$$

and probably $\gamma' \geq 2 - 2(0.345) = 1.31$. It would be most interesting to have an experimental test of this prediction since it differs appreciably from the classical result.

It should be mentioned that Widom and Rice[38] have observed that the critical isotherms of real gases also deviate significantly from the classical prediction being much flatter than the cubic (p, ρ) curve which follows from the van der Waals equation. As yet however, this feature has not been investigated theoretically for lattice gases.

3. PAIR CORRELATION FUNCTION AND CRITICAL SCATTERING

To obtain insight into the microscopic nature of a fluid in the critical region it is natural to consider the many-particle distribution functions $n_s(\mathbf{r}_1, \cdots \mathbf{r}_s)$ which describe the correlations between the constituent molecules. In particular the pair correlation function defined, for a uniform system, by

$$g(\mathbf{r}_{12}) = n_2(\mathbf{r}_1, \mathbf{r}_2)/n_1(\mathbf{r}_1)n_1(\mathbf{r}_2) = n_2(\mathbf{r}_{12})/\rho^2 \tag{3.1}$$

is of central importance. When the system is in one phase $g(r) \to 1$ as $r \to \infty$ and one may introduce the net correlation function

$$G(\mathbf{r}) = g(\mathbf{r}) - 1 \tag{3.2}$$

which decays to zero as $r \to \infty$. The deviation of $G(\mathbf{r})$ from zero is a direct measure of the influence of one molecule on another.

As is well known $G(\mathbf{r})$ is rather directly related to the thermodynamic variables of the system. For our purposes the most important result is the so called fluctuation theorem for the isothermal compressibility,

$$k_B T\left(\frac{\partial \rho}{\partial p}\right)_T = k_B T \rho K_T = 1 + \rho \int G(\mathbf{r})\, d\mathbf{r}, \tag{3.3}$$

which is a quite general consequence of the laws of statistical mechanics. [For a two-dimensional system the integral in (3.3) is restricted appropriately while for a lattice system it is replaced by a sum.]

Now $G(\mathbf{r})$ is essentially a bounded function (more precisely its integral over a finite region is bounded in virtue of the existence of a maximum density arising from the incompressibility of real molecules.) Hence the fact that K_T becomes infinite at the critical point can only be understood if the integral over $G(\mathbf{r})$ diverges at its upper limits. This means that at the critical point the net correlation function becomes *long-range* in the sense that it decays to zero more slowly than $1/r^3$ (or in d-dimensions than $1/r^d$). It is clearly of interest to know the precise nature of this critical decay and to understand the rate of approach of $G(\mathbf{r})$ to its long-range behavior as $T \to T_c$.

Fortunately the pair correlation function can also be studied directly by scattering waves off the system. In practice experiments are usually performed with light or with x rays but thermal neutrons may also be used.[39] The observed angular dissymmetry is then a direct measure of the degree of correlation. To the extent that multiple scattering may be neglected (first Born approximation) we have for the scattering intensity

$$I(\mathbf{k})/I_0(\mathbf{k}) = \chi(\mathbf{k}) = 1 + \rho\hat{G}(\mathbf{k}), \tag{3.4}$$

where $I_0(\mathbf{k})$ is the scattering intensity in the absence of correlation (the molecular form factor), $\mathbf{k}$ is the wave vector, $k = (4\pi/\lambda) \sin \frac{1}{2}\theta$, and where

$$\hat{G}(\mathbf{k}) = \int e^{i\mathbf{k}\cdot\mathbf{r}} G(\mathbf{r})\, d\mathbf{r} \tag{3.5}$$

is the Fourier transform of $G(\mathbf{r})$; for an isotropic three-dimensional system

$$\hat{G}(k) = 4\pi \int_0^\infty \frac{\sin kr}{kr} G(r)\, r^2\, dr. \tag{3.6}$$

The ratio $\chi(k)$ may be regarded as a generalized "susceptibility" since it measures the response of the fluid to an impressed periodic potential of wave number k.[40]

Comparison of (3.5) and (3.4) with the fluctuation relation (3.3) shows that

$$\chi(0) = \lim_{k\to 0} I(k)/I_0(k) = 1 + \rho\hat{G}(0) = k_B T \rho K_T, \tag{3.7}$$

so that the scattering intensity extrapolated to zero angle is proportional to the isothermal compress-

[38] B. Widom and O. K. Rice, J. Chem. Phys. **23**, 1250 (1955).

[39] L. Van Hove, Phys. Rev. **95**, 249 (1954). To discuss neutron scattering fully one must also consider the time dependence of the pair correlation function. Near the critical point, however, the decay of fluctuations probably becomes slower and it is reasonable to neglect this aspect of the problem in first approximation.

[40] P. G. de Gennes, Nuovo Cimento **9**, Suppl. 1, 240 (1958).

ibility. By virtue of the divergence of K_T at the critical point the low-angle scattering must thus become very large as the critical point is approached in the one-phase region. This is the "anomalous" critical scattering long known with visible light as *critical opalescence*. In physical terms one may say that the large compressibility near the critical point allows long-wavelength density fluctuations to grow to large amplitude and these produce visible diffraction.[39]

The classical theory of critical scattering is that developed by Ornstein and Zernike.[41,42] Their results have since been rederived many times and in various ways. In order to discuss the validity of their conclusions we will outline two approaches which characterize most of the derivations: on the one hand the original method of Ornstein and Zernike[41,43] which is perhaps the more mathematical, and on the other hand a semithermodynamic method which concentrates attention on the fluctuations of the free energy and their relation to the gradients of the density deviations. This latter approach was initiated by Rocard[44–46] in the spirit of Einstein's semiphenomenological ideas, but we will follow the presentation of Landau.[46]

Ornstein and Zernike[41] argue on a heuristic basis that the correlation $G(\mathbf{r}_1 - \mathbf{r}_2)$ between molecules 1 and 2 can be regarded as caused by (i) a *direct* influence of 1 on 2 described by the so-called "direct correlation function" $C(\mathbf{r}_1 - \mathbf{r}_2)$ which should be short-ranged [essentially having the range of the pair potential $\phi(r)$], and (ii) an indirect influence propagated directly from 1 to a third molecule at $\mathbf{r}_3$ which in turn exerts its total influence on molecule 2. Integrating over $\mathbf{r}_3$ they thus write the relation

$$G(\mathbf{r}_1-\mathbf{r}_2)=C(\mathbf{r}_1-\mathbf{r}_2)+\rho\int C(\mathbf{r}_1-\mathbf{r}_3)G(\mathbf{r}_3-\mathbf{r}_2)\,d\mathbf{r}_3. \quad (3.8)$$

In the absence of an independent theory enabling one to calculate $C(\mathbf{r})$ in terms of the molecular parameters this relation is really only a *definition* of the direct correlation function: we will, in the main, adopt this attitude. However, Ornstein and Zernike regarded $C(\mathbf{r})$ as the more basic function (in its, presumably, closer relation to the intermolecular forces) and they contemplated the possibility of calculating $C(\mathbf{r})$ directly.

On introducing the Fourier transform of $C(\mathbf{r})$ the relation (3.8) (which states that $G(\mathbf{r}_1 - \mathbf{r}_2)$ and $C(\mathbf{r}_1 - \mathbf{r}_2)$ are reciprocal kernels in the sense of the theory of integral equations) can be solved to yield

$$1 + \rho\hat{G}(\mathbf{k}) = 1/[1 - \rho\hat{C}(\mathbf{k})]. \quad (3.9)$$

On substituting in (3.4) one finds for the inverse scattering intensity

$$1/\chi(\mathbf{k}) = 1 - \rho\hat{C}(\mathbf{k}). \quad (3.10)$$

Consequently the divergence of the compressibility $K_T = \chi(0)/k_B T\rho$ at the critical point is associated with the equation

$$1 - \rho\hat{C}(0) = 1 - \rho\int C(\mathbf{r})\,d\mathbf{r} = 0. \quad (3.11)$$

This shows that the integral of $C(\mathbf{r})$ (i.e., its zeroth moment) remains finite at the cirtical point. Thus $C(\mathbf{r})$ certainly decays to zero more rapidly than $G(\mathbf{r})$ thereby confirming the expectation that it should be relatively short-ranged. To develop the theory further, however, one makes the central assumption that $C(\mathbf{r})$ is strictly short-ranged *at* (and near) the critical point in the sense that its transform $\hat{C}(\mathbf{k})$ has a Taylor series expansion in powers of k^2. In particular, one assumes that the second moment

$$R^2 = \tfrac{1}{2}\rho\langle\cos^2\theta\rangle\int r^2 C(\mathbf{r})\,d\mathbf{r}, \quad (3.12)$$

exists at the critical point and does not vary rapidly in the vicinity. [In three dimensions the average over angles yields $\langle\cos^2\theta\rangle = \frac{1}{3}$.]

We note again that unless $C(r)$ can be calculated in an independent way R^2 will have the status only of a semiphenomenological parameter. The short-range character of $C(r)$ and the existence of R^2 at the critical point constitute a major problem of the theory and we will return to it. We remark here, however, that the virial expansion of $C(r)$ may be obtained quite easily from that for $G(r)$[43,42] and indicates for low densities at least, that $C(r)$ is, in a definite sense, shorter ranged than $G(r)$.

This may be seen in terms of the graphical representations of the respective expansions. All the graphs required[47,48] are connected and have two

[41] L. S. Ornstein and F. Zernike, Proc. Acad. Sci. Amsterdam 17, 793 (1914); Physik. Z. 19, 134 (1918); *ibid.* 27, 761 (1926).

[42] F. Zernike, Proc. Acad. Sci. Amsterdam 18, 1520 (1916).

[43] See also J. Yvon, Nuovo Cimento 9, Suppl. 1, 144 (1958) and Ref. 40.

[44] Y. Rocard, J. Phys. Radium 4, 165 (1933).

[45] See also: M. Fierz in *Theoretical Physics in the Twentieth Century*, edited by M. Fierz and V. F. Weisskopf (Interscience Publishers, Inc., New York, 1960), pp. 175 *et seq.*; M. J. Klein and L. Tisza, Phys. Rev. 76, 1861 (1949).

[46] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon Press, Ltd., London, 1958), Sec. 116.

[47] See, for example, E. Meeron, J. Math. Phys. 1, 192 (1960), J. M. J. Van Leeuwen, J. Groeneveld, and J. De Boer, Physica 25, 792 (1959); T. Morita and K. Hiroike, Progr. Theoret. Phys. (Kyoto) 23, 1003 (1960).

[48] G. E. Uhlenbeck and G. W. Ford, in *Studies in Statistical Mechanics, I*, edited by J. De Boer and G. E. Uhlenbeck, (North-Holland Publishing Company, Amsterdam, 1962), Chap. B.III 4.

fixed points (corresponding to the two molecules fixed at distance $\mathbf{r}$ apart) and $n = 0, 1, 2, \cdots$ field points. Each point is associated with a power of the density ρ while the bonds are, as usual, associated with the Mayer f factors $f_{ij} = \exp[-\phi(r_{ij})/kT] - 1$. If the potential is of strictly finite range b [in the sense that $\phi(r) = 0$ for $r > b$] the factor f_{ij} vanishes unless $r_{ij} \leq b$. This fact restricts the distance up to which a graph of given type can "stretch." Thus in the expansion of $G(r)$ the longest-ranged term in a given order, say ρ^{2m+1}, comes from the open chain of $2m$ bonds which contributes up to distances $r = 2mb$ but not beyond. The graphs entering the expansion of $C(r)$, however, are restricted to be "nonnodal", i.e., none of the field points may be cutting points whose removal would separate the graph into two parts.[47] The open chain of bonds is thus excluded and the longest-ranged contribution to $C(r)$ in order ρ^{2m+1} comes from the graph consisting of two parallel chains of m bonds each. This graph, however, will make no contribution for $r > mb$.

We see therefore that in a given order the range of $C(r)$ is only half that of $G(r)$. It is clear, nevertheless, that one should not expect the virial series to converge well in the critical region (if it converges at all) so that the assumed short-range nature of $C(r)$ cannot be established in this way. For the present, however, we follow Ornstein and Zernike and accept the existence of the second moment R^2 and the possibility of a Taylor series expansion of $\hat{C}(k)$. From (3.10) and (3.12) we then obtain on neglecting terms of order k^4, the scattering formula

$$\chi(k) = 1 + \rho\hat{G}(k) \simeq R^{-2}/(\kappa^2 + k^2), \quad (k^2 \to 0), \tag{3.13}$$

where κ, which has the dimensions of an inverse length, is defined by

$$\kappa^2 = [1 - \rho\hat{C}(0)]/R^2. \tag{3.14}$$

Fourier inversion of the simple Lorentzian scattering curve (3.13) shows that the behavior of $G(r)$ for large r is given in *three* dimensions by the famous result

$$G(r) \simeq \frac{1}{4\pi\rho R^2}\frac{e^{-\kappa r}}{r} \qquad (r \to \infty). \tag{3.15}$$

From this one concludes that the correlations decay exponentially with an inverse range κ which, via the fluctuation theorem (3.3), should be related to the compressibility by

$$K_T = A/\kappa^2 \quad (T \to T_c). \tag{3.16}$$

The constant of proportionality is $A = 1/k_B T\rho R^2$ and is expected to be only slowly varying in the critical region. The divergence of K_T at the critical point thus implies

$$\kappa(T) \to 0 \quad \text{as} \quad T \to T_c, \tag{3.17}$$

so that the critical point correlation function is no longer exponentially damped but is predicted to follow the law

$$G_c(r) \simeq D/r \qquad (r \to \infty, T = T_c). \tag{3.18}$$

In as far as the relation $K_T \sim 1/\kappa^2$ is valid and the classical variation of $K_T(T)$ is accepted, the inverse range will go to zero along the critical isochore as

$$\kappa(T) \simeq \kappa^0 |1 - (T/T_c)|^{\nu} \quad (T \to T_c), \tag{3.19}$$

with $\nu = \frac{1}{2}$. This conclusion was mentioned by Zernike[42] and is accepted by other authors.[49-52] More generally, however, if one recognizes deviations of the isothermal compressibility from van der Waals behavior one would get the "nonclassical" result $\nu = \frac{1}{2}\gamma > \frac{1}{2}$ [see Eq. (2.8)].

The principal alternative approach to the Ornstein–Zernike theory is based on considering the thermodynamic work, or change in free energy, required to establish a density fluctuation in the system,[44-46,49-52] i.e., a local inhomogeneity. One supposes that a local free-energy density $F(\mathbf{r})$ can be defined for an inhomogeneous system and considers the expansion of F (or its integral over a small but macroscopic volume) about its homogeneous mean value $\bar{F}$ in terms of the local deviation $\delta\rho(\mathbf{r})$ of the density from its mean value ρ. In the spirit of the classical theory of the equation of state one assumes that a Taylor series exists even at the critical point. The first power of $\delta\rho$ may be dropped by virtue of the conservation of particles. The coefficient of $\delta\rho^2$ is, thermodynamically, proportional to $1/K_T$ and so this term must be retained.

Since the state of the system will be inhomogeneous, however, one must also expect terms dependent on $\nabla\rho$ the gradient of the density deviation, and on higher derivatives. The necessity for such terms can indeed be seen rather generally from the existence of surface tension which represents, of course, an additive contribution to the free energy directly associated with the density inhomogeneities at an interface. On the grounds of symmetry the leading term will be proportional to $(\nabla\rho)^2$. [The terms $\nabla^2\rho$ and $\rho\nabla^2\rho$ add nothing further after

[49] P. Debye, J. Chem. Phys. **31**, 680 (1959).
[50] E. W. Hart, J. Chem. Phys. **34**, 1471 (1961).
[51] M. Fixman, J. Chem. Phys. **33**, 1357 (1960).
[52] M. Fixman, J. Chem. Phys. **36**, 1965 (1962).

integrating over a small volume.[46]] Consequently, one writes the expansion

$$\Delta F(\mathbf{r}) = \tfrac{1}{2}c\delta\rho^2 + \tfrac{1}{2}d(\nabla\rho)^2 + \cdots \qquad (3.20)$$

and assumes that at least for small slowly varying deviations (i.e., small k), the higher-order terms may be neglected to a good approximation even at the critical point. By stability considerations c and d must be positive. Although c will vanish at the critical point d remains nonzero.

On introducing the Fourier components of the density deviation

$$\delta\hat{\rho}_{\mathbf{k}} = V^{-1}\int e^{i\mathbf{k}\cdot\mathbf{r}}\,\delta\rho(\mathbf{r})\,d\mathbf{r}, \qquad (3.21)$$

substituting in (3.20) and integrating over the volume of the system one obtains for the total free energy fluctuation

$$\Delta F_{\text{total}} = \tfrac{1}{2}V\sum_{\mathbf{k}}(c + dk^2)\,|\delta\hat{\rho}_{\mathbf{k}}|^2. \qquad (3.22)$$

We notice that each density mode contributes additively to the free energy so that the modes are statistically independent or, in other words, effectively noninteracting. This conclusion is, of course, a direct consequence of the truncation of (3.20) although it also has an immediate physical appeal for long wavelength modes.

Now the Boltzmann factor for a fluctuation $\delta\hat{\rho}_{\mathbf{k}}$ is $\exp[-\Delta F_{\mathbf{k}}/k_{\mathrm{B}}T]$ and consequently the mean square fluctuation is predicted, at least for small k, to be

$$\langle|\delta\hat{\rho}_{\mathbf{k}}|^2\rangle = k_{\mathrm{B}}T/V(c + dk^2). \qquad (3.23)$$

Now by the definition (3.21) of the Fourier coefficients we have

$$\langle|\delta\hat{\rho}_{\mathbf{k}}|^2\rangle = \langle\delta\hat{\rho}_{\mathbf{k}}\,\delta\hat{\rho}_{-\mathbf{k}}\rangle = V^{-1}\int e^{i\mathbf{k}\cdot\mathbf{r}}\langle\delta\rho(0)\,\delta\rho(\mathbf{r})\rangle\,d\mathbf{r},$$

where one integration over $\mathbf{r}$ space has been performed using the (approximate) translational invariance of the system. Since the mean-density fluctuation is zero

$$\langle\delta\rho(0)\,\delta\rho(\mathbf{r})\rangle/\rho = \langle[\rho + \delta\rho(0)][\rho + \delta\rho(\mathbf{r})]\rangle/\rho - \rho,$$
$$= \rho g(\mathbf{r}) + \delta(\mathbf{r}) - \rho,$$

where the second line follows from the definition (3.1) of $g(\mathbf{r})$, the delta function being added to allow for the identity of the correlated particles [which was implicitly excluded in (3.1)]. Substitution shows that

$$(V/\rho)\langle|\delta\hat{\rho}_{\mathbf{k}}|^2\rangle = 1 + \rho\hat{G}(\mathbf{k})$$
$$\simeq \frac{(k_{\mathrm{B}}T/\rho)}{c + dk^2}, \qquad (3.24)$$

which is clearly equivalent to the Ornstein–Zernike result (3.13).[53]

The relationship between the two approaches to critical scattering theory is revealed by the more recent development of a complete formal theory of the statistical mechanics of nonuniform systems which shows the significance of the direct correlation function in constructing expansions of the thermodynamic variables of inhomogeneous systems.[54–56] In particular the existence of local thermodynamic variables and the convergence of the corresponding expansions turns out to be dependent on the short range nature of $C(r)$.

A fruitful method developed by Lebowitz and Percus[56,57] is to consider the deviations in density produced by an externally imposed potential $U(\mathbf{r})$ when this is chosen to be the potential $\phi(\mathbf{r})$ which would just correspond to a molecule of the fluid fixed at the origin. The induced singlet-density deviation is then related to the pair correlation function in the homogeneous fluid as follows:

$$n_1(\mathbf{r}|\phi) = \rho + \delta n_1(\mathbf{r}) = n_2^0(\mathbf{r})/n_1^0,$$
$$= \rho + \rho G(\mathbf{r}), \qquad (3.25)$$

where the superscripts zero denote the uniform system.[58]

To obtain an equation for $n_1(\mathbf{r}|\phi)$ and hence for $G(r)$ one assumes the inhomogeneous system can be represented by a grand canonical ensemble and one asks for the relation between the external potential $U(r) = \phi(r)$ and the induced-density deviation. To this end it might be natural to try to expand $n_1(\mathbf{r}|\phi)$ as a functional Taylor series in $\phi(r)$ but Lebowitz and Percus show, on the contrary, that it is possible, and more useful, to expand $\phi(r)$ (in combination with the chemical potential) in terms of the density deviation it produces.[55–57]

[53] J. L. Lebowitz (private communication) has pointed out that by the methods of Refs. 55–57 one may show that in an appropriate grand canonical ensemble the fluctuation of the free energy is given to *second order* in $\delta\hat{\rho}_{\mathbf{k}}/\rho$ exactly by $\Delta F_{\text{total}}/Nk_{\mathrm{B}}T = \tfrac{1}{2}\sum_{\mathbf{k}}[1 - \rho\hat{C}(\mathbf{k})]\,|\delta\hat{\rho}_{\mathbf{k}}/\rho|^2$. Comparison with (3.22) shows even more directly the equivalence to the Ornstein–Zernike theory.

[54] F. H. Stillinger, Jr. and F. P. Buff, J. Chem. Phys. **37**, 1 (1962).

[55] J. L. Lebowitz and J. K. Percus, Phys. Rev. **122**, 1675 (1961); J. Math. Phys. **4**, 116 (1963).

[56] J. K. Percus, Phys. Rev. Letters **8**, 462 (1962).

[57] J. L. Lebowitz and J. K. Percus, J. Math. Phys. **4**, 248 (1963).

[58] We use the notation $\delta n_1(\mathbf{r})$ rather than $\delta\rho(\mathbf{r})$ as previously, since in the thermodynamic arguments one really is considering a macroscopic or coarse-grained density fluctuation whereas here one refers directly to the microscopic distribution functions. By the same token one should replace $\delta\hat{\rho}_{\mathbf{k}}$ in Footnote 53 by $\delta\hat{n}_1(\mathbf{k})$ whereas in Eq. (3.24) and the preceding steps $g(\mathbf{r})$ and $\hat{G}(\mathbf{k})$ represent coarse-grained or macroscopic correlation functions.

One anticipates that such an expansion would converge most rapidly when taken at each point, about a uniform system with the same *local* density $n_1(\mathbf{r}|\phi)$. The successive terms of the expansion, which is conveniently derived by the technique of functional differentiation, are then found to be multiple integrals over the density deviation with kernels which involve $C(\mathbf{r}_1 - \mathbf{r}_2)$ and certain higher-order correlation functions.

Now when $C(\mathbf{r})$ is a short-ranged function one may expand these integrals as a "local" series in the spatial derivatives of the density deviation. The justification for this step must rest, of course, on showing that $C(r)$ decays rapidly to zero for large r [and that $n_1(\mathbf{r}|\phi)$ is not too rapidly varying]. If this is the case one finds

$$\phi(\mathbf{r}) = \mu - \mu^0[n_1(\mathbf{r})] + (R^2 k_B T \rho^2/n_1^2)\nabla^2 G(\mathbf{r}) + \tfrac{1}{6}(l^2\rho^2/k_B T n_1^4 K_T^2)[\nabla G(\mathbf{r})]^2 + \cdots, \quad (3.26)$$

where μ is the chemical potential and R^2, defined already in Eq. (3.12), is the second moment of the direct correlation function. The length l is defined similarly in terms of $C(r)$ and the three-particle distribution function. Further terms of (3.26) can be written down explicitly and involve higher derivatives of $G(r)$ with coefficients depending on higher moments of $C(r)$ and the further many-particle distribution functions.[55-57]

If we now drop the terms in $(\nabla G)^2$ and higher-order derivatives and expand $n_1(r) = \rho[1 + G(r)]$ to first order in $G(r)$ we obtain

$$R^2 k_B T \nabla^2 G(\mathbf{r}) - \rho(\partial\mu^0/\partial\rho)G(\mathbf{r}) = \phi(\mathbf{r}), \quad (3.27)$$

which should be valid in the asymptotic region where $G(r)$ is small. The derivative of μ^0 arises from expanding $\mu^0[n_1(\mathbf{r})]$ and may be eliminated through the thermodynamic relation $\rho(\partial\mu^0/\partial\rho)_T = (\partial p/\partial\rho)_T = 1/\rho K_T$. On introducing the length Λ by

$$\Lambda^2 = R^2 k_B T \rho K_T = \tfrac{1}{2}\langle\cos^2\theta\rangle\rho \int r^2 G(\mathbf{r})\, d\mathbf{r}/[1 + \hat{G}(0)] \quad (3.28)$$

[where the last formula follows from (3.12), (3.10) and (3.9)] we obtain the equation

$$\nabla^2 G(r) - \Lambda^{-2} G(r) = \phi(r)/k_B T R^2 \quad (3.29)$$

first derived by Zernike.[42]

If the potential $\phi(r)$ is negligible for large r the asymptotic solution of (3.29) is, in three dimensions,

$$G(r) \approx De^{-(r/\Lambda)}/r. \quad (3.30)$$

Comparison with (3.15) shows the equivalence to the previous theory and identifies $\Lambda = 1/\kappa$ as the range of correlation. Fourier transformation of Eq. (3.29) leads directly to the previous expression (3.13) for the scattering intensity for small κ^2 and k^2.

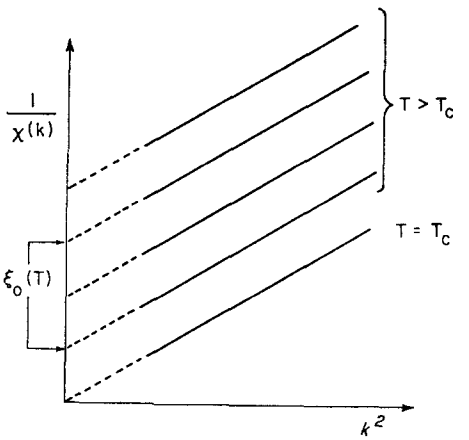


FIG. 5. Dependence of the inverse critical scattering on $k = (4\pi/\lambda)\sin\frac{1}{2}\theta$ according to the Ornstein–Zernike theory. The linear intercept $\xi_0(T)$ is proportional to $1/K_T$.

The advantage of this last derivation is that the neglected terms are explicitly displayed and that the central role of the direct correlation function is apparent. It does not, however, enable us to decide if the assumption that $C(r)$ is short-ranged at the critical point is justified. On the other hand *away* from the critical point, but in a region where K_T is moderately large the analysis indicates that the Ornstein–Zernike theory should be correct. Indeed the exponential part of the asymptotic decay law for $G(r)$ appears to have a rather wide range of validity in a one-phase fluid system since it depends essentially only on the short-range nature of the interactions.[59]

4. VALIDITY OF THE ORNSTEIN-ZERNIKE THEORY

The principal experimental predictions following from the classical theory of critical scattering [i.e., from Eq. (3.13)] are (a) that $1/\chi(k, T)$, the reciprocal of the relative scattering intensity, should vary *linearly* with k^2 with a temperature-independent coefficient of proportionality, and (b) that the extra-

[59] This conclusion follows from a formulation of statistical mechanics in which the system is taken to be in a cylinder of length L and cross section A. For forces with a hard core and strictly finite range b, a nonsingular integral kernel describes the addition of a layer of thickness b to the cylinder. The thermodynamics and correlation functions are related to the resolvent of this kernel: M. E. Fisher, abstract in Proc. Second Eastern Theoretical Physics Conf., University of North Carolina (October, 1963). D. Ruelle and, independently, J. Groeneveld (to be published) have also shown that in the region where the activity expansion can be proved to converge, $G(r)$ [suitably smoothed] decays to zero at least exponentially fast for strictly short-range potentials.

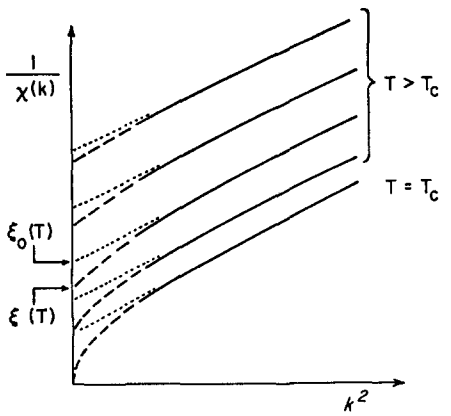


FIG. 6. Schematic variation of the inverse critical scattering expected in view of the limitations of the classical theory. (Compare with Fig. 5.) The apparent linear intercept $\xi_0(T)$ differs from the true intercept $\xi(T)$ which is proportional to $1/K_T$.

polated intercept with the $k^2=0$ axis, $\xi_0(T)$, should be proportional to $1/K_T(T)$ and hence go to zero as $T \to T_c$. The predicted behavior is shown schematically in Fig. 5.

Most modern tests of the theory have been made by light or x-ray scattering measurements on binary fluid mixtures of relatively complex organic molecules.[60-65] For example, Zimm[60] studied a mixture of perfluoromethylcyclohexane in carbon tetrachloride, Brady and Frisch,[62] perfluoroheptane in iso-octane, while Debye and co-workers[63] and McIntyre, Wims and Green[64] have investigated polystryrene–cyclohexane solutions.

Earlier measurements on binary fluids and other systems have been reviewed by Rice.[65] Recently Thomas and Schmidt[66] have made an extensive series of x-ray measurements on argon at various constant pressures in the critical region. They also give references to other more recent work on single-component systems such as carbon dioxide, ethylene and neon.

Accurate experiments are not easy to perform near the critical point and the interpretation can be confused by multiple scattering. Qualitatively, however, the theory seems quite well confirmed by the binary fluid measurements. Plots of the reciprocal scattering intensity versus k^2 in the experimentally accessible range are well represented by sets of parallel straight lines whose intercepts $\xi_0(T)$ fall roughly linearly with $(T - T_c)$ as indicated in Fig. 5. Closer inspection, nevertheless, reveals certain "anomalies", although at present these are not much larger than the experimental uncertainties.[64,67,68] In particular one observes (*a*) a tendency for the scattering curves taken near T_c to be slightly curved and to dip downwards somewhat at the lowest values of k^2 and (*b*) the intercepts $\xi_0(T)$ obtained by extrapolation of the best straight line fits to the data (all necessarily lying above some $k^2_{\min}$) do not seem to approach zero as T goes to T_c: rather plots of $\xi_0(T)$ versus T along the isochore are slightly concave upwards and tend to level off or to extrapolate to a nonzero value at T_c. These deviations are indicated in Fig. 6.

Green[68] reports that significant deviations from the Ornstein–Zernike theory were also found for the, presumably physically simpler, system of nitrogen at its critical point.[69] No deviations were observed by Thomas and Schmidt[66] in argon but they did not measure along the critical isochore and their lowest values of k^2 were relatively large.

As we show below these deviations are in the direction to be expected theoretically on the basis of an analysis of the limitations of the classical theory. Needless to say, however, it would still be desirable to have more accurate and extensive experimental data, especially for low values of k^2 and for simple systems like the noble gases, in order to elucidate fully the true nature of the critical scattering.

As we have seen the main theoretical problem in justifying the derivation of the Ornstein–Zernike result is to establish the short-range nature of the direct correlation function $C(r)$, or, what is equivalent, to show that its Fourier transform $\hat{C}(k)$ has a Taylor series expansion in powers of k^2 *at* the critical point. Our previous discussion of the thermodynamic variables at the critical point, has shown that the hypothesis of a Taylor expansion in temperature or density is probably not tenable. By analogy we should be prepared for a similar failure for the correlation functions.

An obvious defect of the theory can be seen by

[60] B. H. Zimm, J. Phys. & Colloid Chem. 54, 1306 (1950).
[61] Chow Quantie, Proc. Roy. Soc. (London) **A224**, 90 (1954); R. Fürth and C. L. Williams, Proc. Roy. Soc. (London) **A224**, 104 (1954).
[62] G. W. Brady and H. L. Frisch, J. Chem. Phys. **35**, 2234 (1961).
[63] P. Debye, H. Coll, and D. Woermann, J. Chem. Phys. **32**, 939 (1960); *ibid* **33**, 1746 (1960); P. Debye, D. Woermann and B. Chu, J. Chem. Phys. **36**, 851 (1962).
[64] D. McIntyre, A. Wims and M. S. Green, J. Chem. Phys. **37**, 3019 (1962).
[65] O. K. Rice, "Critical Phenomena" Sec. E, in *Thermodynamics and Physics of Matter*, edited by F. D. Rossini (Princeton University Press, Princeton, New Jersey, 1955).
[66] J. E. Thomas and P. W. Schmidt, J. Chem. Phys. **39**, 2506 (1963).
[67] H. L. Frisch and G. W. Brady, J. Chem. Phys. **37**, 1514 (1962).
[68] M. S. Green, J. Chem. Phys. **33**, 1403 (1960).
[69] R. L. Wild, J. Chem. Phys. **18**, 1627 (1950).

considering its application to model systems of dimensionality d different from three. In any number of dimensions the classical expression for the (appropriate) Fourier transform is formally the same, namely

$$\chi(\mathbf{k}) = 1 + \rho\hat{G}(\mathbf{k}) \simeq A/(\kappa^2 + k^2). \quad (4.1)$$

For fixed $\kappa(T) > 0$, that is *away* from the critical point, we find[70] by inverting (4.1) that as r becomes very large

$$G(r) \simeq B_d(e^{-\kappa r}/r^{\frac{1}{2}(d-1)})[1 + O(1/\kappa r)] \quad (r \to \infty,\ \kappa \text{ fixed} > 0). \quad (4.2)$$

As we argued at the end of the previous section it seems probable that this result is generally valid for *fixed* $T > T_c$ and large enough r. However, as $\kappa \to 0$ for *fixed* large r we find a different result,[70] namely, for $d \geq 3$

$$G(r) \simeq D_d(e^{-\kappa r}/r^{d-2})[1 + O(\kappa r)] \quad (\kappa \to 0,\ r \text{ fixed}) \quad (4.3)$$

while for $d = 2$

$$G(r) \simeq D_2(\log r)e^{-\kappa r}[1 + O(1/\log \kappa r)] \quad (\kappa \to 0,\ r \text{ fixed}). \quad (4.4)$$

One notices that $d = 3$ is a rather special case in which both limits (4.2) and (4.3) agree and [by comparison with Eq. (3.15)] the higher-order terms vanish identically!

Now (4.4) implies that at the critical point of a two-dimensional system the correlation function will vary as $D_2 \log r$. This is clearly nonphysical for large r and shows that the assumptions of the theory are certainly not to be trusted for two-dimensional systems.

This defect of the theory can be repaired in an *ad hoc* fashion by retaining further (nonlinear) terms in the density expansions of the free energy [in Eq. (3.20)] or in the Taylor series expansion of $n_1(\mathbf{r})$ [in Eq. (3.26)].[71] It would seem difficult, however, to justify keeping nonlinear terms in $G(r)$ rather than, say nonlinear terms in ∇G or higher-order derivatives of $G(r)$: the more so as the whole question of the convergence of such an expansion is in doubt at the critical point.

The first author to question the theory in respect of the prediction $G(r) \sim 1/r$ at $T = T_c$, for three dimensions was Green[68] who based his arguments on an integral relation for the pair correlation function derived by cluster diagram summation techniques.[47] This so called, hypernetted-chain integral equation may be written

$$1 + G(r) = \exp[-\beta\phi(r) + G(r) - C(r) + E(r)], \quad (4.5)$$

where

$$E(r) = \mathcal{E}\{\rho; G(r)\} \quad (4.6)$$

is a nonlinear integral functional of $G(r)$ known only as an expansion in powers of ρ representable in terms of certain, so called, "basic graphs."[47] [The leading term is of order ρ^2]. The same relations hold for a lattice system (with appropriate definition of the functional $\mathcal{E}$).

Green[68] considered the consequences of assuming that the term $E(r)$ involving the basic graphs, might be neglected at the critical point. Stillinger and Frisch extended his analysis to two-dimensional systems.[72]

To follow the argument let us suppose in greater generality that at the critical point of a d-dimensional system

$$G(r) \simeq D/r^{d-2+\eta} \quad (r \to \infty), \quad (4.7)$$

where the index η ($0 \leq \eta \leq 2$) measures the departure from the Ornstein–Zernike prediction. [As in Eq. (2.7) $1/r^0$ corresponds to $\log r$.] It follows that

$$\rho\hat{G}(k) \simeq \hat{D}/k^{2-\eta}, \quad (k \to 0) \quad (4.8)$$

so that at fixed density one has, through the relation (3.9),

$$\hat{C}(k) \simeq C(0)[1 - c_0k^{2-\eta} + \cdots], \quad (k \to 0). \quad (4.9)$$

If $\eta > 0$ asymptotic inversion yields

$$C(r) \simeq F/r^{d+2-\eta} \quad (r \to \infty). \quad (4.10)$$

Thus when $\eta > 0$, the direct correlation function is *also* "long ranged" in the sense that its second moment does not exist, although it certainly decays to zero more rapidly than $G(r)$, in fact by a factor $1/r^{4-2\eta}$. Notice that if $\eta = 0$ we find instead

$$C(r) \simeq F'e^{-\kappa' r}/r, \quad (\kappa'^2 = 1/c_0) \quad (4.11)$$

so that only in this special case is $C(r)$ short ranged in the sense that $\hat{C}(k)$ has a Taylor series expansion at $k^2 = 0$.

To analyze the hypernetted integral relation it is convenient to define

$$S(r) = G(r) - C(r). \quad (4.12)$$

Clearly $S(r)$ must have the same asymptotic be-

[70] M. E. Fisher, Physica **28**, 172 (1962).
[71] M. Fixman, J. Chem. Phys. **36**, 1965 (1962).
[72] F. H. Stillinger, Jr. and H. L. Frisch, Physica **27**, 751 (1961).

havior as $G(r)$. If the potential $\phi(r)$ is short ranged we may expand the exponential in (4.5) for large r to get

$$C(r) = E(r) + \tfrac{1}{2}[S(r) + E(r)]^2 + \cdots . \qquad (4.13)$$

If we now suppose $E(r)$ can be neglected or, more weakly, that $E(r)$ decays faster than $[S(r)]^2$ it follows that

$$C(r) = \tfrac{1}{2}[S(r)]^2 + \cdots . \qquad (4.14)$$

On substituting (4.7) and (4.10) we obtain the consistency relation

$$\eta = 2 - \tfrac{1}{3}d, \qquad (4.15)$$

which fixes the asymptotic form of the correlation functions.

On this basis we would predict that at the critical point in a three-dimensional system $G(r) \sim 1/r^2$ rather than $1/r$.[68] Correspondingly $\hat{G}(k) \sim 1/k = 1/(k^2)^{\frac{1}{2}}$ so that a plot of inverse scattering versus k^2 at, and near $T = T_c$, should be significantly curved downwards for small k^2 (see Fig. 6).

In two dimensions (4.15) leads to $G(r) \sim 1/r^{4/3}$ ($T = T_c$) which is more reasonable than the Ornstein–Zernike result $\log r$. However as pointed out by Stillinger and Frisch,[72] this prediction can be tested against the rigorous result obtained by Onsager and Kaufman[31] for the correlation function of the nearest-neighbor plane square lattice gas at its critical point. [Note that $G(\mathbf{r})$ at $\rho = \rho_c$ is proportional to the spin pair correlation function $\langle S_0^z S_r^z \rangle$ of the Ising ferromagnet in zero field.] This exact result is[31,32,72,73]

$$G(r) \approx D/r^{\frac{1}{4}} \qquad (T = T_c, d = 2) \qquad (4.16)$$

so that the true value of the index η is $\frac{1}{4}$ rather than $\frac{4}{3}$ or zero.

We conclude that both the Ornstein–Zernike theory and Green's argument are incorrect for a two-dimensional lattice. Consequently both must be suspect for three-dimensional lattice systems.

It is interesting to note that if, as seems to be the case, the true value of η is less than $2 - \frac{1}{3}d$ one really has, as $r \to \infty$

$$E(r) \approx \tfrac{1}{2}[S(r)]^2 \approx \tfrac{1}{2}[G(r)]^2 \approx D^2/r^{2d-4+2\eta} \qquad (4.17)$$

so that the contribution of the basic graphs is also long ranged although it decays faster than $G(r)$. Since after all, $E(r)$ is a functional of $G(r)$ this is not really surprising.

Of course, the rigorous result (4.16) is known only for nearest-neighbor interactions and one should ask to what extent the behavior of lattices with interactions reaching to further neighbors would be similar. It seems plausible that the correct answer is that the behavior sufficiently near the critical point is qualitatively unchanged provided the range of the potential is finite [for example, if $\phi(r) = 0$ for $r > b$]. The reason for this surmise is that the range of correlation near the critical point becomes, as we have seen, very large compared to the lattice spacing and, indeed, very large compared to the potential range b. The asymptotic correlations are then determined by long chains of interactions and should thus be insensitive to the detailed variation of $\phi(r)$. [The same conclusion is really implicit in Ornstein and Zernike's and in Green's approach.]

The independence of the indices α, β, γ and η of the lattice structure is evidence for this conclusion. Further evidence comes from numerical studies of the compressibility of lattice gases with first and *second* neighbor interactions which indicate unchanged values for the index γ in two and three dimensions.[74]

It is more difficult to assess the relationship of the lattices gases to more realistic *continuum* models. At low temperatures and high densities the properties of a lattice gas will always deviate from continuum behavior but in the critical region the long range of the correlations again suggests an insensitivity to the details of the potential and hence, for a lattice subdivision sufficiently fine relative to the range of the potential, one would expect qualitatively very similar behavior.[75] This conclusion is supported by the apparently quite close resemblance between the critical singularities of real gases and of even the simplest nearest neighbor three-dimensional lattice gases discussed already in Sec. 2.

5. MORE GENERAL ANALYSIS OF CRITICAL SCATTERING

It is clear from the foregoing that the Ornstein–Zernike theory is probably not valid at the critical point of a three-dimensional system. On the contrary one should evidently expect that $G(r)$ behaves asymptotically as $1/r^{1+\eta}$ with $0 < \eta < 1$.[70,72]

A natural way to extend the Ornstein–Zernike theory within the same framework is to consider further powers of k^2 in the expansion of the Fourier

[73] See also Ref. 12, pp. 200–201 but notice that an exponent $\frac{1}{2}$ is missing on the left of Eq. (108).

[74] M. F. Sykes and N. Dalton (to be published).

[75] In one dimension one may verify explicitly that the behavior of the lattice gas approaches that of the continuum gas as the lattice spacing is made smaller relative to the scale of the potential.

transform of the direct correlation function $\hat{C}(k)$. Inclusion of such terms leads to a representation of $G(r)$ as a sum of increasingly more rapidly damped exponentials of the form (for $d = 3$)

$$G(r) \approx \frac{De^{-\kappa r}}{r} + \frac{D_1 e^{-\kappa_1 r}}{r} + \frac{D_2 e^{-\kappa_2 r}}{r} + \cdots \qquad (5.1)$$

with $\kappa < \kappa_1 < \kappa_2 \cdots$. At a fixed temperature, however, such an expansion leads to the same asymptotic behavior for $G(r)$ and to similar scattering at small k^2 as does the original theory. (For smaller enough r and large enough k^2 one must, of course, expect derivations from any general theory since the detailed nature of the potential must eventually make itself felt.)

As one considers T approaching T_c along $\rho = \rho_c$, however, it is possible that more and more exponentials in (5.1) become "excited," so that the first exponential is no longer a good approximation except for extremely large r.[76] This certainly represents essentially what happens for the two-dimensional lattice gas. Here the higher-order range parameters $\kappa_1, \kappa_2, \cdots$ obey the relation[12,14,77]

$$\kappa_n(T) - \kappa(T) \sim n\kappa_0 \, |1 - (T/T_c)|, \qquad (5.2)$$

so that the "spectrum" of exponentials closes up as $T \to T_c$ and in fact becomes dense at the critical point.

A similar behavior for a continuum model is suggested by the recent calculations of Hemmer, Kac, and Uhlenbeck[76] for a potential with a strongly repulsive core and a weakly attractive long-range exponential tail. A series similar to (5.1) can be derived but the amplitudes of successive terms are proportional to higher powers of the compressibility K_T and hence the expansion breaks down near the critical point.

These considerations [which really amount to a restatement of our previous conclusion that $\hat{C}(k)$ probably does not have a Taylor series expansion at the critical point although it does for $T > T_c$] indicate that (5.1) is not the best basis for analyzing the deviations from the Ornstein–Zernike theory. Indeed the asymptotic form $1/r^{1+\eta}$ at $T = T_c$ with $\eta \neq 0$ could only arise from (5.1) if the expansion broke down in some way. To investigate the possibilities more generally let us, therefore, extend (4.7) by writing [70] for $T \geq T_c$ and $\rho = \rho_c$

$$G(r) \simeq (De^{-\kappa r}/r^{d-2+\eta})[1 + Q(\kappa r)] \quad (r \to \infty) \qquad (5.3a)$$

where $D = D(T)$ is a relatively slowly varying function of temperature and where $Q(x) \to 0$ as $x \to 1$ and $Q(x)$ does not grow exponentially fast as $x \to \infty$. This expression is still in the spirit of the Ornstein–Zernike theory in as far as the main assumption implicit in (5.3a) is that in the critical region the correlation functions for large r can be described in terms of only *two* lengths: (i) the *range of correlation* $1/\kappa(T)$ which becomes infinite at the critical point, and (ii) an *effective range of direct interaction* $r_0(T)$ which remains finite at the critical point. In (5.3a) r_0 has been absorbed into the coefficient D. [Compare with R of the Ornstein–Zernike theory: Eqs. (3.12) and (3.15).] We could write in analogy with (3.15)

$$D = \bar{d}/\rho r_0^{2+\eta}, \qquad (5.3b)$$

where $\bar{d}$ is a dimensionless constant.

If the classical theory is valid away from the critical point, as concluded in Sec. 3, the formula (5.3a) should reduce to (4.2) when $\kappa > 0$ and r is very large, i.e. only the first term in (5.1) should remain. [For the plane square nearest-neighbor lattice gas at $\rho = \rho_c$ one may verify that (4.2) is indeed valid above T_c.[77]] This would imply that for large x

$$1 + Q(x) \approx qx^{\frac{1}{2}(d-3)+\eta} \quad (x \to \infty). \qquad (5.3c)$$

Sufficiently close to the critical point the nature of this behavior will not matter. It is clear, however, that if $Q(x)$ becomes of order unity for small x the region of significant deviation from classical theory may be rather small. [Stated alternatively, the further exponentials in (5.1) would be significantly excited only very close to the critical point.]

Accepting (5.3) we may calculate the fluctuation integral (3.3) (making the substitution $x = \kappa r$). For the divergence of the compressibility along the critical isochore this yields

$$\chi(0) = k_B T \rho K_T \simeq \hat{D}_0/\kappa^{2-\eta} \quad (T \to T_c), \qquad (5.4)$$

where $\hat{D}_0$ is a slowly varying function of T [of magnitude dependent on $Q(x)$]. Similarly a calculation of the Fourier transform of (5.3) for small k^2 yields, near T_c, the non-Lorentzian critical scattering formula

$$\chi(k) \simeq \hat{D}/(\kappa^2 + k^2)^{1-\frac{1}{2}\eta} \quad (k^2 \to 0), \qquad (5.5)$$

where $\hat{D}$ is a slowly varying function of T and k^2.[78]

[76] P. C. Hemmer, M. Kac and G. E. Uhlenbeck, J. Math. Phys. 5, 60 (1964), P. C. Hemmer, J. Math. Phys. 5, 75 (1964).

[77] The correlation functions may be expressed as a sum of integrals over the complete set of eigenvalues of the basic transition matrix for an Ising lattice (see Ref. 31). The exact limiting density of these eigenvalues is known from Onsager's work (Ref. 14).

[78] If $Q(x)$ is neglected one has for small η

$$\hat{D}(k^2) = D_0\{1 - \eta k^2/(\kappa^2 + k^2) + \cdots\}.$$

As in our discussion of the Ornstein–Zernike theory the way in which the inverse range of correlation $\kappa(T)$ vanishes as $T \to T_c$ for $\rho = \rho_c$ is related to the nature of the corresponding divergence of the compressibility as $1/(T - T_c)^\gamma$. Assuming, as in (3.19) that

$$\kappa(T) \simeq \kappa^0 |1 - (T/T_c)|^\nu \qquad (T \to T_c) \tag{5.6}$$

and substituting in (5.4) shows that the critical indices are related by

$$\gamma = (2 - \eta)\nu. \tag{5.7}$$

This relation may be checked for the plane square lattice gas since, as we have seen,[31,32,72,73] $\eta = \frac{1}{4}$ and $\gamma = 1\frac{3}{4}$ (see Table I). Consequently we should have

$$\nu = 1, \qquad (d = 2) \tag{5.8}$$

in contrast to the classical result $\nu = \frac{1}{2}$. Now the result $\nu = 1$ was in fact derived rigorously by Onsager in his original paper on the Ising model.[14] His derivation is based on the relation

$$e^{-\kappa a} = \lambda_1/\lambda_0, \tag{5.9}$$

where a is the lattice spacing, and where $\lambda_0(T)$ and $\lambda_1(T)$ are the largest and next largest eigenvalues of the basic matrix which adds a row to the lattice at temperature T.[77]

If we *assume*[79] that ν is still unity for the three-dimensional lattice gas and utilize the result[80] $\gamma = 1\frac{1}{4}$ we would predict $\eta = \frac{3}{4}$. Thence the critical point decay law would be $G_c(r) \sim 1/r^{7/4}$ in closer agreement with Green's result. However the assumption $\nu = 1$ in three dimensions is not much better justified *a priori* than the classical assumption $\nu = \frac{1}{2}$. Furthermore, examination of the other critical indices (see Table I) indicates that the classically predicted behavior is more closely approached the larger the dimensionality of the system.[81] Consequently one might anticipate that for the three-dimensional nearest-neighbor lattice gas ν lies between $\frac{1}{2}$ and 1 and η is *less* than $\frac{1}{4}$. [As mentioned previously the *assumption* $\eta = 0$ leads to $\nu = \frac{1}{2}\gamma = \frac{5}{8}$ when $d = 3$.]

To decide between these various speculations it is necessary to calculate $\kappa(T)$ or some other feature of the correlation functions. Fortunately the Ising model is again sufficiently tractable to allow some progress. It is possible at the critical density, to derive a diagrammatic expansion for the decay factor $e^{-\kappa a}$ in powers of $1/T$ via Eq. (5.9).[82] The required graphs consist of an infinite chain of connected bonds stretching right across the lattice together with nonoverlapping closed polygons, as occur in the expansion of the partition function.[12]

Evaluation of the series[83] for $\kappa(T)$ (with the aid of Padé approximants) reveals a behavior near T_c clearly intermediate between the two-dimensional and classical results in a region $T = T_c$ to $2T_c$. Direct estimation of the index ν, however, proves to be not very accurate but indicates a range $\nu = 0.6$ to 0.7. [The coefficients of the series are difficult to calculate and not very smooth.]

An alternative approach is to study the temperature dependence of the higher moments of the correlation function, namely

$$\mu_s(T) = \rho \int r^s G(\mathbf{r})\, d\mathbf{r}. \tag{5.10}$$

The zeroth moment is essentially the compressibility but the second moment is also a direct measure of the range of correlation. It is evident furthermore, that μ_2 is proportional to the curvature of $\rho\hat{G}(k)$ for small k and hence to the true limiting slope of the curve of $1/\chi(k)$ versus k^2 as $k^2 \to 0$.[84] From (5.3) we see that when $T \to T_c$, $\mu_2(T)$ diverges as

$$\begin{aligned} \mu_2(T) &\simeq M_2/\kappa^{4-\eta} \qquad (T \to T_c), \\ &\simeq M_2'/|1 - (T/T_c)|^\delta, \end{aligned} \tag{5.11}$$

with

$$\delta = (4 - \eta)\nu. \tag{5.12}$$

The coefficients M_2 and M_2' are slowly varying functions of T. By comparing (5.12) with the index relation (5.7) we see that η and ν can both be determined if γ and δ are known. [In particular only if $\delta = 2\gamma$ would we have $\eta = 0$.]

For the two- and three-dimensional lattice gases expansions for $\mu_2(T)$ is series of powers of $1/T$ at $\rho = \rho_c$ are not too difficult to calculate.[83] [The labor and graphical analysis is similar to that for the compressibility.] The coefficients prove to be rather smoothly varying and for the plane lattices numerical analysis confirms quite accurately the relation (5.12)

[79] This assumption was made tentatively in Ref. 70 and tested experimentally by Frisch and Brady (Ref. 67). A good fit was obtained but although the data revealed definite nonclassical behavior, they were not sufficiently accurate to distinguish between $\nu = 1$ and some appreciably lower value.

[80] See the discussion in Sec. 2 and Refs. 23 and 28.

[81] From Table I we see $\gamma = 1\frac{3}{4}, 1\frac{1}{4}$ for $d = 2$, $d = 3$ to which we may add $\gamma = 1.094$ ($d = 4$); see M. E. Fisher and D. S. Gaunt, Phys. Rev. **133**, A224 (1964). The classical result $\gamma = 1$ corresponds to $d \to \infty$.

[82] M. E. Fisher (to be published).

[83] M. E. Fisher and R. J. Burford (to be published).

[84] The second moment μ_2 is closely related to the length Λ defined in (3.28) and to the "persistence length" L defined by Debye (Ref. 44) explicitly: $\Lambda^2 = \frac{1}{2}\langle\cos^2\theta\rangle\, \mu_2/(1 + \mu_0)$ and $L^2 = \mu_2/\mu_0$. From (3.28) one sees that $1/\chi(k) = 1/\chi(0)\{1 + \Lambda^2k^2 + O(k^4)\}$.

[which predicts $\delta = 3.75$]. This is an important result since it provides support for the original hypothesis (5.3).

Initial estimates for the simple cubic lattice gas[83] yield $\delta = 2.538 \pm 0.003$ and hence

$$\nu = 0.644 \pm 0.003, \quad \eta = 0.060 \pm 0.007. \quad (5.13)$$

Results for other three-dimensional lattices confirm these estimates with lower accuracy. [One might mention that (5.14) is not inconsistent with the conjecture $\eta = \frac{1}{16} = 0.0625$ which is rather natural in view of the estimate $\beta \simeq \frac{5}{16}$ discussed in Sec. 2.]

The values (5.13) are in accordance with our expectation that the three-dimensional results should be closer to the classical predictions.[81] The magnitude of η for the lattice gas is indeed rather close to zero but it is not unlikely that a more realistic continuum model would lead to somewhat larger value, say $\eta \simeq 0.1$.[85] Until more rigorous theories are developed and more realistic models become soluble we must content ourselves with these rough estimates. Accordingly let us review the nature of the critical scattering to be expected on the basis of our analysis.

The peak in the critical scattering is described by (5.5) and (for $\eta > 0$) should be narrower at a fixed temperature near T_c than the Lorentzian curve of same peak height. Correspondingly a plot of the inverse scattering intensity versus k^2 should be somewhat convex although, as suggested in Fig. 6, for larger values of $k^2 > k^2_{\min}$ the curves might *appear* to be reasonably linear. For small values of η, of the magnitude (5.13), it might indeed be rather difficult experimentally to detect the increasing curvature of the scattering plots as $k^2 \to 0$ even though the curve for $T = T_c$ will theoretically have an infinite slope at $k^2 = 0$. In practice a nonzero value of η can probably best be detected by the observation that the apparent linear intercepts $\xi_0(T)$ [see Fig. 6] would not approach zero when $T \to T_c$ as must the true intercepts $\xi(T) = 1/\chi(0)$. In particular the scattering plot taken *at* $T = T_c$ would tend to extrapolate to a small positive value at $k^2 = 0$. [Of course to detect this behavior it is important to have an independent measurement of T_c and *not* to judge T_c by extrapolating $\xi_0(T)$ to zero with T as would otherwise be tempting!] This behavior is reminiscent of the experimental "anomalies" described in the previous Section although, as yet, these can probably not be regarded as fully established.

The likely fact that γ exceeds unity for a real gas and the consequent curvature of $1/K_T$ versus T would indicate that a plot of the true intercepts, $\xi(T)$, versus T should flatten out as T approaches T_c (and theoretically have zero slope at $T = T_c$). This effect might well be less obvious in a plot of the apparent intercepts $\xi_0(T)$ although it does seem to have been observed.[86]

[85] For example the effectively more "continuum-like" Heisenberg model of ferromagnetism yields $\gamma \simeq 1\frac{1}{3}$ compared with the Ising value $\gamma = 1\frac{1}{4}$ ($d = 3$). (See Refs. 90, 91).

6. SUMMARY AND CONCLUSIONS

We have shown that the classical theories of the gas–liquid critical point are unsatisfactory both on experimental and theoretical grounds. Thus the coexistence curve must be described experimentally by

$$\rho_L - \rho_G \sim (T_c - T)^\beta, \quad (6.1)$$

with $\beta \simeq 0.33$ to 0.36 (rather than with $\beta = \frac{1}{2}$) while for three-dimensional lattice gas models one finds $\beta \simeq 0.31 \simeq \frac{5}{16}$. The specific heat $C_V(T)$ measured along the critical isochore of a fluid becomes infinite at T_c, diverging approximately as $\log|T - T_c|$. A similar result holds for the lattice gas. Theoretically one also expects that the compressibility above and below T_c should diverge as

$$K_T \sim 1/|T - T_c|^\gamma \quad (6.2)$$

with $\gamma > 1$. This prediction while qualitatively correct awaits quantitative experimental verification. (Values of these indices are given in Table I.)

On theoretical grounds the classical (Ornstein–Zernike) theory of critical scattering has been shown to be unsatisfactory close to the critical point (although it probably is valid away from the critical region when the compressibility is still moderately large). More generally one should expect the scattering intensity to vary as

$$I(k) \sim 1/(\kappa^2 + k^2)^{1-\frac{1}{2}\eta}, \quad (6.3)$$

where $\eta > 0$ and where the range parameter vanishes along the critical isotherm as

$$\kappa(T) \sim (T - T_c)^\nu \quad (6.4)$$

with $\nu(2 - \eta) = \gamma$. Theoretical analysis suggests that the index η might be no larger than 0.1 (see Table I). Consequently it is probably not easy experimentally to detect deviations from the classical theory (for which $\eta = 0$). Nevertheless there are experimental indications of the failure of the classical predictions for small k^2 near T_c and these are consistent with $\eta > 0$. Final confirmation of the theory must, however, rest on further, more extensive and accurate measurements on sufficiently simple fluid systems.

[86] D. McIntyre, Ref. 64 and a private communication.

In conclusion one should mention the rather close analogy between the gas–liquid critical point and the Curie point of a ferromagnetic crystal.[87] Indeed almost all our analysis and conclusions apply directly to ferromagnetic systems if appropriately translated. The density deviation $\rho - \rho_c$ should be identified with the magnetization M while the magnetic field H is isomorphic to the chemical potential of the fluid. The critical isochore, $\rho = \rho_c$, corresponds to zero magnetic field (since the mean magnetization then vanishes) and the coexistence curve corresponds to the curve of spontaneous magnetization $M_0(T)$. The specific heat $C_V(T)$ along the critical isochore of a fluid is isomorphic to the specific heat $C_H(T)$ of a ferromagnet in zero field and the compressibilities at condensation and above T_c for $\rho = \rho_c$ correspond essentially to the initial susceptibilities $\chi_0(T) = (\partial M/\partial H)_T$, $(H \to 0)$.

The van der Waals and equivalent classical theories find their precise parallel in the Weiss molecular field theory and its extensions.[88] Similarly the Ornstein–Zernike theory and its developments have been adapted to describe the critical scattering of neutrons by ferromagnets.[89,90] The net pair correlation function $G(\mathbf{r})$ is replaced by the spin–spin correlation functions $\Gamma_{\alpha\beta}(\mathbf{r}) = \langle S_0^\alpha S_r^\beta \rangle$. Theoretically one is able to estimate the susceptibility index γ for the nearest neighbor Heisenberg model above T_c with the approximate result[91,92] $\gamma = \frac{4}{3}$ (independent of spin and lattice structure in three-dimensions). It is interesting that this nonclassical prediction has been quite accurately confirmed recently.[92–95] Furthermore modern neutron scattering experiments[95] have also given a definite suggestion of deviations from the Ornstein–Zernike theory consistent with a small positive value of the index η.

The spontaneous ferromagnetic moment $M_0(T)$ is not easy to measure near T_c but nuclear magnetic resonance experiments by Benedek and Heller[96] have shown that the somewhat analogous sublattice magnetization (or long range order) of an *anti*ferromagnet (actually MnF_2) varies as $(T - T_c)^\beta$ with $\beta = 0.335 \pm 0.010$. The relation holds with remarkable accuracy up to within millidegrees of the critical or Néel point $[\Delta T/T_c = 0.007\%]$.[97] Experiments on antiferromagnets also reveal specific heat infinities at T_c which are approached in approximately logarithmic fashion.[98] The close similarity of these results to the corresponding behavior of fluid systems presents a striking challenge to our theoretical understanding.

ACKNOWLEDGMENTS

I am most grateful to Professor G. E. Uhlenbeck for his criticisms of a first draft of this article and would like to thank Professor Mark Kac and Professor Joel Lebowitz for their comments.

[87] As mentioned in Sec. 2, this relationship is formally exact for a lattice gas and an Ising model ferromagnet. See, for example, T. D. Lee and C. N. Yang, Phys. Rev. **87**, 410 (1952).

[88] See, for example, J. H. Van Vleck, Rev. Mod. Phys. **17**, 27 (1945); P. W. Kasteleijn and J. Van Kranendonk, Physica **22**, 317 (1956).

[89] L. Van Hove, Phys. Rev. **95**, 1374 (1954).

[90] R. J. Elliott and W. Marshall, Rev. Mod. Phys. **30**, 75 (1958).

[91] C. Domb and M. F. Sykes, Phys. Rev. **128**, 168 (1962).

[92] J. L. Gammel, W. Marshall, and L. Morgan, Proc. Roy. Soc. (London) **A275**, 257 (1963).

[93] J. E. Noakes and A. Arrott, J. Appl. Phys. Suppl. **35**, 931 (1964).

[94] J. S. Kouvel [private communication based on an analysis of the measurements by P. Weiss and R. Forrer, Ann. Phys. **5**, 153 (1926).]

[95] L. Passell, K. Blinowski, T. Brun, and P. Nielsen, J. Appl. Phys. (Suppl.) **35**, 933 (1964). Note that the linear intercepts with the $k^2 = 0$ axis of the inverse scattering appear to approach a positeve value at $T = T_c$.

[96] P. Heller and G. B. Benedek, Phys. Rev. Letters **8**, 428 (1962).

[97] More recent NMR experiments by Heller and Benedek (private communication) reveal a similar behavior for the spontaneous magnetization of the ferromagnet EuS.

[98] See, for example, W. K. Robinson, and S. A. Friedberg, Phys. Rev. **117**, 402 (1960).

On the Stability of Flow of a Thermally Stratified Fluid under the Action of Gravity*†

DONALD KOPPEL
Hudson Laboratories, Columbia University, Dobbs Ferry, New York

The equation for the small disturbances from the plane-parallel flow of a thermally stratified fluid under the influence of gravity acting perpendicular to the plane of stratification is derived. It was found necessary to include not only viscosity but also heat conductivity to preclude the resulting differential equation from having a singularity. Asymptotic solutions of the sixth-order differential equation thus derived are obtained. They show the presence of a Stokes point. The limiting form of the differential equation near the Stokes point is next obtained and an exact solution of this equation is derived by means of a Laplace transformation. In the general case the integrand of the Laplace transformation involves Whittaker's confluent hypergeometric functions. In the special case of a Prandtl number of 1, the integrand is considerably simpler and for this case asymptotic representations of the solutions on both sides of the Stokes point have been derived from the Laplace transformation solution by the method of steepest descent. The connection formulas between the solutions are the same as that previously derived by Tollmien and Lin for the case when stratification, gravity, and heat conduction are neglected.

I. INTRODUCTION

THE problem of the stability of the parallel flows of a fluid, neglecting viscosity, was first considered by Lord Rayleigh.[1] He found the following differential equation for the small disturbance:

$$(U + n/k)(d^2w/dz^2 - k^2w) - (d^2U/dz^2)w = 0, \quad (1)$$

where $U(z)$ is the velocity profile of the undisturbed flow, which is assumed to take place along the x direction, and w is that velocity component of the small disturbance that is perpendicular to the undisturbed flow. The small disturbance is assumed to be of the form of a function of z times the quantity $e^{i(kx+nt)}$. Equation (1) has a singularity whenever $(U + n/k) = 0$. The physical significance of Eq. (1) was put in doubt by Lord Rayleigh's further discovery[2] that indeed $(U + n/k)$ must be put equal to zero at some point within the domain of the flow in order to satisfy the boundary conditions in many important physical problems. This implies that u, the velocity component of the small disturbance that is parallel to the undisturbed flow, becomes infinite at the singular plane. In any real fluid, viscosity will intervene to prevent this from happening. It has been found necessary to invoke viscosity in order to get a physically meaningful small-disturbance equation. This equation now has the form

$$(U + n/k)(w'' - k^2w) - U''w = -(i\nu/k)(w'''' - 2k^2w'' + k^4w), \quad (2)$$

where ν is the kinematic viscosity and primes denote differentiation with respect to the coordinate z. The singularity has been removed from the problem. The plane $(U + n/k) = 0$ still has some significance in connection with the asymptotic solutions of Eq. (2), because it is a Stokes point for such solutions. At such a point the asymptotic representation of a given solution of Eq. (2) may change. The key point is then to determine the connection between the various asymptotic solutions as the Stokes point is crossed. This has been done for Eq. (2) by Tollmien[3] and Lin.[4]

If the fluid whose flow is being considered is now assumed not to be homogeneous and incompressible as above, but to have a density stratification, even though it is still assumed to be incompressible, the small-disturbance equation becomes

$$\left(U + \frac{n}{k}\right)\left(\frac{d^2w}{dz^2} - k^2w\right) - U''w + \frac{1}{\sigma}\frac{d\sigma}{dz}\left(U + \frac{n}{k}\right)\frac{dw}{dz} - \frac{1}{\sigma}\frac{d\sigma}{dz}\frac{dU}{dz}w - \frac{g}{(U + n/k)}\frac{1}{\sigma}\frac{d\sigma}{dz}w = 0. \quad (3)$$

Here the z axis is taken to be in the vertical direction and the density $\sigma(z)$ as well as the undisturbed

* Hudson Laboratories, Columbia University Contribution No. 105.

† This work was supported by the Office of Naval Research under Contract Nonr-266(84).

[1] Lord Rayleigh, Proc. London Math. Soc. 11, 57 (1880). [Reprinted in *Scientific Papers* (Cambridge University Press, Cambridge, England, 1899), Vol. 1, p. 474.]

[2] Lord Rayleigh, Phil. Mag. 26, 1001 (1913). [Reprinted in *Scientific Papers* (Cambridge University Press, Cambridge, England, 1920), Vol. VI, p. 197.]

[3] W. Tollmien, Nachr. Ges. Wiss. Göttingen (Math. Phys. Kl.) 21 (1929). (English transl.: U. S. Natl. Adv. Comm. Aeronautics Technical Memorandum No. 609, 1931.)

[4] C. C. Lin, Quart. Appl. Math. 3, 117, 218, 277 (1945–46).

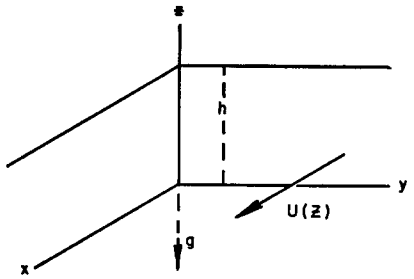


FIG. 1. Coordinate system for a flowing stratified fluid under the action of gravity.

velocity $U(z)$ depends only on the vertical coordinate. The acceleration due to gravity is taken as of magnitude g and is supposed to act in the negative z direction, while w is the vertical component of the velocity of the small disturbance.

The presence of the gravity term in Eq. (3) gives it a worse singularity than Eq. (1). Moreover, the inclusion of viscosity does not now eliminate the singularity. The small-disturbance equation, including viscosity, as used by Schlichting[5] is

$$\sigma\left(U+\frac{n}{k}\right)\left(\frac{d^2w}{dz^2}-k^2w\right)-\sigma U''w + \frac{d\sigma}{dz}\left(U\frac{dw}{dz}-\frac{dU}{dz}w\right) + \frac{n}{k}\frac{d\sigma}{dz}\frac{dw}{dz}-\frac{g}{(U+n/k)}\frac{d\sigma}{dz}w = \frac{\mu}{ik}\left(\frac{d^4w}{dz^4}-2k^2\frac{d^2w}{dz^2}+k^4w\right), \quad (4)$$

where μ is the viscosity.[6] The singularity of this equation is not as bad as that of Eq. (3) but it is still present. Attempts to deal directly with the singularity occurring in the equation for the inviscid, stratified fluid have been made.[7,8]

If the variable density of the fluid is supposed due to a temperature gradient acting on a fluid with a nonvanishing coefficient of thermal expansion, then it will be shown that the inclusion of a nonvanishing thermal conductivity leads to a small-disturbance equation of the sixth order, which is free of singularity. Likewise if the variable density were due to one substance being dissolved in another, say salt in water, then the diffusivity of the salt in the water could be used instead of the thermal conductivity. Once again the plane $(U + n/k) = 0$ has significance as a Stokes point for the asymptotic solutions of the above-mentioned sixth-order differential equation.

The aim of the present paper is to derive the small-disturbance equation including both the effects of viscosity and heat conductivity, to determine asymptotic solutions of this equation, and to find the connection formulas for the asymptotic solutions across the Stokes point for the special case when the Prandtl number is equal to one (the Prandtl number is the ratio of the kinematic viscosity to the thermometric conductivity).

II. DERIVATION OF THE SMALL-DISTURBANCE EQUATION

The small disturbances of the following situation are considered: a fluid is flowing steadily in the x direction with a velocity $U(z)$ that is a function only of the vertical coordinate z. The density $\sigma(z)$ is constant in any horizontal plane but depends on the vertical coordinate because the temperature $\Theta(z)$ is a function of that coordinate. The acceleration of gravity is of magnitude g and acts vertically downward. The boundaries of the fluid are at $z = 0$ and $z = h$ (see Fig. 1). The fluid is assumed to be incompressible. The Navier–Stokes equations of motion are

$$\rho[\partial\mathbf{v}/\partial t + (\mathbf{v}\cdot\nabla)\mathbf{v}] = -\text{grad}\, p + \rho\mathbf{g} + \mu\nabla^2\mathbf{v},$$

where ρ, $\mathbf{v}$ and p are the actual density, velocity and pressure of the fluid. These equations reduce to the following system for the undisturbed motion:

$$-\frac{\partial p_0}{\partial x}+\mu\frac{d^2U}{dz^2}=0,\quad \frac{\partial p_0}{\partial y}=0,\quad -\frac{\partial p_0}{\partial z}-\sigma g=0, \quad (5)$$

where p_0 is the static pressure. Let the velocity $\mathbf{v}$ have the components $U + u$, v, and w, the pressure be $p = p_0 + p'$, and the density be $\rho = \sigma + \rho'$. The explicit form of the equations of motion is then

$$(\sigma+\rho')\left[\frac{\partial u}{\partial t}+(U+u)\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+w\frac{\partial}{\partial z}(U+u)\right] = -\frac{\partial p_0}{\partial x}-\frac{\partial p'}{\partial x}+\mu\frac{d^2U}{dz^2}+\mu\nabla^2u,$$

$$(\sigma+\rho')\left[\frac{\partial v}{\partial t}+(U+u)\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+w\frac{\partial v}{\partial z}\right] = -\frac{\partial p_0}{\partial y}-\frac{\partial p'}{\partial y}+\mu\nabla^2v,$$

[5] H. Schlichting, Z. für Ang. Math., Mech., 15, 313, 1935. (English transl.: NACA Technical Memorandum 1262, 1950.)

[6] This is Schlichting's small-disturbance equation transcribed into the present notation and neglecting a term included by Schlichting because he considers the viscosity to be variable. See Footnote 2 of Schlichting's paper.

[7] A. Eliassen, E. Høiland, and E. Riis, Publication No. 1, Institute for Weather and Climate Research, The Norwegian Academy of Sciences (1953).

[8] L. A. Dikii, PMM 24, 249 (1960) [English transl.: J. Appl. Math. Mech. 24, 357 (1960).]

$$(\sigma + \rho')\left[\frac{\partial w}{\partial t} + (U + u)\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} + w\frac{\partial w}{\partial z}\right]$$
$$= -\frac{\partial p_0}{\partial z} - \frac{\partial p'}{\partial z} - (\sigma + \rho')g + \mu\nabla^2 w.$$

The velocities u, v, and w are assumed to be small compared with U, while the pressure p' and the density ρ' are assumed to be small compared with p_0 and σ, respectively. Under these conditions, linearization of the above equations yields the set

$$\sigma\left(\frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} + w\frac{dU}{dz}\right) = -\frac{\partial p'}{\partial x} + \mu\nabla^2 u,$$
$$\sigma\left(\frac{\partial v}{\partial t} + U\frac{\partial v}{\partial x}\right) = -\frac{\partial p'}{\partial y} + \mu\nabla^2 v, \qquad (6)$$
$$\sigma\left(\frac{\partial w}{\partial t} + U\frac{\partial w}{\partial x}\right) = -\frac{\partial p'}{\partial z} - \rho' g + \mu\nabla^2 w,$$

where use has been made of Eq. (5) to cancel out the lowest-order terms.

The equation of state of the fluid is taken to be

$$\rho = \rho_0[1 - \alpha(\theta - \theta_0)], \qquad (7)$$

where θ denotes the actual temperature of the fluid at a given goint, α is the coefficient of thermal expansion, and ρ_0 denotes the density of the fluid at the standard temperature θ_0. Applied to the undisturbed situation the equation of state becomes

$$\sigma = \rho_0[1 - \alpha(\Theta - \theta_0)]. \qquad (8)$$

Let θ' denote the departure of the temperature from the value Θ, so that

$$\theta = \Theta + \theta'. \qquad (9)$$

Substitution in Eq. (7) yields

$$\sigma + \rho' = \rho_0[1 - \alpha(\Theta + \theta' - \theta_0)].$$

Using Eq. (8), several terms can be canceled, giving

$$\rho' = -\alpha\rho_0\theta'. \qquad (10)$$

Now substitute this in the last of Eq. (6):

$$\sigma\left(\frac{\partial w}{\partial t} + U\frac{\partial w}{\partial x}\right) = -\frac{\partial p'}{\partial z} + \alpha g\rho_0\theta' + \mu\nabla^2 w.$$

The fundamental approximation of Boussinesq will be made[9]: the variation of density is neglected in all terms in the equations of motion except the term involving gravity. It cannot be neglected in the latter term because it is precisely this term that may give rise to the stability or instability, but if the temperature difference between the boundaries is not very great, neither will the density difference be very great and it may be safely neglected. Let the standard temperature θ_0 be taken between the temperatures of the boundaries. Then we may approximately identify σ with the constant ρ_0. The Eqs. (6) become

$$\frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} + w\frac{dU}{dz} = -\frac{1}{\rho_0}\frac{\partial p'}{\partial x} + \nu\nabla^2 u,$$
$$\frac{\partial v}{\partial t} + U\frac{\partial v}{\partial x} = -\frac{1}{\rho_0}\frac{\partial p'}{\partial y} + \nu\nabla^2 v, \qquad (11)$$
$$\frac{\partial w}{\partial t} + U\frac{\partial w}{\partial x} = -\frac{1}{\rho_0}\frac{\partial p'}{\partial z} + \alpha g\theta' + \nu\nabla^2 w,$$

where $\nu = \mu/\rho_0$. The equation for the conduction of heat may be written

$$\partial\theta/\partial t + (\mathbf{v}\cdot\nabla)\theta = \kappa\nabla^2\theta,$$

where κ is the thermomentric conductivity, and both the generation of heat by viscosity and the work done in the expansion and contraction of the fluid elements have been neglected. This equation reduces to the following form for the undisturbed flow:

$$\nabla^2\Theta = 0. \qquad (12)$$

In the general case it may be written

$$\frac{\partial\theta'}{\partial t} + (U + u)\frac{\partial\theta'}{\partial x} + v\frac{\partial\theta'}{\partial y} + w\frac{\partial}{\partial z}(\Theta + \theta')$$
$$= \kappa\nabla^2\Theta + \kappa\nabla^2\theta'.$$

The linearization of this equation is

$$\frac{\partial\theta'}{\partial t} + U\frac{\partial\theta'}{\partial x} + w\frac{d\Theta}{dz} = \kappa\nabla^2\theta', \qquad (13)$$

where use has been made of Eq. (12).

Finally the equation of continuity is

$$\frac{\partial\rho}{\partial t} + \operatorname{div}(\rho\mathbf{v}) = 0.$$

Written out in full this becomes

$$\frac{\partial\rho'}{\partial t} + (\sigma + \rho')\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right)$$
$$+ (U + u)\frac{\partial\rho'}{\partial x} + v\frac{\partial\rho'}{\partial y} + w\left(\frac{\partial\sigma}{\partial z} + \frac{\partial\rho'}{\partial z}\right) = 0,$$

with the subsequent linearization:

$$\frac{\partial\rho'}{\partial t} + \sigma\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) + U\frac{\partial\rho'}{\partial x} + w\frac{d\sigma}{dz} = 0. \qquad (14)$$

Take the derivative of Eq. (8):

$$d\sigma/dz = -\alpha\rho_0\, d\Theta/dz.$$

[9] Lord Rayleigh, Phil. Mag. 32, 529 (1916). [Reprinted in *Scientific Papers* (Cambridge University Press, Cambridge, England, 1920), Vol. VI, p. 432.]

Now multiply Eq. (13) by $(-\alpha\rho_0)$, and after using Eq. (10), substract it from Eq. (14):

$$\sigma(\partial u/\partial x + \partial v/\partial y + \partial w/\partial z) = -\kappa\nabla^2\rho'.$$

Using Eq. (10) again this becomes

$$\sigma(\partial u/\partial x + \partial v/\partial y + \partial w/\partial z) = \alpha\rho_0\kappa\nabla^2\theta'.$$

Now the approximation of Boussinesq involves putting $\alpha = 0$ but $\alpha g \neq 0$, since the change in density, which is assumed to be small, is proportional to α. This allows the right-hand side of the above equation to be neglected. The form of the equation of continuity used is thus

$$\partial u/\partial x + \partial v/\partial y + \partial w/\partial z = 0. \tag{15}$$

The basic equations are (11), (13), and (15). Eliminate the variables u and v by differentiating the first of Eqs. (11) with respect to x, the second of Eqs. (11) with respect to y, adding, and making use of Eq. (15):

$$-\frac{\partial^2 w}{\partial z\,\partial t} - U\frac{\partial^2 w}{\partial x\,\partial z} + \frac{dU}{dz}\frac{\partial w}{\partial x} = -\frac{1}{\rho_0}\left(\frac{\partial^2 p'}{\partial x^2} + \frac{\partial^2 p'}{\partial y^2}\right) - \nu\nabla^2\frac{\partial w}{\partial z}. \tag{16}$$

The notation

$$\nabla_1^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$$

is used hereafter. Operate on the last of Eqs. (11) with ∇_1^2 to give

$$\frac{\partial}{\partial t}\nabla_1^2 w + U\frac{\partial}{\partial x}\nabla_1^2 w = -\frac{1}{\rho_0}\frac{\partial}{\partial z}\nabla_1^2 p' + \alpha g\nabla_1^2\theta' + \nu\nabla^2\nabla_1^2 w.$$

Now the pressure p' can be eliminated between this equation and Eq. (16):

$$\frac{\partial}{\partial t}\nabla_1^2 w + U\frac{\partial}{\partial x}\nabla_1^2 w = \alpha g\nabla_1^2\theta' + \nu\nabla^2\nabla_1^2 w + \frac{\partial}{\partial z}\left[\nu\nabla^2\frac{\partial w}{\partial z} - \frac{\partial^2 w}{\partial z\,\partial t} - U\frac{\partial^2 w}{\partial x\,\partial z} + \frac{dU}{dz}\frac{\partial w}{\partial x}\right].$$

This can be rearranged to become

$$\frac{\partial}{\partial t}\nabla^2 w + U\frac{\partial}{\partial x}\nabla^2 w - \frac{d^2U}{dz^2}\frac{\partial w}{\partial x} = \alpha g\nabla_1^2\theta' + \nu\nabla^4 w. \tag{17}$$

Operate with $\alpha g\nabla_1^2$ on Eq. (13):

$$\frac{\partial}{\partial t}(\alpha g\nabla_1^2\theta') + U\frac{\partial}{\partial x}(\alpha g\nabla_1^2\theta') + \alpha g\frac{d\Theta}{dz}\nabla_1^2 w = \kappa\nabla^2(\alpha g\nabla_1^2\theta').$$

The temperature θ' can be eliminated between the above equation and Eq. (17):

$$\frac{\partial}{\partial t}\left[\frac{\partial}{\partial t}\nabla^2 w + U\frac{\partial}{\partial x}\nabla^2 w - U''\frac{\partial w}{\partial x} - \nu\nabla^4 w\right] + U\frac{\partial}{\partial x}\left[\frac{\partial}{\partial t}\nabla^2 w + U\frac{\partial}{\partial x}\nabla^2 w - U''\frac{\partial w}{\partial x} - \nu\nabla^4 w\right] + \alpha g\frac{d\Theta}{dz}\nabla_1^2 w = \kappa\nabla^2\left[\frac{\partial}{\partial t}\nabla^2 w + U\frac{\partial}{\partial x}\nabla^2 w - U''\frac{\partial w}{\partial x} - \nu\nabla^4 w\right].$$

This equation can be rewritten as

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)\left[\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)\nabla^2 w - U''\frac{\partial w}{\partial x} - \nu\nabla^4 w\right] + \alpha g\frac{d\Theta}{dz}\nabla_1^2 w = \kappa\nabla^2\left[\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)\nabla^2 w - U''\frac{\partial w}{\partial x} - \nu\nabla^4 w\right].$$

The identity

$$\nabla^2(fg) = f\nabla^2 g + g\nabla^2 f + 2\left(\frac{\partial f}{\partial x}\frac{\partial g}{\partial x} + \frac{\partial f}{\partial y}\frac{\partial g}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial g}{\partial z}\right),$$

where f and g are arbitrary functions of the coordinates, allows the above formula to be written more explicitly:

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)\left[\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)\nabla^2 w - U''\frac{\partial w}{\partial x} - (\kappa+\nu)\nabla^4 w\right] + \alpha g\frac{d\Theta}{dz}\nabla_1^2 w = -\kappa\nu\nabla^6 w + 2\kappa U'\nabla^2\frac{\partial^2 w}{\partial x\,\partial z} - 2\kappa U'''\frac{\partial^2 w}{\partial x\,\partial z} - \kappa U''''\frac{\partial w}{\partial x}. \tag{18}$$

This is a partial differential equation whose coefficients are independent of x, y, and t. Separation of variables is possible and the vertical velocity w will be assumed to have the form

$$e^{i(kx+ly+nt)}$$

times a function of z. The operator ∇_1^2 becomes

$$\nabla_1^2 = -(k^2 + l^2),$$

and

$$(\partial/\partial t + U\,\partial/\partial x) = i(n + kU).$$

Equation (18) becomes

$$i(n + kU)[i(n + kU)\nabla^2 w - ikU''w - (\kappa + \nu)\nabla^4 w] - \alpha g(d\Theta/dz)(k^2 + l^2)w = -\kappa\nu\nabla^6 w + 2ik\kappa U'\nabla^2 \frac{dw}{dz} - 2ik\kappa U''' \frac{dw}{dz} - ik\kappa U''''w. \quad (19)$$

Also

$$\nabla^2 = d^2/dz^2 - (k^2 + l^2).$$

Putting

$$c = -n/k \quad \text{and} \quad \beta = d\Theta/dz,$$

Eq. (19) can be written

$$(U - c)\left[(U - c)\nabla^2 w - U''w + \frac{i(\kappa + \nu)}{k}\nabla^4 w\right] + \frac{\alpha\beta g(k^2 + l^2)}{k^2} w = \frac{\kappa\nu}{k^2}\nabla^6 w - \frac{2i\kappa U'}{k}\nabla^2 \frac{dw}{dz} + \frac{2i\kappa U'''}{k}\frac{dw}{dz} + \frac{i\kappa}{k} U''''w. \quad (20)$$

When $\kappa = 0$, $l = 0$, and the identification

$$d\sigma/dz = -\alpha\rho_0 \, d\Theta/dz$$

is made, the above equation reduces to

$$(U - c)\left[(U - c)\nabla^2 w - U''w + \frac{i\nu}{k}\nabla^4 w\right] - \frac{g}{\rho_0}\frac{d\sigma}{dz} w = 0,$$

which is to be compared with Eq. (4), when $d\sigma/dz$ is neglected in the latter, except in the term involving the gravity. Next put $U = 0$ in Eq. (20):

$$c^2\nabla^2 w - \frac{i(\kappa + \nu)c}{k}\nabla^4 w + \frac{\alpha\beta g(k^2 + l^2)}{k^2} w - \frac{\kappa\nu}{k^2}\nabla^6 w = 0,$$

which becomes

$$\left[\left(c - \frac{i\nu}{k}\nabla^2\right)\left(c - \frac{i\kappa}{k}\nabla^2\right)\nabla^2 + \frac{\alpha\beta g(k^2 + l^2)}{k^2}\right] w = 0.$$

This is the usual form of small-disturbance equation for a heated layer of fluid.[10]

If the velocity profile $U(z)$ is symmetric around some plane, then all the even-order derivatives of U are symmetric and all the odd-order derivatives are antisymmetric around that plane. Equation (20) contains even-order derivatives of w in conjunction with even-order derivatives of U and odd-order derivatives of w in conjunction with odd-order derivatives of U. Therefore, it is invariant to the substitution $z \to -z$, provided that β is also symmetric with respect to the given plane, which is taken as $z = 0$. If the boundary conditions are also symmetric all solutions may be classified into even and odd solutions.

III. THE BOUNDARY CONDITIONS. SQUIRE'S THEOREM

At a rigid wall the boundary conditions are that the fluid adhere to the wall, or symbolically $u = v = w = 0$, since the walls are assumed to be at rest. Since this is true for all x and y, it follows that

$$\partial u/\partial x = \partial v/\partial y = 0$$

at the wall, and from Eq. (15),

$$\partial w/\partial z = 0.$$

The behavior of the temperature at the boundaries must also be prescribed. The simplest situation is one where the temperature at the boundary is held fixed, say by contact with a heat reservoir. This gives $\theta' = 0$, independent of x and y, so that $\nabla_1^2\theta' = 0$. Equation (17) gives

$$(\partial/\partial t)\nabla^2 w - \nu\nabla^4 w = 0,$$

when account is taken of the fact that U and $\partial w/\partial x$ vanish at the boundary. The following equations summarize the boundary conditions for a rigid wall maintained at a constant temperature:

$$w = \partial w/\partial z = (\partial/\partial t)\nabla^2 w - \nu\nabla^4 w = 0. \quad (21)$$

At a free surface, upon which there are assumed to be no surface waves, we still have $w = 0$, but now

$$T_{zz} = \mu(\partial u/\partial z + \partial w/\partial x),$$

$$T_{zy} = \mu(\partial v/\partial z + \partial w/\partial y),$$

which are components of the stress tensor, must vanish for all x and y. This leads to the equations

$$\partial T_{zz}/\partial x = \partial T_{zy}/\partial y = 0,$$

which, when combined with

$$\partial^2 w/\partial x^2 = \partial^2 w/\partial y^2 = 0,$$

give

$$\partial^2 u/\partial z\,\partial x = \partial^2 v/\partial z\,\partial y = 0.$$

[10] A. Pellew and R. V. Southwell, Proc. Roy. Soc. (London) **A176**, 312 (1940).

Differentiating the equation of continuity (15) with respect to z we see that

$$\partial^2 w/\partial z^2 = 0.$$

When combined with $w = 0$, this gives $\nabla^2 w = 0$ on the boundary. If the same behavior of the temperature at the boundary is assumed as before, Eq. (17) gives

$$\nabla^4 w = 0.$$

The following equations summarize the boundary conditions for a free surface, with no surface waves, maintained at a constant temperature:

$$w = \partial^2 w/\partial z^2 = \partial^4 w/\partial z^4 = 0. \tag{22}$$

Consider the case of a parabolic velocity profile

$$U(z) = (4U_0/h^2)z(h - z),$$

where U_0 denotes the maximum velocity of the undisturbed flow. With the notation

$$a^2 = k^2 + l^2,$$

the small-disturbance equation (20) becomes

$$\begin{aligned}(U - c)\Big[(U - c)\Big(\frac{d^2}{dz^2} - a^2\Big)w + \frac{8U_0}{h^2}w \\ + \frac{i(\kappa + \nu)}{k}\Big(\frac{d^2}{dz^2} - a^2\Big)^2 w\Big] \\ + \frac{\alpha\beta g a^2}{k^2}w = \frac{\kappa\nu}{k^2}\Big(\frac{d^2}{dz^2} - a^2\Big)^3 w \\ - \frac{8i\kappa U_0}{hk}\Big(1 - 2\frac{z}{h}\Big)\cdot\Big(\frac{d^2}{dz^2} - a^2\Big)\frac{dw}{dz}.\end{aligned} \tag{23}$$

The dimensionless variables

$$\zeta = z/h \qquad b = ha$$

will be needed. The velocity profile assumes the dimensionless form

$$U/U_0 = 4(\zeta - \zeta^2).$$

The Reynolds number Re, and the Prandtl number Pr are needed:

$$\mathrm{Re} = U_0 h/\nu \qquad \mathrm{Pr} = \nu/\kappa. \tag{24}$$

In addition the Richardson number Ri is defined by the equation

$$\mathrm{Ri} = -\frac{g}{(dU/dz)^2}\frac{1}{\sigma}\frac{d\sigma}{dz}.$$

The quantity (dU/dz) varies with position, but attains its maximum value at $z = 0$, where it is given by

$$dU/dz = 4U_0/h.$$

The Richardson number for the flow as a whole is defined by inserting into the above formula the maximum value of (dU/dz). When in addition the identification

$$\alpha\beta = -\sigma^{-1}\, d\sigma/dz$$

is made, the Richardson number is defined by

$$\mathrm{Ri} = \alpha\beta g h^2/16U_0^2. \tag{25}$$

Using the quantities defined above, Eq. (23) can be written in dimensionless form

$$\begin{aligned}\Big(\frac{U}{U_0} - \frac{c}{U_0}\Big)\Big[\Big(\frac{U}{U_0} - \frac{c}{U_0}\Big)\Big(\frac{d^2}{d\zeta^2} - b^2\Big)w + 8w \\ + \frac{i}{hk\,\mathrm{Re}}\Big(1 + \frac{1}{\mathrm{Pr}}\Big)\Big(\frac{d^2}{d\zeta^2} - b^2\Big)^2 w\Big] \\ + \frac{16\,\mathrm{Ri}\,b^2}{h^2k^2}w = \frac{1}{\mathrm{Pr}\,h^2k^2\,\mathrm{Re}^2}\Big(\frac{d^2}{d\zeta^2} - b^2\Big)^3 w \\ - \frac{8i}{\mathrm{Pr}\,hk\,\mathrm{Re}}(1 - \zeta)\Big(\frac{d^2}{d\zeta^2} - b^2\Big)\frac{dw}{d\zeta}.\end{aligned} \tag{26}$$

Separating variables in the boundary conditions (21) gives

$$w = \frac{dw}{dz} = in\Big(\frac{d^2}{dz^2} - a^2\Big)w - \nu\Big(\frac{d^2}{d\zeta^2} - a^2\Big)^2 w = 0.$$

These equations can also be written in dimensionless form

$$\begin{aligned}w = \frac{dw}{d\zeta} = i\Big(\frac{c}{U_0}\Big)\Big(\frac{d^2}{d\zeta^2} - b^2\Big)w \\ + \frac{1}{hk\,\mathrm{Re}}\Big(\frac{d^2}{d\zeta^2} - b^2\Big)^2 w = 0.\end{aligned} \tag{27}$$

Now define the quantities

$$\mathrm{Re}^* = hk\,\mathrm{Re}/b, \qquad \mathrm{Ri}^* = \mathrm{Ri}\,b^2/h^2k^2.$$

Note that $\mathrm{Re}^* \leq \mathrm{Re}$ and $\mathrm{Ri}^* \geq \mathrm{Ri}$. In terms of these quantities the small-disturbance equation (26) becomes

$$\begin{aligned}\Big(\frac{U}{U_0} - \frac{c}{U_0}\Big)\Big[\Big(\frac{U}{U_0} - \frac{c}{U_0}\Big)\Big(\frac{d^2}{d\zeta^2} - b^2\Big)w \\ + 8w + \frac{i}{b\,\mathrm{Re}^*}\Big(1 + \frac{1}{\mathrm{Pr}}\Big)\Big(\frac{d^2}{d\zeta^2} - b^2\Big)^2 w\Big] \\ + 16\,\mathrm{Ri}^*\,w = \frac{1}{\mathrm{Pr}\,b^2\mathrm{Re}^{*2}}\Big(\frac{d^2}{d\zeta^2} - b^2\Big)^3 w \\ - \frac{8i}{\mathrm{Pr}\,b\,\mathrm{Re}^*}(1 - \zeta)\Big(\frac{d^2}{d\zeta^2} - b^2\Big)\frac{dw}{d\zeta},\end{aligned} \tag{28}$$

while the boundary conditions (27) reduce to

$$w = \frac{dw}{d\zeta} = i\left(\frac{c}{U_0}\right)\left(\frac{d^2}{d\zeta^2} - b^2\right)w + \frac{1}{b\,\mathrm{Re}^*}\left(\frac{d^2}{d\zeta^2} - b^2\right)^2 w = 0. \quad (29)$$

Equations (28) and (29) are the small-disturbance equation and boundary conditions for the case when the small disturbance is assumed to have no component of velocity in the y direction and all quantities are assumed to be independent of the y coordinate. For this identification to hold, the parameters Re* and Ri* must be identified with the Reynolds number and Richardson number of the two-dimensional flow through Eqs. (24) and (25). Hence, we have the following generalization of Square's theorem[11,12]: the three-dimensional small-disturbance problem is equivalent to a two-dimensional problem with a smaller Reynolds number and a larger Richardson number. In this comparison the Prandtl number and magnitude of the wave number must be unchanged. In view of this theorem attention will henceforth be confined to two-dimensional disturbances.

A small-disturbance equation for the temperature θ' can be obtained by eliminating w between Eqs. (13) and (17). The result, assuming a constant temperature gradient β, and only two-dimensional motion $l = 0$, is

$$\begin{aligned}(U-c)\Big[&(U-c)\nabla^2\theta' + 2U'\frac{d\theta'}{dz} + \frac{i(\nu+\kappa)}{k}\nabla^4\theta'\Big] + \alpha\beta g\theta' \\ &= \frac{\kappa\nu}{k^2}\nabla^6\theta' - \frac{4i\nu U'}{k}\frac{d}{dz}\nabla^2\theta' + \frac{i\kappa U''}{k}\nabla^2\theta' \\ &\quad - \frac{i\nu}{k}\Big[6U''\frac{d^2\theta'}{dz^2} + 4U'''\frac{d\theta'}{dz} + U''''\theta' - 2k^2U''\theta'\Big]. \end{aligned} \quad (30)$$

IV. ASYMPTOTIC SOLUTIONS

Solutions of the small-disturbance equation valid for large Reynolds numbers (vanishing viscosity) are of interest. This is because in any stability calculation where the density gradient is stabilizing, the instability would first enter at rather large values of the Reynolds number. The following trial function is used for substitution in the differential equation:

$$W = A(z)\exp\left[(k/i\nu)^{\frac{1}{2}}B(z)\right]. \quad (31)$$

First put $l = 0$ in Eq. (20) and then substitute the above expression in it. Equating to zero the terms of order $(k/i\nu)$ gives

$$(U-c)^2AB'^2 + \frac{i(\kappa+\nu)}{k}\left(\frac{k}{i\nu}\right)AB'^4(U-c) = \frac{\kappa\nu}{k^2}\left(\frac{k}{i\nu}\right)^2 AB'^6.$$

Supposing that $A \neq 0$, this becomes

$$B'^6 + (1+\mathrm{Pr})(U-c)B'^4 + \mathrm{Pr}\,(U-c)B'^2 = 0.$$

The following solutions of this equation exist:

$$\text{(i)}\ B'^2 = 0, \quad \text{(ii)}\ B'^2 = -(U-c), \quad \text{(iii)}\ B'^2 = -\mathrm{Pr}\,(U-c). \quad (32)$$

Now equate to zero the terms of order $(k/i\nu)^{\frac{1}{2}}$:

$$\begin{aligned}[2\,\mathrm{Pr}\,B'(U-c)^2 &+ (1+\mathrm{Pr})(U-c)4B'^3 + 6B'^5]A'/A \\ &= -15B'^4B'' - U'B'^3 - (U-c)^2B''\,\mathrm{Pr} \\ &\quad - 6(1+\mathrm{Pr})(U-c)B'^2B''.\end{aligned}$$

If $B' = 0$ this equation is identically satisfied. The case $B'^2 = -(U-c)$ leads to the following equation:

$$(1-\mathrm{Pr})A'/A = -(1-\mathrm{Pr})\tfrac{5}{2}B''/B'.$$

If $\mathrm{Pr} = 1$ this equation is identically satisfied. Otherwise it can be integrated to obtain the relations

$$A = (B')^{-5/2}, \qquad A \propto (U-c)^{-5/4}.$$

The case $B'^2 = -\mathrm{Pr}\,(U-c)$ leads to the equation

$$(1-\mathrm{Pr})A'/A = -\tfrac{3}{2}(1-\mathrm{Pr})B''/B'.$$

It also is identically satisfied if $\mathrm{Pr} = 1$, and likewise can be integrated if $\mathrm{Pr} \neq 1$:

$$A = (B')^{-9/2}, \qquad A \propto (U-c)^{-9/4}.$$

When $\mathrm{Pr} \neq 1$ four asymptotic solutions have been determined above:

$$\begin{aligned}\psi_3 &= (U-c)^{-5/4}\exp\left[-\left(\frac{ik}{\nu}\right)^{\frac{1}{2}}\int (U-c)^{\frac{1}{2}}\,dz\right], \\ \psi_4 &= (U-c)^{-5/4}\exp\left[+\left(\frac{ik}{\nu}\right)^{\frac{1}{2}}\int (U-c)^{\frac{1}{2}}\,dz\right], \\ \psi_5 &= (U-c)^{-9/4}\exp\left[-\left(\frac{ik}{\kappa}\right)^{\frac{1}{2}}\int (U-c)^{\frac{1}{2}}\,dz\right], \\ \psi_6 &= (U-c)^{-9/4}\exp\left[+\left(\frac{ik}{\kappa}\right)^{\frac{1}{2}}\int (U-c)^{\frac{1}{2}}\,dz\right].\end{aligned} \quad (33)$$

Since for two cases the equation determined by

[11] H. B. Squire, Proc. Roy. Soc. (London) **A142**, 621 (1933).
[12] C. Yih, Quart. Appl. Math. **12**, 434 (1955).

letting the terms of order-of-magnitude $(k/i\nu)^{\frac{1}{2}}$ vanish is identically satisfied, we must go to terms of order (1) in Eq. (20) for these cases. When $B'^2 = 0$ this gives

$$(U - c)[(U - c)(A'' - k^2A) - U''A] + \alpha\beta gA = 0.$$

This is the small-disturbance equation neglecting viscosity and heat conductivity. The two solutions determined from it will be called ψ_1 and ψ_2.

The case $\Pr = 1$, $B'^2 = -(U - c)$ leads to the following equation in third order:

$$4B'^4A'' + 24B'^3B''A' + [\alpha\beta g + 25B'^2B''^2 + 10B'^3B''']A = 0.$$

The trial solution $A = (B')^s$ is used for this equation. Two cases are solvable by means of this trial solution. First put $\alpha\beta g = 0$ and substitute the trial function in the resulting equation:

$$2(2s + 5)B'^3B''' + (2s + 5)^2B'^2B''^2 = 0.$$

Hence if

$$2s + 5 = 0, \qquad s = -\tfrac{5}{2},$$

the equation is identically satisfied. This gives

$$A = (B')^{-5/2}$$

as a solution when $\alpha\beta g = 0$.

When $\alpha\beta g \neq 0$, substituting the trial function in the differential equation gives

$$(4s + 10)B'^3B''' + (4s^2 + 20s + 25)B'^2B''^2 + \alpha\beta g = 0.$$

The case of a linear velocity profile will be treated. Differentiating the second of Eqs. (32) gives

$$B'B'' = -\tfrac{1}{2}U', \qquad B''^2 + B'B''' = 0,$$

in that case. Using these formulas in the above equation gives

$$(4s^2 + 16s + 15) + 4\alpha\beta g/(U')^2 = 0.$$

Since U' is a constant, this equation can be satisfied by an appropriate choice of s. Defining the Richardson number

$$\mathrm{Ri} = \alpha\beta g/(U')^2,$$

we get the following quadratic equation for s:

$$4s^2 + 16s + (15 + 4\,\mathrm{Ri}) = 0.$$

Note that when $\mathrm{Ri} = \frac{1}{4}$ this quadratic equation has a double root. When $\mathrm{Ri} < \frac{1}{4}$ the roots are real and when $\mathrm{Ri} > \frac{1}{4}$ the roots are complex.

The solutions (33) have been determined to the second order. However in order to determine the temperature θ' to second order from the vertical velocity w, it will be found necessary to determine the latter quantity to third order. Assume a trial function

$$w = \left[A + A_1\left(\frac{i\nu}{k}\right)^{\frac{1}{2}}\right] \exp\left[\left(\frac{k}{i\nu}\right)^{\frac{1}{2}} B\right], \tag{34}$$

where A and B are given by the second-order theory. Then for $B'^2 = -(U - c)$, $\Pr \neq 1$, third-order theory gives the equation

$$\frac{75}{4}\left(\frac{1}{\Pr} - 1\right)B'^2B''^2 + \frac{13}{2}\left(1 - \frac{1}{\Pr}\right)B'^3B''' + \left(1 - \frac{1}{\Pr}\right)k^2B'^4 + \alpha\beta g + 2\left(\frac{1}{\Pr} - 1\right)B'^5\frac{d}{dz}\left(\frac{A_1}{A}\right) = 0. \tag{35}$$

Since A and B' are known, this equation allows the determination of A_1 by a quadrature.

When $B'^2 = -\Pr\,(U - c)$, $\Pr \neq 1$, the following equation is obtained:

$$\left(195 - \frac{131}{\Pr}\right)\frac{B'^2B''^2}{4} + \left(\frac{33}{\Pr} - 49\right)\frac{B'^3B'''}{2} + \alpha\beta g\,\Pr + \left(\frac{1}{\Pr} - 1\right)k^2B'^4 + 2\left(1 - \frac{1}{\Pr}\right)B'^5\frac{d}{dz}\left(\frac{A_1}{A}\right) = 0.$$

Now the temperature θ' can be obtained. Separating variables in Eqs. (13) and (17) gives us a pair of equations connecting θ' and w:

$$\begin{aligned} ik(U - c)\theta' + \beta w &= \kappa\nabla^2\theta', \\ ik(U - c)\nabla^2 w - ikU''w &= -\alpha gk^2\theta' + \nu\nabla^4 w, \end{aligned} \tag{36}$$

where l has been set equal to zero. Corresponding to the inviscid solutions ψ_1 and ψ_2 for w there are two inviscid solutions ϕ_1 and ϕ_2 for θ'. These are obtained by putting $\nu = \kappa = 0$ in the above equations. This gives

$$\phi_1 = \frac{i\beta}{k(U - c)}\psi_1, \qquad \phi_2 = \frac{i\beta}{k(U - c)}\psi_2.$$

The solutions ϕ_5 and ϕ_6 corresponding to the solutions ψ_5 and ψ_6 are obtained by substituting formula (34) for w into Eq. (36) and retaining only the terms of highest order of magnitude, which are of order $(k/i\nu)$. The result is

$$\phi_5 = \frac{\Pr\,(1 - \Pr)}{\alpha g\nu}(U - c)^2\psi_5,$$

$$\phi_6 = \frac{\Pr\,(1 - \Pr)}{\alpha g\nu}(U - c)^2\psi_6.$$

The solutions ϕ_3 and ϕ_4 corresponding to ψ_3 and ψ_4 may be obtained in a similar manner, except that now the terms of order-of-magnitude $(k/i\nu)$ and $(k/i\nu)^{\frac{1}{2}}$ vanish. The first nonvanishing term gives the result

$$\alpha g k^2 \theta' = -ikA\left[\frac{75}{4}B''^2 - \frac{13}{2}B'B'' - k^2B'^2 + 2B'^3\frac{d}{dz}\left(\frac{A_1}{A}\right)\right]\exp\left[\left(\frac{k}{i\nu}\right)^{\frac{1}{2}}B\right].$$

Using Eq. (35) this can be reduced to

$$\theta' = \frac{i\beta}{k(1 - 1/\mathrm{Pr})}\frac{w}{(U - c)},$$

which gives

$$\phi_3 = \frac{i\beta}{k(1 - 1/\mathrm{Pr})}\frac{\psi_3}{(U - c)},$$

$$\phi_4 = \frac{i\beta}{k(1 - 1/\mathrm{Pr})}\frac{\psi_4}{(U - c)}.$$

V. BEHAVIOR OF SOLUTIONS NEAR THE STOKES POINT

The asymptotic solutions (33) are singular at a point z_c where $(U - c) = 0$, even though the original differential equation is not singular at such a point. In order to examine the solutions in more detail near this point expand the velocity profile in powers of $(z - z_c)$:

$$(U - c) = U_c'(z - z_c) + \tfrac{1}{2}U_c''(z - z_c)^2 + \cdots$$

Also the substitution

$$z - z_c = \epsilon\eta$$

will be made. The quantity ϵ will be taken as small, so that this substitution magnifies the region around $z = z_c$. Since ϵ is small we can expand w in powers of ϵ:

$$w = w_0 + w_1\epsilon + w_2\epsilon^2 + \cdots.$$

The quantities κ and ν will be assumed to be of the same order of magnitude, while ϵ^3 will be taken to be of the order of magnitude of ν. Thus indeed $\epsilon \to 0$ when the viscosity is taken as small.

Using only the first term of the above expansion of w in powers of ϵ, Eq. (20) becomes

$$(U_c')^2\eta^2\frac{d^2w}{d\eta^2} + \frac{iU_c'(\kappa + \nu)}{k\epsilon^3}\eta\frac{d^4w}{d\eta^4} + \alpha\beta g w = -\frac{2i\kappa U_c'}{k\epsilon^3}\frac{d^3w}{d\eta^3} + \frac{\kappa\nu}{k^2\epsilon^6}\frac{d^6w}{d\eta^6},$$

where l has been put equal to zero, and the subscript has been dropped on w.

Defining the following Richardson number:

$$\mathrm{Ric} = \alpha\beta g/(U_c')^2,$$

the above equation becomes

$$\eta^2\frac{d^2w}{d\eta^2} + \frac{i(\kappa + \nu)}{k\epsilon^3U_c'}\eta\frac{d^4w}{d\eta^4} + \mathrm{Ric}\, w = -\frac{2i\kappa}{k\epsilon^3U_c'}\frac{d^3w}{d\eta^3} + \frac{\kappa\nu}{k^2\epsilon^6(U_c')^2}\frac{d^6w}{d\eta^6}. \tag{37}$$

Two possible definitions of ϵ will be considered. First define it as

$$\epsilon = (\nu/kU_c')^{\frac{1}{3}},$$

so we get

$$\eta^2\frac{d^2w}{d\eta^2} + i\left(1 + \frac{1}{\mathrm{Pr}}\right)\eta\frac{d^4w}{d\eta^4} + \mathrm{Ric}\, w = \frac{1}{\mathrm{Pr}}\frac{d^6w}{d\eta^6} - \frac{2i}{\mathrm{Pr}}\frac{d^3w}{d\eta^3}. \tag{38}$$

Leaving ν fixed let $\kappa \to 0$ so that $\mathrm{Pr} \to \infty$. The limiting form of Eq. (38) is then

$$\eta^2\frac{d^2w}{d\eta^2} + i\eta\frac{d^4w}{d\eta^4} + \mathrm{Ric}\, w = 0. \tag{39}$$

This equation holds when heat conduction is neglected but viscosity is not.

Next define ϵ as

$$\epsilon = (\kappa/kU_c')^{\frac{1}{3}}.$$

Equation (37) becomes

$$\eta^2\frac{d^2w}{d\eta^2} + i(1 + \mathrm{Pr})\eta\frac{d^4w}{d\eta^4} + \mathrm{Ric}\, w = -2i\frac{d^3w}{d\eta^3} + \mathrm{Pr}\frac{d^6w}{d\eta^6}. \tag{40}$$

Leaving κ fixed let $\nu \to 0$ so that $\mathrm{Pr} \to 0$. Equation (40) becomes

$$\eta^2\frac{d^2w}{d\eta^2} + i\eta\frac{d^4w}{d\eta^4} + \mathrm{Ric}\, w = -2i\frac{d^3w}{d\eta^3}. \tag{41}$$

This equation holds when viscosity is neglected but heat conduction is not.

Equation (38) for the solutions near the Stokes point can be reduced to an equation of the second order by means of the following Laplace transformation:

$$w(\eta) = \int_a^b e^{\eta t}v(t)\,dt.$$

Substituting this expression in Eq. (38) we find the

following conditions that the Laplace transformation be a solution:

$$\frac{d^2(t^2v)}{dt^2} - i\left(1 + \frac{1}{\mathrm{Pr}}\right)\frac{d(t^4v)}{dt} + \left(\mathrm{Ric} + \frac{2i}{\mathrm{Pr}}\,t^3 - \frac{t^6}{\mathrm{Pr}}\right)v = 0, \quad (42a)$$

$$\left[t^2\eta e^{\eta t}v - e^{\eta t}\frac{d(t^2v)}{dt} + i\left(1 + \frac{1}{\mathrm{Pr}}\right)t^4e^{\eta t}v\right]_a^b = 0. \quad (42b)$$

Change the dependent variable in Eq. (42a) to u, defined by

$$v = t^{-2}e^{ct^3}u,$$

where c is an arbitrary constant. Then the equation transforms to

$$\frac{d^2u}{dt^2} + \left[6c - i\left(1 + \frac{1}{\mathrm{Pr}}\right)\right]t^2\frac{du}{dt} + \left\{\left[9c^2 - \frac{1}{\mathrm{Pr}} - 3ic\left(1 + \frac{1}{\mathrm{Pr}}\right)\right]t^4 + (6c - 2i)t + \frac{\mathrm{Ric}}{t^2}\right\}u = 0. \quad (43)$$

Choose c to satisfy the equation

$$9c^2 - 3i(1 + 1/\mathrm{Pr})c - 1/\mathrm{Pr} = 0.$$

This has the solutions

$$c = i/3, \qquad i/3\ \mathrm{Pr}.$$

The solution $c = i/3$ will be used, since this makes the term $(6c - 2i)tu$ in the above differential equation vanish. Then Eq. (43) becomes

$$\frac{d^2u}{dt^2} + i\left(1 - \frac{1}{\mathrm{Pr}}\right)t^2\frac{du}{dt} + \frac{\mathrm{Ric}}{t^2}u = 0. \quad (44)$$

When Pr = 1 this reduces to

$$t^2\,d^2u/dt^2 + \mathrm{Ric}\,u = 0.$$

Using the trial function

$$u = t^m,$$

the condition that this equation be satisfied is

$$m(m - 1) + \mathrm{Ric} = 0.$$

The solutions for m are

$$\begin{aligned} m &= \tfrac{1}{2} \pm (\tfrac{1}{4} - \mathrm{Ric})^{\frac{1}{2}} \qquad (\mathrm{Ric} < \tfrac{1}{4}), \\ &= \tfrac{1}{2} \pm i(\mathrm{Ric} - \tfrac{1}{4})^{\frac{1}{2}} \qquad (\mathrm{Ric} > \tfrac{1}{4}). \end{aligned} \quad (45)$$

Thus the function $w(\eta)$ is given by

$$w(\eta) = \int_a^b t^{m-2}\exp\left(\frac{i}{3}\,t^3 + \eta t\right)dt, \quad (46)$$

when Pr = 1.

When Ric = 0 Eq. (44) reduces to

$$\frac{d^2u}{dt^2} + i\left(1 - \frac{1}{\mathrm{Pr}}\right)t^2\frac{du}{dt} = 0.$$

This equation has the solutions

$$u_1 = 1, \qquad u_2 = \int_0^t \exp\left[-\frac{i}{3}\left(1 - \frac{1}{\mathrm{Pr}}\right)t^3\right]dt,$$

valid for arbitrary Pr and Ric = 0. The solution u_1 gives the following expression for $w(\eta)$:

$$w(\eta) = \int_a^b t^{-2}\exp\left(\frac{i}{3}\,t^3 + \eta t\right)dt. \quad (47)$$

To deal with Eq. (44) for arbitrary Pr and Ric, first make the following transformation of the independent variable:

$$t' = Dt^3,$$

where D is an arbitrary constant. The transformed equation is

$$9t'^2\frac{d^2u}{dt'^2} + \left[6t' + \frac{3i}{D}\left(1 - \frac{1}{\mathrm{Pr}}\right)t'^2\right]\frac{du}{dt'} + \mathrm{Ric}\,u = 0.$$

Next use the new dependent variable V, defined by

$$u = (t')^{-\frac{1}{3}}\exp\left[-\frac{i}{6D}\left(1 - \frac{1}{\mathrm{Pr}}\right)t'\right]V.$$

The result is

$$9t'^2\frac{d^2V}{dt'^2} + \left\{(\mathrm{Ric} + 2) - \frac{i}{D}\left(1 - \frac{1}{\mathrm{Pr}}\right)t' + \frac{1}{4D^2}\left(1 - \frac{1}{\mathrm{Pr}}\right)^2t'^2\right\}V = 0.$$

Making the identifications

$$\frac{1}{4} = -\frac{1}{36D^2}\left(1 - \frac{1}{\mathrm{Pr}}\right)^2, \quad K = -\frac{i(1 - 1/\mathrm{Pr})}{9D}, \quad \tfrac{1}{4} - m^2 = (\mathrm{Ric} + 2)/9, \quad (48)$$

the above differential equation becomes

$$\frac{d^2V}{dt'^2} + \left[-\frac{1}{4} + \frac{K}{t'} + \frac{(\frac{1}{4} - m^2)}{t'^2}\right]V = 0,$$

which is Whittaker's equation for the confluent hypergeometric function, with solution $W_{K,m}(t')$.

Solve the set of Eqs. (48):

$$D = \pm\frac{i}{3}\left(1 - \frac{1}{\mathrm{Pr}}\right), \quad K = \mp\frac{1}{3}, \quad m = \pm\frac{(1 - 4\,\mathrm{Ric})^{\frac{1}{2}}}{6},$$

where the signs of D and K are correlated. Finally we can go back to get the solution in terms of the set of variables t, v:

$$v = t^{-2} \exp\left[\frac{i}{6}\left(1 + \frac{1}{\mathrm{Pr}}\right)t^3\right] \times W_{\mp\frac{1}{2},\, \pm(1-4\,\mathrm{Ric})^{\frac{1}{2}}/6}\left[\pm\frac{i}{3}\left(1 - \frac{1}{\mathrm{Pr}}\right)t^3\right],$$

where a constant factor has been discarded.

VI. SINGULARITIES IN THE ABSENCE OF HEAT CONDUCTION OR VISCOSITY

Equation (39), which holds for vanishing heat conduction, has a singularity at $\eta = 0$.

When Ric = 0, it reduces to

$$\eta\, d^2w/d\eta^2 + i\, d^4w/d\eta^4 = 0,$$

after division by η. This no longer has a singularity. Therefore when gravity is absent viscosity is sufficient to remove the singularity.

Using a series solution of the form

$$w = \sum_{s=0}^{\infty} a_s \eta^{(\rho+s)},$$

for trial, reveals three solutions of Eq. (39) that are not singular at $\eta = 0$. The fourth solution has the behavior

$$w = 1 + \tfrac{1}{6} i\,\mathrm{Ric}\, \eta^3 \ln \eta$$

near $\eta = 0$.

Equation (41), which holds for vanishing viscosity, also has a singularity at $\eta = 0$. When Ric = 0, it reduces to

$$\eta^2 \frac{d^2w}{d\eta^2} + i\eta \frac{d^4w}{d\eta^4} = -2i \frac{d^3w}{d\eta^3}.$$

This equation still has a singularity. Therefore when gravity is absent heat conduction is not sufficient by itself to remove the singularity. Equation (41) has a solution behaving as

$$w = \eta \ln \eta$$

near $\eta = 0$.

VII. CHOICE OF CONTOURS

Condition (42b) must still be satisfied for the definite integrals to be solutions of the differential equation. Confining attention to the case Pr = 1, use can be made of the explicit expression for v obtained above. Then condition (42b) becomes

$$\left\{t^{m-1} \exp\left(\frac{i}{3} t^3 + \eta t\right)\left[\eta t - m + \frac{i}{\mathrm{Pr}} t^3\right]\right\}_a^b = 0.$$

Thus it is sufficient that $it^3 \to -\infty$ at the end points for the above condition to be satisfied. This happens for arg $t = \pi/6, 5\pi/6, 3\pi/2, 13\pi/6, \cdots$ as $|t| \to \infty$.

Since m is not in general an integer, it follows that the integrand of Eq. (46) will have a branch point at $t = 0$. The complex t-plane is cut along the line arg $t = \pi/6$, and the sheet $\pi/6 \leq \arg t < 2\pi + \pi/6$, will be referred to as the principal plane. Similarly the sheets

$$2\pi + \frac{\pi}{6} \leq \arg t < 4\pi + \frac{\pi}{6} \quad \text{and} \quad \frac{\pi}{6} - 2\pi \leq \arg t < \frac{\pi}{6}$$

are referred to as the upper and lower planes, respectively.

Writing

$$t = re^{i\theta},$$

the contour C_1 is defined as follows: the contour starts at infinity and goes to $r = r_1$ along the straight line $\theta = 5\pi/6$; it next describes an arc of the circle $r = r_1$ from $\theta = 5\pi/6$ clockwise to $\theta = \pi/6$; finally it goes out to infinity again along the line $\theta = \pi/6$. The value of the function defined by this contour is independent of the value of r_1. The contour C_2 starts at infinity and goes to $r = r_1$ along the straight line $\theta = 3\pi/2$; then it follows an arc of the circle $r = r_1$ from $\theta = 3\pi/2$ to $\theta = 5\pi/6$; finally it goes out to infinity along the straight line $\theta = 5\pi/6$. The contour C_3 starts at infinity and goes to $r = r_1$ along the straight line $\theta = 13\pi/6$; then it follows an arc of the circle $r = r_1$ from $\theta = 13\pi/6$ to $\theta = 3\pi/2$; finally it goes out to infinity again along the straight line $\theta = 3\pi/2$. The contour C_4 starts at infinity and goes to $r = r_1$ along the straight line $\theta = \pi/6$; then it follows the circle $r = r_1$ from $\theta = \pi/6$ to $\theta = 13\pi/6$; finally it goes out to infinity again along the straight line $\theta = 13\pi/6$.

Denote the integral (46) taken along the path C_i by $w_i(\eta)$:

$$w_i(\eta) = \int_{C_i} t^{m-2} \exp\left(\frac{i}{3} t^3 + \eta t\right) dt. \qquad (49)$$

Then, by the definitions of the contours given above, the following relation holds between the four solutions thus defined:

$$w_1 + w_2 + w_3 + w_4 = 0. \qquad (50)$$

The three contours C_1, C_2, and C_3 together with the two values of m given by Eq. (45) yield six independent solutions of the sixth-order differential equation (38) (with Pr = 1 in the latter). Nevertheless, it is convenient to define here several more solutions which will be needed later.

The contour C_5 starts at infinity and goes to $r = r_1$ along the straight line $\theta = 5\pi/6$; then it describes a circle of radius $r = r_1$ from $\theta = 5\pi/6$ to

$\theta = 17\pi/6$, crossing from the principal plane to the upper plane in the process; finally it goes out to infinity again along the straight line $\theta = 17\pi/6$. The contour C_6 starts at infinity and goes to $r = r_1$ along the straight line $\theta = 5\pi/6 - 2\pi$; then it traverses the circle $r = r_1$ from $\theta = 5\pi/6 - 2\pi$ to $\theta = 5\pi/6$, crossing from the lower plane into the principal plane in the process; finally it goes out to infinity along the straight line $\theta = 5\pi/6$. The contour C_7 starts at infinity and goes to $r = r_1$ along the straight line $\theta = \pi/6 - 2\pi$; then it traverses the circle $r = r_1$ from $\theta = \pi/6 - 2\pi$ to $\theta = \pi/6$; finally it goes out to infinity along the straight line $\theta = \pi/6$.

VIII. POWER SERIES EXPANSIONS OF THE SOLUTIONS

The contour integrals of Eq. (49) can be used to expand the functions $w_i(\eta)$ defined by them in ascending powers of η. The function $w_1(\eta)$ is used as an example. The equations of the various parts of the contour C_1 are written

$$t = t_0 e^{i5\pi/6} \text{ (along the incoming line)},$$

$$t = r_1 e^{i\theta} \text{ (along the arc of the circle)},$$

$$t = t_1 e^{i\pi/6} \text{ (along the outgoing line)}.$$

Then Eq. (49) becomes, more explicitly

$$\begin{aligned} w_1(\eta) = {} & \int_\infty^{r_1} t_0^{m-2} e^{i(m-2)5\pi/6} \\ & \times e^{(i\frac{1}{3}t_0^3 e^{i5\pi/2} + \eta t_0 e^{i5\pi/6})} e^{i5\pi/6}\, dt_0 \\ & + \int_{5\pi/6}^{\pi/6} r_1^{m-2} e^{i(m-2)\theta} \\ & \times \exp\left(\frac{i}{3} r_1^3 e^{3i\theta} + \eta r_1 e^{i\theta}\right) i r_1 e^{i\theta}\, d\theta \\ & + \int_{r_1}^{\infty} t_1^{m-2} e^{i(m-2)\pi/6} \\ & \times \exp\left(\frac{i}{3} t_1^3 e^{i\pi/2} + \eta t_1 e^{i\pi/2}\right) e^{i\pi/6}\, dt_1. \end{aligned}$$

The next step is to expand the exponentials containing η in powers of η. After some rearrangement we get

$$\begin{aligned} w_1(\eta) = {} & e^{i(m-1)5\pi/6} \sum_{k=0}^{\infty} \frac{\eta^k e^{i5\pi k/6}}{k!} \int_\infty^{r_1} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 \\ & + i r_1^{m-1} \sum_{k=0}^{\infty} \frac{\eta^k r_1^k}{k!} \int_{5\pi/6}^{\pi/6} e^{i(m-1+k)\theta} \exp\left(\frac{i}{3} r_1^3 e^{3i\theta}\right) d\theta \\ & + e^{i(m-1)\pi/6} \sum_{k=0}^{\infty} \frac{\eta^k e^{i\pi k/6}}{k!} \int_{r_1}^{\infty} t_1^{m+k-2} e^{-t_1^3/3}\, dt_1. \end{aligned}$$

Now combine the first and third terms on the right-hand side of the above equation by replacing the variable of integration t_1 by the variable t_0:

$$\begin{aligned} w_1(\eta) = {} & \sum_{k=0}^{\infty} \frac{\eta^k}{k!} \left[e^{i(m+k-1)5\pi/6} - e^{i(m+k-1)\pi/6}\right] \\ & \times \int_\infty^{r_1} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 + i r_1^{m-1} \sum_{k=0}^{\infty} \frac{\eta^k r_1^k}{k!} \\ & \times \int_{5\pi/6}^{\pi/6} e^{i(m+k-1)\theta} \exp\left(\frac{i}{3} r_1^3 e^{3i\theta}\right) d\theta. \end{aligned}$$

Since $w_1(\eta)$ is independent of the value chosen for r_1, we may conveniently evaluate the above series in the limit as $r_1 \to 0$. Only the case Ric ≥ 0 will be considered. For this case Re $m \geq 0$, by Eq. (45). Then

$$\exp\left(\frac{i}{3} r_1^3 e^{3i\theta}\right) \to 1, \quad r_1^{m+k-1} \to 0 \quad \text{for} \quad k \geq 1,$$

and

$$\int_\infty^{r_1} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 \to \int_\infty^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0$$

for $k \geq 1$, all in the limit as $r_1 \to 0$.

The above expression for $w_1(\eta)$ reduces to

$$\begin{aligned} w_1(\eta) = {} & \sum_{k=1}^{\infty} \frac{\eta^k}{k!} \left[e^{i(m+k-1)5\pi/6} - e^{i(m+k-1)\pi/6}\right] \\ & \times \int_\infty^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 + \left[e^{i(m-1)5\pi/6} - e^{i(m-1)\pi/6}\right] \\ & \times \int_\infty^{r_1} t_0^{(m-2)} e^{-t_0^3/3}\, dt_0 + i r_1^{m-1} \int_{5\pi/6}^{\pi/6} e^{i(m-1)\theta}\, d\theta. \end{aligned}$$

The integral

$$\int_{5\pi/6}^{\pi/6} e^{i(m-1)\theta}\, d\theta = \frac{1}{i(m-1)} \left[e^{i(m-1)\pi/6} - e^{i(m-1)5\pi/6}\right]$$

is needed. Also, integration by parts gives

$$\begin{aligned} & \int_\infty^{r_1} t_0^{(m-2)} e^{-t_0^3/3}\, dt_0 \\ & \quad = \frac{1}{(m-1)} \left[r_1^{m-1} e^{-r_1^3/3} + \int_\infty^{r_1} t_0^{m+1} e^{-t_0^3/3}\, dt_0\right]. \end{aligned}$$

In the limit $r_1 \to 0$, this becomes

$$\begin{aligned} & \int_\infty^{r_1} t_0^{(m-2)} e^{-t_0^3/3}\, dt_0 \\ & \quad \to \frac{1}{(m-1)} \left[r_1^{m-1} + \int_\infty^{0} t_0^{m+1} e^{-t_0^3/3}\, dt_0\right]. \end{aligned}$$

Substitution of these integrals in the last expression

for $w_1(\eta)$ gives

$$w_1(\eta) = \sum_{k=1}^{\infty} \frac{\eta^k}{k!} [e^{i(m+k-1)5\pi/6} - e^{i(m+k-1)\pi/6}] \int_{\infty}^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 + \frac{[e^{i(m-1)5\pi/6} - e^{i(m-1)\pi/6}]}{(m-1)} \int_{\infty}^{0} t_0^{m+1} e^{-t_0^3/3}\, dt_0.$$

This is required expansion of $w(\eta)$ in ascending powers of η. It is manifestly independent of r_1.

The method used above gives the following power series expansions of the other solutions that have been defined:

$$w_2(\eta) = \sum_{k=1}^{\infty} \frac{\eta^k}{k!} [e^{(3\pi i/2)(m+k-1)} - e^{(5\pi i/6)(m+k-1)}] \times \int_{\infty}^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 + \frac{1}{(m-1)} [e^{(5\pi i/6)(m-1)} - e^{(3\pi i/2)(m-1)}] \int_{0}^{\infty} t_0^{m+1} e^{-t_0^3/3}\, dt_0,$$

$$w_3(\eta) = \sum_{k=1}^{\infty} \frac{\eta^k}{k!} [e^{(\pi i/2)(3m+3k-1)} - e^{(\pi i/6)(13m+k-7)}] \times \int_{\infty}^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 + \frac{1}{(m-1)} [e^{(\pi i/6)(13m-7)} - e^{(\pi i/2)(3m-1)}] \int_{0}^{\infty} t_0^{m+1} e^{-t_0^3/3}\, dt_0,$$

$$w_4(\eta) = \sum_{k=1}^{\infty} \frac{\eta^k}{k!} e^{(i\pi/6)(m+k-1)} \times (1 - e^{2\pi mi}) \int_{\infty}^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 - \frac{1}{(m-1)} e^{(i\pi/6)(m-1)} (1 - e^{2\pi mi}) \int_{0}^{\infty} t_0^{m+1} e^{-t_0^3/3}\, dt_0,$$

$$w_5(\eta) = \sum_{k=1}^{\infty} \frac{\eta^k}{k!} e^{(i5\pi/6)(m+k-1)} \times (1 - e^{2\pi mi}) \int_{\infty}^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 - \frac{1}{(m-1)} e^{(5\pi i/6)(m-1)} (1 - e^{2\pi mi}) \int_{0}^{\infty} t_0^{m+1} e^{-t_0^3/3}\, dt_0,$$

$$w_6(\eta) = -\sum_{k=1}^{\infty} \frac{\eta^k}{k!} e^{(5\pi i/6)(m+k-1)} \times (1 - e^{-2\pi mi}) \int_{\infty}^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 + \frac{1}{(m-1)} e^{(5\pi i/6)(m-1)} (1 - e^{-2\pi mi}) \int_{0}^{\infty} t_0^{m+1} e^{-t_0^3/3}\, dt_0,$$

$$w_7(\eta) = -\sum_{k=1}^{\infty} \frac{\eta^k}{k!} e^{(\pi i/6)(5+k-11m)} \times (1 - e^{2\pi im}) \int_{\infty}^{0} t_0^{m+k-2} e^{-t_0^3/3}\, dt_0 - \frac{1}{(m-1)} e^{(\pi i/6)(5-11m)} (1 - e^{2\pi mi}) \int_{\infty}^{0} t_0^{m+1} e^{-t_0^3/3}\, dt_0.$$

The following relations between the various solutions are obtained from these power series expansions:

$$w_4 = e^{2\pi mi} w_7, \tag{51}$$

$$w_5 = e^{2\pi mi} w_6, \tag{52}$$

$$w_5 = (1 - e^{2\pi mi}) w_1 + w_4. \tag{53}$$

IX. SADDLE POINTS

Asymptotic formulas for the functions $w_i(\eta)$ are needed. For this purpose, write Eq. (49) in the form

$$w_i(\eta) = \int_{C_i} e^{F(t)}\, dt, \tag{54}$$

where

$$F(t) = (i/3)t^3 + \eta t + (m-2) \ln t. \tag{55}$$

The saddle points, denoted by t_1 here, are defined by the equation

$$F'(t_1) = 0,$$

where primes indicate differentiation with respect to t. They satisfy the cubic equation

$$it_1^3 + \eta t_1 + (m-2) = 0.$$

Thus, there are three saddle points for a given value of η and m.

When $|\eta| \to \infty$, a pair of saddle points are given approximately by

$$t_1 = \pm(i\eta)^{\frac{1}{2}}, \tag{56}$$

while the remaining saddle point is given by the formula

$$t_1 = (2 - m)/\eta. \tag{57}$$

Thus, $|t_1| \to \infty$ for two of the saddle points, and $|t_1| \to 0$ for the remaining, as $|\eta| \to \infty$.

It will be observed that near the saddle point t^{m-2} varies slowly compared with

$$\exp\left(\frac{i}{3} t^3 + \eta t\right)$$

in the former case. For this pair of saddle points the problem is simplified by writing Eq. (49) in the form

$$w_i(\eta) = \int_{C_i} t^{m-2} e^{f(t)}\, dt, \tag{58}$$

with

$$f(t) = \tfrac{1}{3}it^3 + \eta t. \tag{59}$$

The saddle points of $f(t)$ are given exactly by Eq. (56).

Decompose $f(t)$ into real and imaginary parts:

$$f(t) = u + iv.$$

Remembering that $t = re^{i\theta}$, this gives

$$u = \eta r \cos\theta - \tfrac{1}{3}r^3 \sin 3\theta, \tag{60}$$

$$v = \eta r \sin\theta + \tfrac{1}{3}r^3 \cos 3\theta. \tag{61}$$

The curve of steepest descent is the curve $v =$ constant that passes through the saddle point. The curves $u =$ constant are the orthogonal trajectories of the family of curves $v =$ constant.

When $\eta > 0$, the choice of phase $\arg \eta = 0$ will be made, while when $\eta < 0$, the choice of phase is $\arg \eta = \pi$, so that

$$\eta = |\eta| e^{i\pi}.$$

With this choice of phases the saddle points (56) become

$$t_1 = \eta^{\frac{1}{2}} e^{i\pi/4}, \qquad t_1 = \eta^{\frac{1}{2}} e^{i5\pi/4} \tag{62a, b}$$

for positive η, and

$$t_1 = |\eta|^{\frac{1}{2}} e^{i3\pi/4}, \qquad t_1 = |\eta|^{\frac{1}{2}} e^{i7\pi/4} \tag{63a, b}$$

for negative η. Therefore, all the saddle points are in the principal plane.

Substituting these values of t_1 in Eq. (59) gives the values of v (say v_1) for the curves passing through the saddle points

$$v_1 = \tfrac{1}{3}\sqrt{2}\,\eta^{\frac{3}{2}}, \qquad v_1 = -\tfrac{1}{3}\sqrt{2}\,\eta^{\frac{3}{2}} \tag{64a, b}$$

for positive η, and

$$v_1 = -\tfrac{1}{3}\sqrt{2}\,|\eta|^{\frac{3}{2}}, \qquad v_1 = \tfrac{1}{3}\sqrt{2}\,|\eta|^{\frac{3}{2}} \tag{65a, b}$$

for negative η.

The saddle points given by Eq. (57) require the use of the function $F(t)$. Separate this function into real and imaginary parts:

$$F(t) = u + iv.$$

When Ric $< \frac{1}{4}$, so that m is real, we get

$$u = \eta r \cos\theta - \tfrac{1}{3}r^3 \sin 3\theta + (m-2)\ln r, \tag{66}$$

$$v = \eta r \sin\theta + \tfrac{1}{3}r^3 \cos 3\theta + (m-2)\theta. \tag{67}$$

When Ric $> \frac{1}{4}$, so that m is complex, the quantity

$$n = (\text{Ric} - \tfrac{1}{4})^{\frac{1}{2}}$$

is used. According to the second equation of (45), m is given by

$$m = \tfrac{1}{2} \pm in.$$

Now u and v are given by the formulas

$$u = \eta r \cos\theta - \tfrac{1}{3}r^3 \sin 3\theta - \tfrac{3}{2}\ln r \mp n\theta, \tag{68}$$

$$v = \eta r \sin\theta + \tfrac{1}{3}r^3 \cos 3\theta \pm n \ln r - \tfrac{3}{2}\theta. \tag{69}$$

Note that $(2 - m)$ is positive when Ric is in the range $0 \leq \text{Ric} \leq \frac{1}{4}$. From Eq. (57) we see that $\arg t_1 = 0$ when $\eta > 0$. Therefore, t_1 is in the lower plane. A second approximation to t_1 may be obtained by substituting formula (57) for t_1 into the cubic term in the cubic equation for t_1:

$$t_1 = [(2-m)/\eta][1 - i(2-m)^2/\eta^3]. \tag{70}$$

This value of t_1 gives the following value of v_1 when substituted into Eq. (55):

$$v_1 = (2-m)^3/3\eta^3, \tag{71}$$

when m is real.

When $\eta < 0$ Eq. (57) gives $\arg t_1 = -\pi$ when the usual choice of the phase of η is made. Therefore, t_1 is in the lower plane in this case also. A more accurate value of t_1 is

$$t_1 = \frac{(2-m)}{|\eta|} e^{-\pi i}\left[1 + \frac{(2-m)^2}{|\eta|^3} i\right]. \tag{72}$$

The value of v_1 corresponding to this is

$$v_1 = (2-m)\pi - (2-m)^3/3\,|\eta|^3, \tag{73}$$

when m is real. When Ric $> \frac{1}{4}$ so that m is complex, the saddle point for the case $\eta < 0$ remains in the lower plane. However, as Ric increases, the saddle point for the case $\eta > 0$ moves from the lower plane to the principal plane.

X. DISCUSSION OF THE CURVES OF STEEPEST DESCENT

When r is large the two dominant terms in Eqs. (61), (67), and (69) are the same and are

$$\eta r \sin\theta + \tfrac{1}{3}r^3 \cos 3\theta = 0.$$

Solving this for r gives

$$r^2 = -3\eta \sin\theta / \cos 3\theta \tag{74}$$

when $r \neq 0$. Thus there are asymptotes when $\cos 3\theta = 0$ so that $r \to \infty$ for these values. This happens at $\theta = \pi/6, \pi/2, 5\pi/6, 7\pi/6, 3\pi/2, 11\pi/6, 13\pi/6$ in the principal plane and at

$$\theta = -\frac{\pi}{6}, -\frac{\pi}{2}, -\frac{5\pi}{6}, -\frac{7\pi}{6}, -\frac{3\pi}{2}, -\frac{11\pi}{6},$$

in the lower plane. Equation (74) shows that the

asymptotes

$$\theta = \frac{\pi}{6}, \frac{5\pi}{6}, \frac{7\pi}{6}, \frac{11\pi}{6}, \frac{13\pi}{6}, -\frac{\pi}{6}, -\frac{5\pi}{6}, -\frac{7\pi}{6}, -\frac{11\pi}{6}, -\frac{13\pi}{6}$$

are approached through decreasing values of θ and the asymptotes

$$\theta = \pi/2, 3\pi/2, -\pi/2, -3\pi/2$$

through increasing values of θ when $\eta > 0$. When $\eta < 0$ the situation is reversed.

Equations (60), (66), and (68) show that u takes the following approximate form for large r:

$$u = -\tfrac{1}{3}r^3 \sin 3\theta.$$

Thus $u \to -\infty$ along the branches asymptotic to

$$\theta = \frac{\pi}{6}, \frac{5\pi}{6}, \frac{3\pi}{2}, \frac{13\pi}{6}, -\frac{\pi}{2}, -\frac{7\pi}{6}, -\frac{11\pi}{6},$$

while $u \to +\infty$ along the branches asymptotic to

$$\theta = \frac{\pi}{2}, \frac{7\pi}{6}, \frac{11\pi}{6}, -\frac{\pi}{6}, -\frac{5\pi}{6}, -\frac{3\pi}{2}.$$

Note that the curves given by Eq. (74) have no intercepts with the asymptotes. If this equation were exactly true the branches asymptotic to $\theta = \pi/6$ and $\theta = \pi/2$ would connect as would the branches asymptotic to $\theta = 7\pi/6$ and $\theta = 3\pi/2$, when $\eta > 0$. Similarly the branches asymptotic to $\theta = \pi/2$ and $\theta = 5\pi/6$ would connect as would the branches asymptotic to $\theta = 3\pi/2$ and $\theta = 11\pi/6$, when $\eta < 0$. Corresponding results would hold for the lower plane. The minimum value of r for these branches would be of the order of magnitude of $|\eta|^{\frac{1}{3}}$. Equations (67) and (69) have no other terms that are of magnitude comparable with those already retained, when this minimum value of r is approached. Consequently the branches mentioned will be connected when the full equations (67) and (69) are taken into account. On the other hand, Eq. (61) does contain an additional term that must be taken into account, as shown by Eqs. (64) and (65). It can no longer be assumed that the branches mentioned will join up, when the full equation (61) is taken into account.

Next, the behavior of the curves near the saddle point must be obtained. To this end, expand the function $f(t)$ in a Taylor series around the saddle point t_1:

$$f(t) = f(t_1) + f'(t_1)(t - t_1) + \tfrac{1}{2}f''(t_1)(t - t_1)^2 + \cdots . \quad (75)$$

This equals

$$f(t) = \tfrac{1}{3}it_1^3 + \eta t_1 + it_1(t - t_1)^2 + \cdots .$$

The variables ρ and ϕ given by

$$t - t_1 = \rho e^{i\phi}$$

are used. Substitution of t_1 from Eqs. (62) and (63) results in the following expressions for $f(t)$:

$$f(t) = \tfrac{2}{3}\eta^{\frac{3}{2}}e^{i\pi/4} + \eta^{\frac{1}{2}}\rho^2 e^{i(2\phi+3\pi/4)}, \quad (76a)$$

$$f(t) = \tfrac{2}{3}\eta^{\frac{3}{2}}e^{i5\pi/4} + \eta^{\frac{1}{2}}\rho^2 e^{i(2\phi+7\pi/4)} \quad (76b)$$

when η is positive, and

$$f(t) = \tfrac{2}{3}|\eta|^{\frac{3}{2}}e^{i7\pi/4} + |\eta|^{\frac{1}{2}}\rho^2 e^{i(2\phi+5\pi/4)}, \quad (77a)$$

$$f(t) = \tfrac{2}{3}|\eta|^{\frac{3}{2}}e^{i11\pi/4} + |\eta|^{\frac{1}{2}}\rho^2 e^{i(2\phi+9\pi/4)} \quad (77b)$$

when η is negative.

Separate Eq. (76a) into real and imaginary parts:

$$u = \tfrac{1}{3}\sqrt{2}\,\eta^{\frac{3}{2}} + \eta^{\frac{1}{2}}\rho^2 \cos(2\phi + 3\pi/4),$$

$$v = \tfrac{1}{3}\sqrt{2}\,\eta^{\frac{3}{2}} + \eta^{\frac{1}{2}}\rho^2 \sin(2\phi + 3\pi/4).$$

Using Eq. (64a) we see that near the saddle point the curve of steepest descent satisfies the condition

$$\sin(2\phi + 3\pi/4) = 0.$$

Thus the curve is tangent to the lines $\phi = \pi/8, 5\pi/8, 9\pi/8, 13\pi/8$ at t_1. From the above equation for u we see that this quantity decreases as we go away from the saddle point along the lines $\phi = \pi/8, 9\pi/8$; and likewise it increases when $\phi = 5\pi/8, 13\pi/8$. Similarly, Eq. (76b) determines curves tangent to the lines $\phi = \pi/8, 9\pi/8$ along which u increases, and curves tangent to the lines $\phi = 5\pi/8, 13\pi/8$ along which u decreases. Equation (77a) determines curves tangent to the lines $\phi = 3\pi/8, 11\pi/8$ along which u increases; and curves tangent to the lines $\phi = 7\pi/8, 15\pi/8$ along which u decreases. Equation (77b) determines curves tangent to the lines $\phi = 3\pi/8, 11\pi/8$ along which u decreases as we recede from t_1; and curves tangent to the lines $\phi = 7\pi/8, 15\pi/8$ along which u increases as we recede from t_1.

The behavior of the curves given by Eqs. (67) and (69) may be obtained by expanding $F(t)$ in a Taylor series around t_1:

$$F(t) = F(t_1) + F'(t_1)(t - t_1) + \tfrac{1}{2}F''(t_1)(t - t_1)^2 + \cdots .$$

Using Eqs. (55) and (57) this becomes approximately

$$F(t) = (2 - m) - (2 - m)\ln t_1 + [\eta^2/2(2 - m)](t - t_1)^2 + \cdots . \quad (78)$$

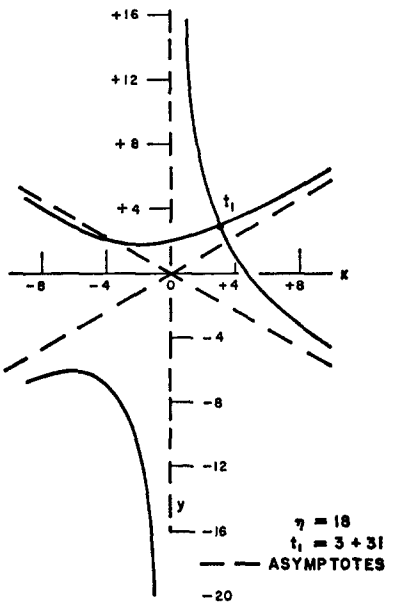


FIG. 2. Curves of steepest descent passing through the saddle point.

Putting

$$t_1 = r_1 e^{i\theta_1} \tag{79}$$

and using the variables ρ and ϕ, allow $F(t)$ to be separated into real and imaginary parts, when m is real:

$$u = (2 - m) - (2 - m)\ln r_1 + [\eta^2/2(2 - m)]\rho^2 \cos 2\phi,$$

$$v = -(2 - m)\theta_1 + [\eta^2/2(2 - m)]\rho^2 \sin 2\phi.$$

Thus the curves of steepest descent are tangent to the lines $\phi = 0, \pi/2, \pi, 3\pi/2$ at the saddle point. The equation for u shows that it increases going away from the saddle point along the lines $\phi = 0, \pi$ and decreases along the lines $\phi = \pi/2, 3\pi/2$.

When m is complex, write

$$2 - m = \tfrac{3}{2} \mp in = (\tfrac{9}{4} + n^2)^{\frac{1}{4}} \times \exp(\mp i \tan^{-1} 2n/3), \tag{80}$$

where $n = (\mathrm{Ric} - \tfrac{1}{4})^{\frac{1}{2}}$. Inserting this expression for $(2 - m)$ into Eq. (78) and separating $F(t)$ into real and imaginary parts we get

$$u = \tfrac{3}{2} - \tfrac{3}{2}\ln r_1 \mp n\theta_1 + \frac{\eta^2\rho^2}{2(\frac{9}{4} + n^2)^{\frac{1}{2}}} \cos\left(2\phi \pm \tan^{-1}\frac{2n}{3}\right),$$

$$v = \mp n - \tfrac{3}{2}\theta_1 \pm n \ln r_1 + \frac{\eta^2\rho^2}{2(\frac{9}{4} + n^2)^{\frac{1}{2}}} \sin\left(2\phi \pm \tan^{-1}\frac{2n}{3}\right),$$

where use has been made of Eq. (79). The curves of steepest descent at the saddle point are tangent to the lines given by

$$\sin(2\phi \pm \tan^{-1}(2n/3)) = 0.$$

The solutions

$$2\phi \pm \tan^{-1}(2n/3) = 0, 2\pi$$

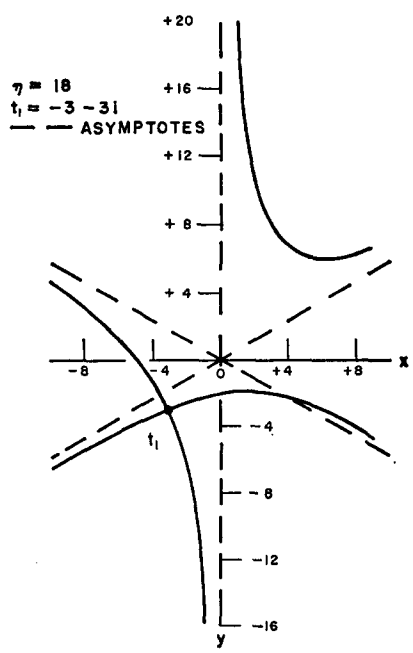


FIG. 3. Curves of steepest descent passing through the saddle point.

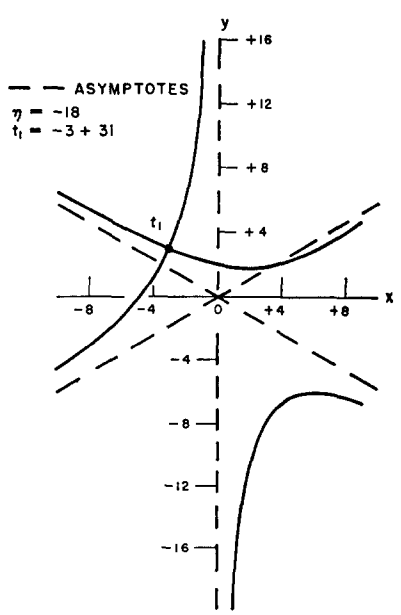


FIG. 4. Curves of steepest descent passing through the saddle point.

when inserted into the above expression for u show that this quantity increases when going away from t_1. Likewise the solutions

$$2\phi \pm \tan^{-1}(2n/3) = \pi, 3\pi$$

are seen to give decreasing values for u.

The final task is to connect up the branches of the curve near the asymptotes with the branches near the saddle point. For this purpose consideration of the intercepts of the curves $v = v_1$ with the asymptotes is helpful. For Eq. (61) we see that these intercepts are given by

$$r = v_1/\eta \sin\theta.$$

Hence there is at most one intercept with each asymptote. When v_1 and η have the same sign there is an intercept with each asymptote in the upper half-plane, and none with the asymptotes in the lower half-plane. When v_1 and η have opposite signs the situation is reversed.

Note also that since Eq. (61) is cubic in r there are at most three values of r for every value of θ. Also u is monotonic along each branch of the curve starting from the saddle point.

Figures 2 through 5 illustrate the curves of steepest descent corresponding to Eq. (61). The curve of Fig. 2 passing through the saddle point and asymptotic to the lines $\theta = \pi/6, 5\pi/6$ is called C_1'. The curve of Fig. 3 passing through the saddle point and asymptotic to the lines $\theta = 5\pi/6, 3\pi/2$ is called C_2'. The curve of Fig. 4 passing through the saddle point and asymptotic to the lines $\theta = \pi/6, 5\pi/6$ is called C_1''. Finally the curve of Fig. 5 passing through the saddle point and asymptotic to the lines $\theta = 3\pi/2, 13\pi/6$ will be called C_3'. These curves pass through the saddle points given by Eqs. (62) and (63), respectively.

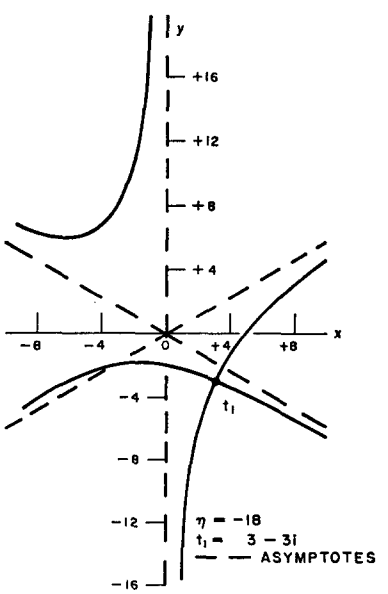


FIG. 5. Curves of steepest descent passing through the saddle point.

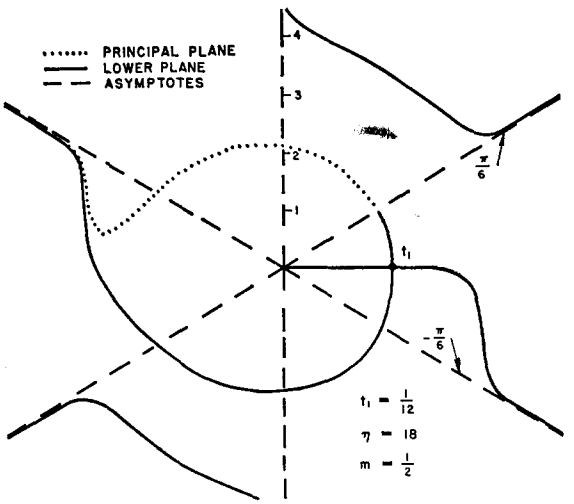


FIG. 6. Curves of steepest descent. The radial coordinate is $\log_{10} 1000r$ plotted vs θ.

The curves of Eqs. (67) and (69) have branches near the origin. Putting $r = 0$ in Eq. (67) gives

$$\theta = v_1/(m - 2).$$

From Eq. (71) we see that $\theta \to 0$ when η is large and positive. From Eq. (73) we see that $\theta \to -\pi$ when η is large and negative. Hence these branches are in the lower plane. Equation (66) shows that $u \to -\infty$ as we approach the origin along these branches. Letting $r \to 0$ in Eq. (69) shows that the curve spirals toward the origin. Equation (68)

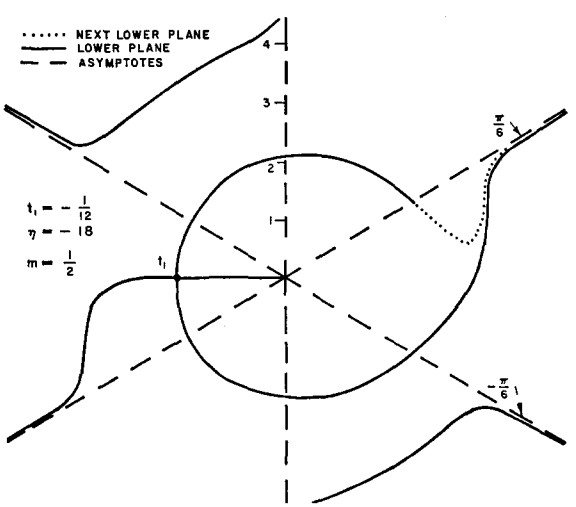


FIG. 7. Curves of steepest descent. The radial coordinate is $\log_{10} 1000r$ plotted vs θ.

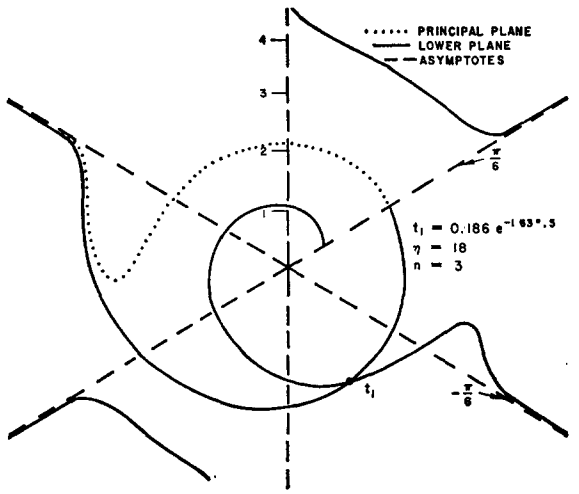


FIG. 8. Curves of steepest descent. The radial coordinate is $\log_{10} 1000r$ plotted vs θ.

shows that $u \to +\infty$ as we approach the origin along the spiral.

Next the various branches of the curve of Eq. (67) must be joined up. For this purpose the intercepts of the curve with the asymptotes are once again useful. When $\eta > 0$ they are given by

$$r = [(2 - m)/\eta](\theta/\sin\theta),$$

where use has been made of Eq. (71). There are intercepts with the lines $\theta = 5\pi/6, \pi/2, \pi/6, -\pi/6, -\pi/2, -5\pi/6$ and none with the lines $\theta = 11\pi/6, 3\pi/2, 7\pi/6, -7\pi/6, -3\pi/2, -11\pi/6$.

When $\eta < 0$, the use of Eq. (73) gives the following formula for the intercepts:

$$r = -[(2 - m)/|\eta|][(\pi + \theta)/\sin\theta].$$

There is an intercept with every asymptote in the lower plane. In the principal plane there are intercepts with the asymptotes $\theta = 7\pi/6, 3\pi/2, 11\pi/6$, and none with the asymptotes $\theta = 5\pi/6, \pi/2, \pi/6$.

Figure 6 illustrates the curves of steepest descent for $m = \frac{1}{2}$ and $\eta = 18$, given by Eq. (67). The curve of Fig. 6 passing through the saddle point and asymptotic to the lines $\theta = -7\pi/6, 5\pi/6$ is called C_6'.

Figure 7 illustrates the curves of steepest descent for the case $m = \frac{1}{2}$, $\eta = -18$. The curve of Fig. 7 passing through the saddle point and asymptotic to the lines $\theta = -11\pi/6, \pi/6$ is called C_7'. It passes out of the lower plane into the plane given by

$$\pi/6 - 4\pi \leq \theta \leq \pi/6 - 2\pi,$$

which is called the next lower plane.

When Ric $> \frac{1}{4}$ so that m is complex, the discussion of the curves is more complicated. The quantity v_1 is no longer given by Eqs. (71) and (73). Figures 8, 9, 10 and 11 illustrate the curves corresponding to Eq. (69). These figures show that nothing radical happens to the curves of steepest descent as Ric passes through the value $\frac{1}{4}$.

The curves of Figs. 8 and 9 passing through the saddle point and asymptotic to the lines $\theta = -7\pi/6, 5\pi/6$ are called C_6'', and C_6''', respectively. The curves of Figs. 10 and 11 passing through the saddle point and asymptotic to the lines $\theta = -11\pi/6, \pi/6$ are called C_7'' and C_7''', respectively.

XI. ASYMPTOTIC FORMULAS

The contours C_i in Eq. (49) can be deformed to the contours C_i' without changing the value of the functions $w_i(\eta)$:

$$w_i(\eta) = \int_{C_i'} t^{(m-2)} \exp\left(\frac{i}{3}t^3 + \eta t\right) dt. \qquad (81)$$

The same holds for the contours C_i'' and C_i'''. Consider first the saddle points (62) and (63). Near the saddle point the factor $t^{(m-2)}$ varies slowly compared with the other factor of the integrand, and accordingly it will be taken as constant. The contour C_i' will be replaced by a straight line through the saddle point tangent to C_i' at that point. Using the Taylor expansion (75) for $f(t)$ we get

$$w_i(\eta) \simeq t_1^{(m-2)} e^{f(t_1)} \int_{C_i'} \exp[\tfrac{1}{2}f''(t_1)(t - t_1)^2]\, dt.$$

Use of Eqs. (76) and (77) for $f(t)$ gives the following asymptotic formulas:

$$w_1(\eta) \simeq \pi^{\frac{1}{2}} \exp\left[\frac{i\pi}{4}(m - \tfrac{3}{2})\right] \eta^{\frac{1}{2}(m-5/2)} \times \exp[\tfrac{2}{3}i^{\frac{1}{2}}\eta^{\frac{3}{2}}], \qquad (82a)$$

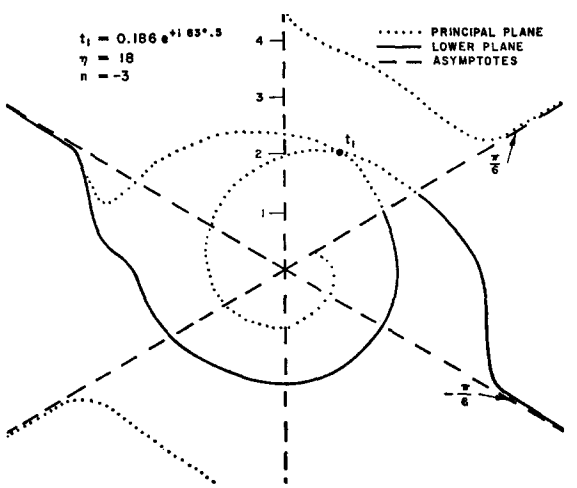


FIG. 9. Curves of steepest descent. The radial coordinate is $\log_{10} 1000r$ plotted vs θ.

$$w_2(\eta) \simeq \pi^{\frac{1}{2}} \exp\left[\frac{i5\pi}{4}(m - \tfrac{3}{2})\right]\eta^{\frac{1}{2}(m-5/2)} \times \exp\left[-\tfrac{2}{3}i^{\frac{1}{2}}\eta^{\frac{3}{2}}\right] \quad (82b)$$

as $\eta \to +\infty$, and

$$w_1(\eta) \simeq \pi^{\frac{1}{2}} \exp\left[\frac{i3\pi}{4}(m + \tfrac{1}{2})\right] |\eta|^{\frac{1}{2}(m-5/2)} \times \exp\left[-\tfrac{2}{3}i^{\frac{3}{2}}|\eta|^{\frac{3}{2}}\right], \quad (83a)$$

$$w_3(\eta) \simeq \pi^{\frac{1}{2}} \exp\left[i\frac{\pi}{4}\left(7m - \frac{17}{2}\right)\right]\eta^{\frac{1}{2}(m-5/2)} \times \exp\left[\tfrac{2}{3}i^{\frac{3}{2}}|\eta|^{\frac{3}{2}}\right] \quad (83b)$$

as $\eta \to -\infty$.

Next consider the saddle points given by Eqs. (70) and (72). When $|\eta|$ is large the only appreciable contribution to the integral in Eq. (81) comes from the neighborhood of the saddle point. Since $|t_1|$ is small there, the cubic term in $F(t)$ is neglected, and the following formula used:

$$F(t) = \eta t + (m - 2)\ln t.$$

Expanding $F(t)$ in a Taylor series around the point t_1 and retaining no terms higher than the quadratic ones give the following approximation to Eq. (81):

$$w_i(\eta) \simeq e^{(2-m)}(2 - m)^{-(2-m)}\eta^{(2-m)} \times \int_{C_i'} \exp\left[\frac{\eta^2}{2(2 - m)}(t - t_1)^2\right] dt,$$

when $i = 6, 7$. Replacing the contour C_i' by a straight line through the saddle point, tangent to the contour there, we finally get

$$w_6(\eta) \simeq (2\pi)^{\frac{1}{2}} i e^{(2-m)}(2 - m)^{(m-\frac{1}{2})}\eta^{(1-m)} \quad (84)$$

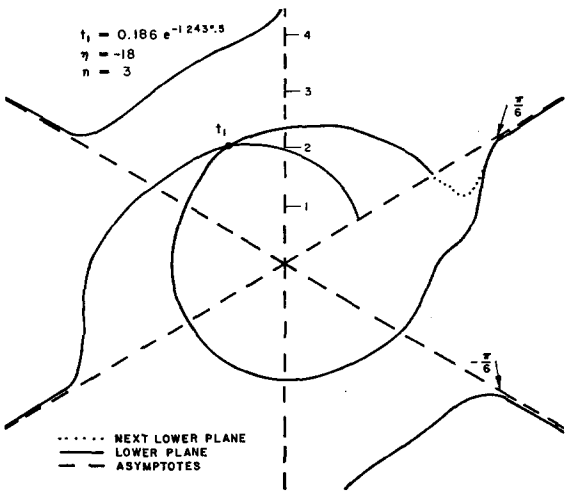


FIG. 10. Curves of steepest descent. The radial coordinate is $\log_{10} 1000r$ plotted vs θ.

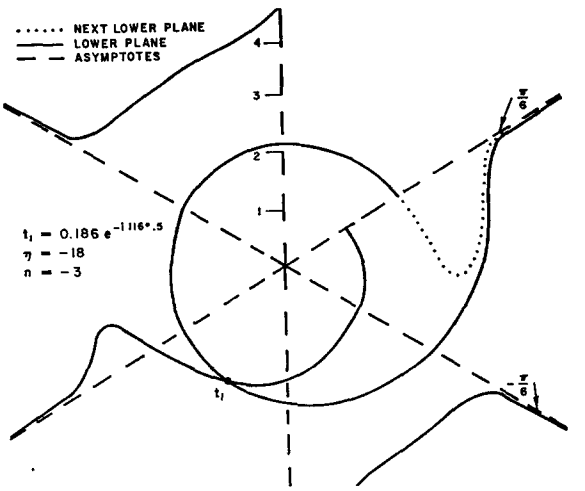


FIG. 11. Curves of steepest descent. The radial coordinate is $\log_{10} 1000r$ plotted vs θ.

as $\eta \to +\infty$, and

$$w_7(\eta) \simeq -(2\pi)^{\frac{1}{2}} i e^{(2-m)}(2 - m)^{(m-\frac{1}{2})}|\eta|^{(1-m)}e^{-m\pi i} \quad (85)$$

as $\eta \to -\infty$.

The same expressions are obtained if the contour C_6'' or C_6''' is used instead of C_6', and if C_7'' or C_7''' is used instead of C_7'. It follows that the above expressions are valid for m real or complex.

The behavior of $w_2(\eta)$ as $\eta \to -\infty$ can be obtained from Eqs. (85), (51), (83), and (50). The result is

$$w_2(\eta) \simeq -\pi^{\frac{1}{2}} \exp\left[i\frac{3\pi}{4}(m + \tfrac{1}{2})\right]|\eta|^{\frac{1}{2}(m-5/2)} \times \exp\left[-\tfrac{2}{3}i^{\frac{3}{2}}|\eta|^{\frac{3}{2}}\right]. \quad (86)$$

Similarly the behavior of $w_3(\eta)$ as $\eta \to +\infty$ can be obtained from Eqs. (84), (52), (82), (53), and (50):

$$w_3(\eta) \simeq -\pi^{\frac{1}{2}} \exp\left[i\frac{3\pi}{4}(3m - \tfrac{1}{2})\right]\eta^{\frac{1}{2}(m-5/2)} \times \exp\left[\tfrac{2}{3}i^{\frac{1}{2}}\eta^{\frac{3}{2}}\right]. \quad (87)$$

All of the solutions defined thus far behave as an increasing exponential on at least one side of the Stokes point. A solution, $w_8(\eta)$, may be defined, however, which does not have this property. It is given by

$$w_8 = w_4 + (1 - e^{-2\pi m i})w_3. \quad (88)$$

An alternative expression for w_8 may be obtained by the use of Eqs. (50), (52), and (53):

$$w_8 = (e^{-2\pi m i} - 1)w_2 + w_6. \quad (89)$$

From Eqs. (82b), (84), and (89) we obtain the

asymptotic behavior

$$w_8 \simeq (2\pi)^{\frac{1}{2}} i e^{(2-m)} (2 - m)^{(m-\frac{3}{2})} \eta^{(1-m)}, \qquad (90)$$

as $\eta \to +\infty$. From Eqs. (83b), (85), (51), and (88), the behavior

$$w_8 \simeq -(2\pi)^{\frac{1}{2}} i e^{(2-m)} (2 - m)^{(m-\frac{3}{2})} |\eta|^{(1-m)} e^{m\pi i} \qquad (91)$$

follows as $\eta \to -\infty$.

Although the choice of phase $\eta = |\eta| e^{i\pi}$ was made when η is negative, the asymptotic formulas obtained in this section are independent of the choice of phase of η. Note however that formulas (86), (83b), and (91) may be formally obtained from (82b), (87), and (90) by substituting $\eta = |\eta| e^{-i\pi}$ in the latter. This is the same rule obtained in the case when gravity, density stratification, and heat conduction are neglected.[4]

XII. DISCUSSION

The main results obtained here are perhaps the connection formulas for the asymptotic solutions across the turning (Stokes) point deduced in the last section. However, this was explicitly obtained only for the case of a Prandtl number of one, the more general case requiring a detailed examination of Whittaker's confluent hypergeometric functions. A logical sequel to obtaining the connection formulas would be the detailed solution of an eigenvalue problem using the asymptotic formulas. This has not been attempted as yet.

Bergman's Integral Operator Method in Generalized Axially Symmetric Potential Theory

R. P. GILBERT
U. S. Naval Ordnance Laboratory, White Oak, Silver Spring, Maryland and University of Maryland, College Park, Maryland*

(Received 5 June 1963)

This paper contains a study of properties of solutions to the equation of generalized axially symmetric potentials. These potentials play an important role in many aspects of mathematical physics, in particular to an understanding of compressible flow in the transonic region. The ideas that have been basic in this investigation are contained in the integral operator method of Bergman. This method allows one to transplant certain properties of analytic functions to the solutions of linear partial differential equations. Results are obtained concerning singularities, residues, bounds, and growth of entire solutions, which are analogous to those found in classical function theory.

I. INTRODUCTORY REMARKS

GENERALIZED Axially Symmetric Potential Theory (GASPT) is the name that Weinstein[1-4] has given to the study of solutions of the partial differential equation

$$L_K[u] = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{K}{y}\frac{\partial u}{\partial y} = 0, \quad K > 0. \tag{1.1}$$

One is initially led to consider a differential equation of this type, when one considers those solutions of the n-dimensional Laplace equation $\partial^2 u/\partial x_1^2 + \cdots + \partial^2 u/\partial x_n^2 = 0$, which depend solely on the variables $x = x_1, y = (x_2^2 + \cdots + x_n^2)^{\frac{1}{2}}$. In this case $K = n - 2$.

Recently, this author has written several papers on GASPT[5-9] using function-theoretic methods similar to those developed by S. Bergman in the study of harmonic functions in three variables.[10-13] In this paper these results shall be reformulated and presented as a unified approach to the study of the GASPT equation. It will be clear from the development, that these methods are immediately extendable to certain other singular partial differential equations, such as the Euler–Poisson–Darboux equation,

$$\frac{\partial^2 u}{dx^2} + \frac{\nu}{x}\frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial y^2} + \frac{\mu}{y}\frac{\partial u}{\partial y}, \tag{1.2}$$

the doubly symmetric equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\mu}{x}\frac{\partial u}{\partial x} + \frac{\nu}{y}\frac{\partial u}{\partial y} = 0, \tag{1.3}$$

and the symmetric Helmholtz equation[14]

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\nu}{y}\frac{\partial u}{\partial y} + K^2 u = 0. \tag{1.4}$$

It should also be mentioned that the methods used here employ integral operators which are not of the same type that Bergman[15] uses for partial differential equations in two variables with *entire coefficients*. That is, there is no obvious connection between Bergman's Integral Operator of the first kind, and our operators; however, there is a close connection with the Whittaker–Bergman Operator, which transforms functions of two complex variables into solutions of the Laplace equation in three variables

$$H(X) = B_3[f], \qquad B_3[f] = \frac{1}{2\pi i}\int_{\mathcal{L}} f(t, \zeta)\frac{d\zeta}{\zeta},$$

$$t = [-(x_1 - ix_2)\tfrac{1}{2}\zeta + x_3 + (x_1 + ix_2)(1/2\zeta)], \tag{1.5}$$

$$||X - X^0|| < \epsilon, \; X \equiv (x_1, x_2, x_3), \; X^0 \equiv (x_1^0, x_2^0, x_3^0),$$

where $\mathcal{L}$ is a closed differentiable arc in the ζ plane, and $\epsilon > 0$ is sufficiently small. The connection, between the operator $B_3[f]$, and the operators we may introduce for Eqs. (1.1)–(1.4) is that there

* Supported in part by Air Force Office of Scientific Research under Grant 400-63.

[1] A. Weinstein, Trans. Am. Math. Soc. **63**, 342 (1948).
[2] A. Weinstein, Bull. Am. Math. Soc. **59**, 20 (1953).
[3] A. Weinstein, *The Method of Axial Symmetry in Partial Differential Equations* (Atti del convegno internazionale sulle Equazioni alle derivate parziali, Trieste, 1954), pp. 1–10.
[4] A. Weinstein, Proc. Naval Ord. Lab. Aeroballistics Symp. **73**, 73 (1949).
[5] R. Gilbert, Arch. Ratl. Mech. Anal. **6**, 171 (1960).
[6] R. Gilbert, J. Reine Angew. Math. **212**, 158 (1963).
[7] R. Gilbert, Am. J. Math. **84**, 475 (1962).
[8] R. Gilbert, Ann. Math. Pura Appl. **61**, 337 (1963).
[9] R. Gilbert, Technical Note BN-298, Univ. Maryland (1962), to appear in Journal Math. Mech., July (1964).
[10] S. Bergman, Ergeb. Math. Grenzgeb. **23**, 40 (1960).
[11] S. Bergman, Math. Ann. **99**, 629 (1928); **101**, 534 (1929).
[12] S. Bergman, Scripta Math. **26**, 5 (1961).
[13] S. Bergman, Bull. Amer. Math. Soc. **49**, 163 (1943).
[14] Extensions of the methods presented in this paper are presently being investigated by the author and his associates for the cases of Eqs. (1.2), (1.3), and (1.4).
[15] S. Bergman, Arch. Ratl. Mech. Anal. **8**, 207 (1961).

exists a *special* or *generating variable*, such that powers of this variable are transformed into solutions of the partial differential equation by the integral operator. For instance, in the case of $B_3[f]$ we see that the powers of t correspond to the following solutions[16-18]:

$$r^n P_n(\cos\theta) = B_3[t^n] \equiv \frac{1}{2\pi i}\int_{|\zeta|=1} t^n \frac{d\zeta}{\zeta}, \qquad (1.6)$$

$$r^2 = x^2 + y^2, \quad \cos\theta = x/r.$$

For the GASPT equation the set $r^n C_n^{\frac{1}{2}K}(\cos\theta)$, where the $C_n^{\frac{1}{2}K}(\xi)$ are Gegenbauer or ultraspherical harmonics,[15,18,19] forms a complete system of solutions regular about the origin. This fact, along with the integral representation for the Gegenbauer polynomials,

$$r^n C_n^{\mu}(\cos\theta) = \frac{2^{1-2\mu}}{n!}\frac{\Gamma(2\mu+n)}{\Gamma(\mu)^2} \times \int_0^{\pi} [x + iy\cos\varphi]^n (\sin\varphi)^{2\mu-1}\, d\varphi, \qquad (1.7)$$

suggests the introduction of the following integral operator, where we have as the *generating variable* $\sigma = x + iy\cos\varphi$:

$$u(z) \equiv u(x, y) = \mathfrak{A}_K[f],$$

$$\mathfrak{A}_K[f] \equiv \alpha_K \int_{\mathcal{L}} f(\sigma)[\zeta - \zeta^{-1}]^{K-1}\frac{d\zeta}{\zeta},$$

$$\alpha_K = \frac{4}{(4i)^K\Gamma(\frac{1}{2}K)^2}, \quad \sigma = x + \tfrac{1}{2}iy[\zeta + \zeta^{-1}], \qquad (1.8)$$

$$\mathcal{L} \equiv \{\zeta \mid \zeta = e^{i\varphi}, 0 \le \varphi \le \pi\},$$

$$|z - z^0| < \epsilon, \qquad z \equiv x + iy,$$

where z^0 is an initial point of definition,[20] and $\epsilon > 0$ is sufficiently small. We recognize immediately from (1.7) and (1.8), that the powers of σ, $\{\sigma^n\}$, are mapped onto the complete system of functions $[n!/\Gamma(K+n)]r^n C_n^{\frac{1}{2}K}(\cos\theta)$; consequently, if $f(\sigma)$ is an analytic function, regular about the origin, $f(\sigma) = \sum_{n=0}^{\infty} a_n\sigma^n$, then it is mapped by $\mathfrak{A}_K[f]$ onto the solution,[21]

$$u(z) = w(r, \cos\theta) = \sum_{n=0}^{\infty}\frac{n!}{\Gamma(K+n)} a_n r^n C_n^{\frac{1}{2}K}(\cos\theta). \qquad (1.9)$$

We should like to stress here that this representation for $u(z)$ is valid only *in the small* of the origin, and it is part of the theory of the integral operator method to extend the representation *in the large*.

It should be mentioned that Bateman,[22] Mackie,[23] Erdélyi,[24] and Henrici,[25] have considered integral representations of the form (1.8). In general, however, their approach to the study of integral operators has been different from ours. We shall, nevertheless, attempt to point out any similarities which may exist.

The integral representations of Vekua,[26,27] on the other hand, have no obvious connections with ours. Vekua obtains a representation for elliptic equations by formally transforming them into hyperbolic form and using the Riemann function representation. This method has the advantage of permitting one to solve boundary value problems; however, the representation formulas are somewhat more complicated than those given by the expression (1.8). Consequently, it would be advantageous if a connection could be found between the two methods,[28] as has been done by Diaz and Ludford[29,30] for Bergman's operator of the first kind and Riemann function representations for *regular* linear hyperbolic equations. (For a discussion of Bergman's operator method as applied to compressible fluid flow see Krzywoblocki.[31])

As a concluding remark, we mention that the theory of integral operators, as has been developed by Bergman and is presented here, formalizes a procedure by which analytic functions of a complex variable are transformed into solutions of partial

---

[16] Bateman Manuscript Project, *Higher Transcendental Functions*, edited by A. Erdélyi (McGraw-Hill Book Company, Inc., New York, (1953), Vol. II.

[17] R. Gilbert, Pac. J. Math. 10, 1243 (1960).

[18] E. W. Hobson, The Theory of Spherical and Ellipsoidal Harmonics, (Cambridge University Press, London, 1931).

[19] See Ref. 16, Vol. I.

[20] We assume the existence of a point z_0, such that in a suitably small neighborhood of this point, $N(z) \equiv \{z \mid |z - z_0| < \epsilon\}$, the integral representation is defined.

[21] We refer to $u(z) = u(x, y)$ as the $\mathfrak{A}_K$ *associate* of $f(\sigma)$.

[22] H. Bateman, *Partial Differential Equations of Mathematical Physics* (Dover Publications, Inc., New York, 1944).

[23] A. G. Mackie, J. Rat. Mech. Anal. 4, 733 (1955).

[24] A. Erdélyi, Commun. Pure Appl. Math. 9, (1956).

[25] P. Henrici, *Complete Systems of Solutions for a Class of Singular Elliptic Partial Differential Equations, Boundary Value Problems in Differential Equations*, edited by R. E. Langer (University of Wisconsin Press, Madison, Wisconsin 1960).

[26] I. N. Vekua, *Novye Metody rešenija Ellipličeskikh Uravnenij*, (OGIZ, Mockow and Leningrad, 1948).

[27] P. Henrici, Z. Angew. Math. Physik 8, 169 (1957).

[28] This problem is currently being investigated by the author.

[29] J. B. Diaz and G. S. S. Ludford, Quart. Appl. Math. 14, 428 (1957).

[30] J. B. Diaz and G. S. S. Ludford, Quart. Appl. Math. 12, 422 (1955).

[31] M. Z. V. Kryzwoblocki, *Bergman's Linear Integral Operator Method in the Theory of Compressible Fluid Flow* (Springer-Verlag, Vienna, 1960).

differential equations. The importance of this theory is three-fold, (i) to transplant theorems concerning analytic functions into theorems about solutions of partial differential equations, (ii) to obtain representation theorems for solutions, and (iii) to consider the analytic continuation of solutions.

II. INVERSE OPERATORS OF THE FIRST AND SECOND KIND

As we have just mentioned at the close of the previous section, a major use of the integral operator method is the transplanting of theorems concerning analytic functions into theorems about solutions of partial differential equations. These theorems are usually stated in terms of the $\mathfrak{a}_K$ associates, and not expressed solely in terms of the solutions. In order to overcome this is it necessary to obtain an inverse operator to $\mathfrak{a}_K[f]$, which maps solutions of the GASPT equation back onto their $\mathfrak{a}_K$ associates, This may be done in several ways.

Inverse Operators of the First Kind

In order to obtain the inverse operator $\mathfrak{a}_K^{-1}[u]$, we continue the arguments x, y to complex values, and introduce the complex variables $r = +(x^2+y^2)^{\frac{1}{2}}$, $\xi = x/r$. (When x, y are real, $\xi = \cos\theta$.) Next, we define the kernel

$$K\left(\frac{\sigma}{r}, \xi\right) = \left[\frac{\Gamma(K)}{\Gamma[\frac{1}{2}(K+1)]}\right]^2 \frac{(1-\xi^2)^{\frac{1}{2}(K-1)}}{2^K} \times \sum_{m=0}^{\infty} (2m+K)\left(\frac{\sigma}{r}\right)^m C_m^{\frac{1}{2}K}(\xi). \tag{2.1}$$

Then, from the orthogonality relation for the Gegenbauer polynomials[16]

$$\int_{-1}^{+1} (1-\xi^2)^{\frac{1}{2}K} C_n^{\frac{1}{2}K}(\xi) C_m^{\frac{1}{2}K}(\xi)\, d\xi = \frac{2^K \Gamma(K+n)}{(2n+K)n!}\left[\frac{\Gamma[\frac{1}{2}(K+1)]}{\Gamma(K)}\right]^2 \delta_{mn}, \tag{2.2}$$

it may be seen, that

$$\sigma^m = \int_{-1}^{+1} \left\{\frac{m!}{\Gamma(m+K)} r^m C_m^{\frac{1}{2}K}(\xi)\right\} K\left(\frac{\sigma}{r}, \xi\right) d\xi. \tag{2.3}$$

Consequently, if $f(\sigma) = \sum_{n=0}^{\infty} a_n\sigma^n$, and

$$W(r, \xi) = \mathfrak{a}_K[f] = \sum_{n=0}^{\infty} \frac{a_n n!}{\Gamma(n+K)} r^n C_n^{\frac{1}{2}K}(\xi),$$

then one has

$$f(\sigma) = \sum_{n=0}^{\infty} a_n\sigma^n = \int_{-1}^{+1} W(r, \xi) K\left(\frac{\sigma}{r}, \xi\right) d\xi, \tag{2.4}$$

where $W(r, \xi) = u(x, y)$ with x, y replaced by $r\xi$, $r(1-\xi^2)^{\frac{1}{2}}$, respectively. To justify this procedure we note that $K(\sigma/r, \xi)$ may be summed formally, whenever $|\sigma/r| \leq \rho < 1$;

$$K\left(\frac{\sigma}{r}, \xi\right) = 2^{-K} \frac{\Gamma(K)^2}{\Gamma[\frac{1}{2}(K+1)]^2} (1-\xi^2)^{\frac{1}{2}(K-1)} \times \left\{t^{1-K} \frac{\partial}{\partial t}\left[t^K \sum_{n=0}^{\infty} t^{2n} C_n^{\frac{1}{2}K}(\xi)\right]\right\}_{t=(\sigma/r)^{\frac{1}{2}}}$$
$$= K 2^{-K} \frac{\Gamma(K)^2}{\Gamma[\frac{1}{2}(K+1)]^2} \frac{(1-\xi^2)^{\frac{1}{2}(K-1)}(1-\sigma^2/r^2)}{[1-2\xi(\sigma/r)+\sigma^2/r^2]^{\frac{1}{2}K+1}}. \tag{2.5}$$

This follows from the classical identity[16]

$$\sum_{n=0}^{\infty} t^n C_n^{\frac{1}{2}K}(\xi) = (1-2\xi t+t^2)^{-\frac{1}{2}K}, \text{ for } |t| < 1. \tag{2.6}$$

The constant coefficient in (2.5) may be reduced by means of the Legendre duplication formula

$$\frac{\Gamma(2K)}{\Gamma(K)} = \frac{2^{2K-1}}{\pi^{\frac{1}{2}}} \Gamma(K+\tfrac{1}{2}),$$

such that one obtains

$$K\left(\frac{\sigma}{r}, \xi\right) = \frac{\beta_K (1-\xi^2)^{\frac{1}{2}(K-1)}}{[1-2\xi(\sigma/r)+\sigma^2/r^2]^{\frac{1}{2}K+1}}, \qquad \beta_K = \frac{K}{\pi}\Gamma(\tfrac{1}{2}K)^2 2^{K-2}. \tag{2.7}$$

Lemma 2.1. *Let* $W(r, \cos\theta) = u(x, y)$ *be a* GASPT *function element defined about the origin by the integral operator* $\mathfrak{a}_K[f]$, *where* $f(\sigma) = \sum_{n=0}^{\infty} a_n\sigma^n$. *Then there exists an inverse integral operator*

$$\mathfrak{a}_K^{-1}[u] = \int_{\mathfrak{C}^{-1}}^{+1} W(r, \xi) K\left(\frac{\sigma}{r}, \xi\right) d\xi,$$

$K(\sigma/r, \xi)$ *is the kernel defined in* (2.6) *and* $\mathfrak{C}$ *is a smooth curve joining* -1 *to* $+1$, *which maps* $u(x, y)$ *back onto its* $\mathfrak{a}_K$ *associate* $f(\sigma)$.

It is possible to find another inverse operator for $\mathfrak{a}_K$ by considering $u(x, y)$ on the space $r^2 = zz^* = 0$. We assume as before, that

$$f(\sigma) = \sum_{n=0}^{\infty} a_n\sigma^n, \quad \text{and} \quad u = \mathfrak{a}_K[f] = \sum_{n=0}^{\infty} \frac{a_n n!}{\Gamma(K+n)} r^n C_n^{\frac{1}{2}K}(\xi).$$

Since,

$$C_n^{\frac{1}{2}K}(\xi) = \sum_{m=0}^{n} \frac{(-1)^m \Gamma(\frac{1}{2}K+m)\Gamma(n+K+m)(\frac{1}{2}-\frac{1}{2}\xi)^m}{m!\,\Gamma(\frac{1}{2}K)\Gamma(2m+K)(n-m)!}, \qquad \xi = \frac{x}{r}, \tag{2.8}$$

as $r \to 0$, we have

$$\lim_{r\to 0} r^n C_n^{\frac{1}{2}K}(\xi) = \frac{\Gamma(\frac{1}{2}K + n)}{n!\,\Gamma(\frac{1}{2}K)} (\tfrac{1}{2}x)^n \quad \text{on} \quad zz^* = 0; \quad (2.9)$$

we may rewrite this as

$$\left[\frac{\Gamma(\frac{1}{2}K)^2 n!}{\Gamma(K+n)} (2r)^n C_n^{\frac{1}{2}K}(\xi)\right]_{zz^*=0} = [B(\tfrac{1}{2}K + n, \tfrac{1}{2}K)x^n]_{zz^*=0}, \quad (2.10)$$

for $K > 0$, where $B(p, q)$ is the beta function.

Consequently, one has on $E\{zz^* = 0\}$,

$$\begin{aligned}\Gamma(\tfrac{1}{2}K)^2 W(2r, \xi) &= \Gamma(\tfrac{1}{2}K)^2 \sum_{n=0}^{\infty} \frac{a_n n!}{\Gamma(K+n)} (2r)^n C_n^{\frac{1}{2}K}(\xi) \\ &= \sum_{n=0}^{\infty} B(\tfrac{1}{2}K + n, \tfrac{1}{2}K) a_n x^n \\ &= \sum_{n=0}^{\infty} \int_0^1 t^{n+\frac{1}{2}K-1}(1-t)^{\frac{1}{2}K-1} a_n x^n \, dt \\ &= \int_0^1 \left[\sum_{n=0}^{\infty} a_n (xt)^n\right] t^{\frac{1}{2}K-1}(1-t)^{\frac{1}{2}K-1} \, dt. \end{aligned} \quad (2.11)$$

The inversion of the order of summation and integration is valid for all $|x|$, $\ni x \leq \rho_0 <$ the radius of convergence of $f(\sigma)$; this result follows by considering Hadamard's theorem concerning the multiplication of singularities with respect to the functions

$$\sum_{n=0}^{\infty} a_n x^n, \quad \sum_{n=0}^{\infty} \frac{\Gamma(K+n)}{n!} x^n,$$

and

$$\sum_{n=0}^{\infty} a_n \frac{\Gamma(K+n)}{n!} x^n.$$

On the set $E\{zz^* = 0\}$, we may consider (2.10) as an integral equation for $f(xt)$, that is,

$$u(2x, 2y) = \Gamma(\tfrac{1}{2}K)^{-2} \int_0^1 f(xt) t(1-t)^{\frac{1}{2}K-1} \, dt, \quad (2.12)$$

where $u(2x, 2y) = w(2r, \xi)$. If we write $F(x) = u(2x, 2y)$ on $E\{zz^* = 0\}$, then (2.12) may be written as

$$F(x) = \Gamma(\tfrac{1}{2}K)^{-2} \int_0^1 f(xt) t(1-t)^{\frac{1}{2}K-1} \, dt,$$

and by setting $\tau = xt$, $G(x) = x^{K-1}F(x)$, $g(\tau) = \tau^{\frac{1}{2}K-1} f(\tau)$, (2.12) may be rewritten as a convolution integral,

$$G(x) = \Gamma(\tfrac{1}{2}K)^{-2} \int_0^x g(\tau)(x-\tau)^{\frac{1}{2}K-1} \, d\tau. \quad (2.13)$$

Equation (2.13) may be solved by means of Laplace transforms when $0 < K < 2$ in the form

$$f(\tau) = C_K \tau^{1-\frac{1}{2}K} \frac{d}{d\tau} \int_0^\tau (\tau - x)^{-\frac{1}{2}K} x^{K-1} F(x) \, dx, \qquad C_K \equiv \Gamma(\tfrac{1}{2}K)^2 \frac{1}{\pi} \sin \tfrac{1}{2}(\pi K), \quad (2.14)$$

which may be expressed as

$$f(\sigma) = C_K \sigma^{1-\frac{1}{2}K} \frac{d}{d\sigma} \left\{\sigma^{\frac{1}{2}K} \int_0^1 (1-\eta)^{-\frac{1}{2}K} \eta^{K-1} \times [u(2\eta\sigma, 2\xi\sigma)] \, d\eta\right\}, \qquad \eta^2 + \xi^2 = 0. \quad (2.15)$$

Lemma 2.2. Let $u(x, y)$ be a GASPT *function element defined about the origin by the integral operator $\mathfrak{A}_K[f]$, $(0 < K < 2)$, where $f(\sigma)$ is analytic-regular about the origin. Then there exists an inverse operator for $\mathfrak{A}_K[f]$ of the form*

$$\mathfrak{A}_K^{-1} = C_K \sigma^{1-\frac{1}{2}K} \frac{d}{d\sigma} \left\{\sigma^{\frac{1}{2}K} \int_0^1 (1-\eta)^{-\frac{1}{2}K} \eta^{K-1} \times [u(2\eta\sigma, 2\xi\sigma)] \, d\eta\right\}, \qquad \eta^2 + \xi^2 = 0.$$

which maps $u(x, y)$ back on to its $\mathfrak{A}_K$-associate $f(\sigma)$.

The reason for introducing inverse operators is that by employing an inverse operator it is sometimes possible to obtain theorems concerning GASPT functions, which are independent of their $\mathfrak{A}_K$ associates. To see how this may be done, let $\mathfrak{F}\{f(\sigma); P\}$ be a class of analytic functions with property P; furthermore, let $\mathfrak{G}\{u(x, y); P^*\}$ be the class of GASPT-function elements, which is the image of $\mathfrak{F}$ under the transformation $u(x, y) = \mathfrak{A}_K[f]$, $f \in \mathfrak{F}$. Let $\mathfrak{A}_K^{-1}[u]$ be an inverse operator for $\mathfrak{A}_K[f]$, that is $\mathfrak{A}_K^{-1}[u]$ is an integral operator, which transforms u into its $\mathfrak{A}_K$ associate. Now, suppose $\mathfrak{IC}\{f; P^{**}\}$ is the image of $\mathfrak{G}$ under the inverse mapping $f = \mathfrak{A}_K^{-1}[u]$; if $\mathfrak{F} \cap \mathfrak{IC} \neq 0$, we may then extract a result concerning GASPT-function elements, which is independent of the $\mathfrak{A}_K$ associates. This principle is illustrated in the later sections.

III. THEOREMS CONCERNING SINGULARITIES OF GASPT-FUNCTION ELEMENTS

Let us assume that $f(\sigma)$ is analytic about the origin; then

$$u(x, y) = \alpha_K \int_{\mathfrak{L}} f\left(x + \frac{iy}{2}[\zeta + \zeta^{-1}]\right)(\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}, \quad (3.1)$$

where $\mathfrak{L} = E\{\zeta \mid \zeta = e^{i\alpha}; 0 \leq \alpha \leq \pi\}$ defines a GASPT-function element in some neighborhood of the origin, $\mathfrak{N}(0)$. The representation (3.1) is actually

valid for all points $z = x + iy$, which may be reached by continuation along a contour γ, providing that no point of γ corresponds to a singularity of the integrand on the path of integration, $\mathfrak{L}$. *We refer to this set of points* $\mathfrak{D}$ as the *initial domain of definition* for $u(x, y) = \mathfrak{A}_K[f]$. It is possible, however, sometimes to extend this domain of definition by continuously deforming the path of integration $\mathfrak{L}$ into $\mathfrak{L}'$ in such a manner, that it does not cross over a singularity of the integrand. For instance, as we continue $u(x, y)$ along γ in the z plane, the singularities of $f(x + \frac{1}{2}(iy)[\zeta + \zeta^{-1}])$ move in the ζ plane; however, as long as these singularities do not cross over $\mathfrak{L}'$ we may continue $u(x, y)$ further.

Let us consider the special case, where the only finite singularity of $f(\sigma)$ is at $\sigma = \alpha$; here we may represent the singularity manifold of $u(x, y)$ in complex two-space $\mathbf{C}^2$ as

$$S(x, y; \zeta) \equiv (x - \alpha)\zeta + \tfrac{1}{2}(iy)(\zeta^2 + 1) = 0. \tag{3.2}$$

Suppose we have been able to continue $u(x, y)$ up to the point z_0 on γ, but at this point there is a singularity (one of the two roots $\zeta = -((x-\alpha)_{\pm}[(x-\alpha)^2+y^2]^{\frac{1}{2}}/iy)$ about to cross over $\mathfrak{L}'$. Let us call the singularity $\zeta = \zeta_0$; then in a suitably small neighborhood of ζ_0, say $|\zeta - \zeta_0| < \epsilon$, $S(x_0, y_0; \zeta)$ vanishes only at $\zeta = \zeta_0$. If ζ_0 is a simple zero, we may describe the singularity locally by setting

$$S(x_0, y_0; \zeta) \cong (\zeta - \zeta_0)(\partial S/\partial\zeta)(x_0, y_0; \zeta_0). \tag{3.3}$$

In this case it is always possible to avoid the singularity at $\zeta = \zeta_0$, by deforming $\mathfrak{L}'$ so that it follows a portion of the circle $|\zeta - \zeta_0| = \frac{1}{2}\epsilon$ about ζ_0. From this we have the following results:

Lemma 3.1. *Let $f(\sigma)$ be an analytic function, whose only finite singularity is at $\sigma = \alpha$. Then the* GASPT-*function element, $u(x, y) = \mathfrak{A}_K[f]$, is regular for all points which do not lie on the intersection,*

$$E\{Z \mid S(x, y; \zeta) = 0\} \bigcap E\{Z \mid (\partial S/\partial\zeta)(x, y; \zeta) = 0\} = 0. \tag{3.4}$$

Theorem 3.2. *If the only finite singularities of the $\mathfrak{A}_K$ associate for a* GASPT-*function element $u(x, y)$ is at α, then the only possible singularities of $u(x, y)$ lie at α, and $\bar{\alpha}$.*

We may obtain the last result by computing

$$\partial S/\partial\zeta \equiv (x - \alpha) + iy\zeta = 0,$$

and eliminating ζ from (3.2).

It is possible for us, however, to show that $z = \alpha$, $\bar{\alpha}$, are the actual singularities of $u(x, y)$. This may be done by using the inverse operator,

$$f(\sigma) = \mathfrak{A}_K^{-1}[u(x, y)] = \beta_K\left(1 - \frac{\sigma^2}{r^2}\right) \times \int_{\mathfrak{C}^{-1}}^{+1} \frac{W(r, \xi)(1 - \xi^2)^{\frac{1}{2}K}\, d\xi}{(1 - 2\xi\sigma/r + \sigma^2/r^2)^{\frac{1}{2}K+1}}, \tag{3.5}$$

$W(r, \xi) = u(r\xi, r(1 - \xi^2)^{\frac{1}{2}})$, and considering what are the possible singularities of $f(\sigma)$ if $u(x, y)$ is singular at $z = \alpha, \bar{\alpha}$.

Lemma 3.3. *Let $u(x, y) = \mathfrak{A}_K[f]$ be a* GASPT-*function element defined in the small of the origin. Furthermore, let $W(r, \xi) = u(r\xi, r(1 - \xi^2)^{\frac{1}{2}})$, and let $Z = E^1[\xi = \psi(r)]$ be the singularity manifold of $W(r, \xi)$ in $\mathbf{C}^2$ (complex two-space); then the function $f(\sigma) = \mathfrak{A}_K^{-1}[u(x, y)]$ is regular at σ providing this point does not lie on the intersection*

$$E\{\Phi(\sigma; r) \equiv r^2 - 2\sigma r\psi(r) + \sigma^2 = 0\} \bigcap E\{\partial\psi/\partial r = 0\}. \tag{3.6}$$

The proof of this result parallels that of Lemma 3.1. We notice that the singularities of the integrand of (3.4) are of two different kinds; for instance, the singularities of the kernel $K(\sigma/r, \xi)$ (see Eq. 2.7) move in the ξ plane as we attempt to continue $f(\sigma)$ in the σ plane; whereas, the singularities of $W(r, \xi)$ remain fixed. By the same reasoning as Hadamard used in his celebrated theorem concerning the multiplication of singularities,[17,32] we realize that, unless the singularities of $W(r, \xi)$ and $K(\sigma/r, \xi)$ coincide in the ξ plane, it is always possible to deform the contour of integration $\mathfrak{C}$, in order to avoid having a singularity of the kernel cross over it as $f(\sigma)$ is continued along a path in the σ plane. This means, that the only possible singularities of $f(\sigma)$ must lie on the set

$$E\{r^2 - 2\sigma r\xi + \sigma^2 = 0\} \bigcap E\{\xi = \psi(r)\} \equiv E\{r^2 - 2\sigma r\psi(r) + \sigma^2 = 0\}.$$

We may now use the arguments of Lemma 3.1, in order to complete the proof.

Theorem 3.4. *If $W(r, \xi) = U(r\xi, r(1 - \xi^2)^{\frac{1}{2}}) = \mathfrak{A}_K[f]$ has for its only finite singularities the points on the set*

$$Z' = E\{r^2 - 2\alpha\xi r + \alpha^2 = 0\},$$

then $f(\sigma)$ is singular only at $\sigma = \alpha$.

This may be seen from the fact that the only possible singularities of $f(\sigma)$ must lie on the intersection (3.5). However, if $r^2 - 2\alpha\xi r + \alpha^2 = 0$, then $\xi = \psi(r) \equiv \frac{1}{2}(r/\alpha + \alpha/r)$. Now, eliminating r between

[32] J. Hadamard, Acta Math. 22, 55 (1898).

$$\Phi(\sigma; r) = r^2 - \sigma(r^2/\alpha + \alpha) + \sigma^2 = 0, \quad (3.7)$$

and

$$\partial\Phi/\partial r \equiv 2r(1 - \sigma/\alpha) = 0$$

yields $\sigma = \alpha$ or $\sigma = 0$. We disregard $\sigma = 0$, since $u(x, y)$ [and then consequently $f(\sigma)$] is regular at the origin.

We notice that this result can be rephrased in terms of z, and $\bar{z}$. For instance, since

$$(z - \alpha)(\bar{z} - \alpha)z\bar{z} - \alpha(\bar{z} + z) + \alpha^2$$
$$= x^2 + y^2 + 2\alpha x + \alpha^2 = r^2 - 2\alpha r\xi + \alpha^2,$$

if $u(x, y)$ is singular at $z = \alpha$, $\bar{\alpha}$, $f(\sigma)$ may have a singularity at $\sigma = \alpha$. However, one may rewrite the singularity manifold of $u(x, y)$ as

$$(z - \bar{\alpha})(\bar{z} - \bar{\alpha}) = 0, \quad \text{or as} \quad r^2 - 2\bar{\alpha}r\xi + \bar{\alpha}^2 = 0,$$

which implies $f(\sigma)$ may have a singularity at $\sigma = \bar{\alpha}$. Consequently, we have the following results:

Theorem 3.5. *If $u(x, y) = \mathfrak{A}_K[f]$ has a singularity at $z = \alpha$, $\bar{\alpha}$, then $f(\sigma)$ may be singular at $\sigma = \alpha$, $\bar{\alpha}$.*

Theorem 3.6. *The necessary and sufficient conditions for $u(x, y) = \mathfrak{A}_K[f]$ to be singular at $z = \alpha$ is that $f(\sigma)$ be singular at $\sigma = \alpha$, or $\bar{\alpha}$.*[33]

The validity of Theorem (3.6) is realized as follows: each "possible" singularity of $u(x, y)$ is seen to correspond to an actual singularity of its $\mathfrak{A}_K$ associate under the inverse mapping $f(\sigma) = \mathfrak{A}_K^{-1}[u]$.

We are now in a position to *transplant* certain results of analytic function theory to the case of GASPT. We list some theorems which were proven earlier by this author, and indicate their proofs below.[6,7,9]

Theorem 3.7. *Let $\phi(x, y)$, and $\psi(x, y)$ be two* GASPT-*function elements with the series expansion*

$$\phi \equiv \sum_{n=0}^{\infty} a_n r^n C_n^{\frac{1}{2}K}(\cos\theta), \quad \psi \equiv \sum_{n=0}^{\infty} b_n r^n C_n^{\frac{1}{2}K}(\cos\theta).$$

Furthermore, let us suppose, that ϕ and ψ have singularities, respectively, at the point pairs $\{\alpha, \bar{\alpha}\}$, and $\{\beta, \bar{\beta}\}$. Then the GASPT-*function element defined by the development*

$$F \equiv \sum_{n=0}^{\infty} a_n b_n r^n C_n^{\frac{1}{2}K}(\cos\theta),$$

has singularities at either the point pair $\{\alpha\beta, \bar{\alpha}\bar{\beta}\}$, or at $\{\alpha\bar{\beta}, \bar{\alpha}\beta\}$.

Proof: Let $f(\sigma)$, $g(\sigma)$, and $h(\sigma)$ be the $\mathfrak{A}_K$ associates of ϕ, ψ, and F, respectively. From Hadamard's multiplication of singularities theorem we realize that if $f(\sigma)$, $g(\sigma)$ have singularities at δ, γ, respectively, then $h(\sigma)$ may be singular only at $\sigma = \delta\gamma$. In this case, the corresponding singularities of ϕ, ψ, and F are then at the point pairs $\{\delta, \bar{\delta}\}$, $\{\gamma, \bar{\gamma}\}$, and $\{\delta\gamma, \bar{\delta}\bar{\gamma}\}$, respectively. This completes the proof.

Theorem 3.8. *Let $u(x, y)$ be a* GASPT-*function element with the following series development about the origin:*[34]

$$u = \sum_{n=0}^{\infty} a_n r^n C_n^{\frac{1}{2}K}(\cos\theta).$$

Then, u converges uniformly and absolutely in any compact subset of the disk of convergence $|z| < R$, where

$$R^{-1} = \overline{\lim_{n\to\infty}} |a_n|^{1/n}.$$

Proof: We first recognize, that the $\mathfrak{A}_K$ associate for $u(x, y)$ is the function

$$f(\sigma) = \sum_{n=0}^{\infty} \frac{\Gamma(K + n)}{n!} a_n \sigma^n,$$

and that the radius of convergence for $f(\sigma)$ may be computed by the Hadamard formula to be

$$R = \left(\overline{\lim_{n\to\infty}} \left|\frac{\Gamma(K + n)}{n!} a_n\right|^{1/n}\right)^{-1}$$
$$= \left(\lim_{n\to\infty} \left[\frac{\Gamma(K + n)}{n!}\right]^{1/n} \overline{\lim_{n\to\infty}} |a_n|^{1/n}\right)^{-1}$$
$$= (\overline{\lim_{n\to\infty}} |a_n|^{1/n})^{-1}.$$

This may be seen by using the fact, that $\Gamma(K + n)/\Gamma(n + 1) \approx n^{K-1}$ for large n.[35]

Next, since $|\sigma| = |x + iy\cos\alpha| \leq |z|$, when $\zeta = e^{i\alpha}$, we have that

$$\left|u(x, y) - \sum_{n=0}^{N} a_n r^n C_n^{\frac{1}{2}K}(\cos\theta)\right|$$
$$= \left|\sum_{n=N+1}^{\infty} a_n r^n C_n^{\frac{1}{2}K}(\cos\theta)\right|$$
$$\leq \sum_{n=N+1}^{\infty} |a_n r^n C_n^{\frac{1}{2}K}(\cos\theta)|$$
$$\leq \sum_{n=N+1}^{\infty} \frac{|a_n|\,\Gamma(K + n)}{2^{K-1} n!} \int_0^{\pi} |\sigma|^n \sin^{K-1}\alpha\, d\alpha$$
$$\leq \frac{\pi}{2^{K-1}} \sum_{\nu=0}^{\infty} \frac{\Gamma(N + 1 + K + \nu)}{(N + 1 + \nu)!}$$
$$\times |a_{N+1+\nu}| \cdot r^{N+1+\nu} \cdot \left|\frac{z}{r}\right|^{N+1+\nu}.$$

[33] Henrici has also given a similar result; however, he uses a different approach to this problem. P. Henrici, Proc. Am. Math. Soc. 8, 29 (1957).

[34] Erdélyi[24] has also given a similar theorem concerning the disk of convergence for a Gegenbauer series.

[35] E. T. Copson, *Theory of a Function of a Complex Variable* (Oxford University Press, New York, 1935).

where $|z| < r \leq R$. Since $f(\sigma)$ is analytic-regular in the disk $|\sigma| < R$, the terms $[\Gamma(N+1+K+\nu)/(N+1+\nu)!]\,|a_{N+1+\nu}|\,r^{N+1+\nu}$ are bounded above, say, by M. Consequently, we have

$$\left|u(x,y) - \sum_{n=0}^{N} a_n r^n C_n^{\frac{1}{2}K}(\cos\theta)\right| \leq \frac{\pi M}{2^{K-1}} \left|\frac{z}{r}\right|^{N+1} \frac{1}{1-|z/r|},$$

which tends to zero as $N \to \infty$, for $|z| < r \leq R$.

It is clear from Theorem 3 that $u(x, y)$ is regular in the disk $|z| < r \leq R$, and that the first singularity of $u(x, y)$ must occur on the circle $|z| = R$.

Theorem 3.9. *Let $u(x, y)$ be a* GASPT-*function element defined about $z = 0$, with the series development*

$$u(x,y) = \sum_{n=0}^{\infty} a_n r^n C_n^{\frac{1}{2}K}(\cos\theta).$$

Furthermore, let

$$D_\lambda^{(\mu)} \equiv \begin{vmatrix} a_\lambda, & a_{\lambda+1}, & \cdots, & a_{\lambda+\mu} \\ a_{\lambda+1}, & a_{\lambda+2}, & \cdots, & a_{\lambda+\mu+1} \\ \vdots & & & \\ a_{\lambda+\mu}, & a_{\lambda+\mu+1}, & \cdots, & a_{\lambda+2\mu} \end{vmatrix},$$

and $l_\mu \overline{\lim}_{\lambda\to\infty} |D_\lambda^{(\mu)}|^{\frac{1}{\lambda}}$. Then we have the following possibilities:

(a) *If there is a ν such that $l_\nu/l_{\nu-1} = 0$, $u(x, y)$ has 2ν singularities (pole-like branch points) in the entire z plane.*

(b) *If $l_\mu/l_{\mu-1} \to 0$, $u(x, y)$ has just a finite number of singularities in every finite region.*

(c) *If $l_\mu/l_{\mu-1} \to 1/R$, $u(x, y)$ has just a finite number of singularities in the disk $|z| \leq \rho < R$, but an infinite number of singularities in the neighborhood of $|z| = R$.*

Proof: This result is essentially a transplanting of Hadamard's theorem concerning the singularities of meromorphic functions.[36] In order to see that it is true we first note, that the $\mathfrak{A}_K$ associate of $u(x, y)$ is

$$f(\sigma) = \sum_{n=0}^{\infty} \frac{\Gamma(n+K)}{n!} a_n \sigma^n.$$

From our fundamental theorem concerning singularities of GASPT functions, we recall that the necessary and sufficient criteria for $u(x, y)$ to be singular at $z = \alpha$, $\bar{\alpha}$ is for $f(\sigma)$ to be singular at either $\sigma = \alpha$ or $\bar{\alpha}$. With this in mind we apply the Hadamard criteria to $f(\sigma)$ as given above and consider the determinants

$$\Delta_\lambda^{(\mu)} \equiv \begin{vmatrix} \frac{\Gamma(K+\lambda)}{\lambda!} a_\lambda, & \cdots, & \frac{\Gamma(K+\lambda+\mu)}{(\lambda+\mu)!} a_{\lambda+\mu} \\ \vdots & & \\ \frac{\Gamma(K+\lambda+\mu)}{(\lambda+\mu)!} a_{\lambda+\mu}, & \cdots, & \frac{\Gamma(K+\lambda+2\mu)}{(\lambda+2\mu)!} a_{\lambda+2\mu} \end{vmatrix},$$

and the limits $L_\mu = \overline{\lim}_{\lambda\to\infty} |\Delta_\lambda^{(\mu)}|^{1/\lambda}$, If $f(\sigma)$ is meromorphic, then one of the situations (a), (b), or (c) holds, and in this case L_μ exists. Next, we note as before that for large λ, $\Gamma(K+\lambda+\mu)/\Gamma(\lambda+\mu+1) \approx \lambda^{K-1}$. Consequently, as $\lambda \to \infty$, we have

$$\lim_{\lambda\to\infty} |\lambda^{K-1}|^{1/\lambda} \to 1,$$

and

$$L_\mu = \overline{\lim_{\lambda\to\infty}} |\Delta_\lambda^{(\mu)}|^{1/\lambda} = \lim_{\lambda\to\infty} |\lambda^{K-1}|^{1/\lambda} \overline{\lim_{\lambda\to\infty}} |D_\lambda^{(\mu)}|^{1/\lambda} = \overline{\lim_{\lambda\to\infty}} |D_\lambda^{(\mu)}|^{1/\lambda} = l_\mu.$$

This concludes our proof.

IV. GASPT-FUNCTION ELEMENTS WITH MEROMORPHIC ASSOCIATES

In this section we consider the case where the $\mathfrak{A}_K$ associate of $u(x, y)$ is a meromorphic function in the finite σ plane. Then we may express $f(\sigma)$ in terms of its Mittag–Leffler expansion[7,15,37] as

$$f(\sigma) = \sum_{\nu=1}^{\infty} \left[P_\nu\left(\frac{1}{\sigma - b_\nu}\right) - p_\nu(\sigma)\right] + e(\sigma), \qquad (4.1)$$

where the b_ν are the poles of $f(\sigma)$, $P_\nu(1/(\sigma - b_\nu))$ the corresponding principal parts, the $p_\nu(\sigma)$ suitably chosen polynomials to ensure convergence, and $e(\sigma)$ an entire function. Since (4.1) converges uniformly in every compact subset of the σ plane which does not contain the b_ν, one may evaluate the integral representation for $u(x, y) = \mathfrak{A}_K[f]$ by inverting the orders of summation and integration. One has in this instance

$$\begin{aligned} u(x,y) = \mathfrak{A}_K[f] &\equiv \alpha_K \int_{\mathfrak{L}} f(\sigma)(\zeta + \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta} \\ &= \alpha_K \int_{\mathfrak{L}} \left\{\sum_{\nu=1}^{\infty} \left[P_\nu\left(\frac{1}{\sigma - b_\nu}\right) - p_\nu(\sigma)\right] + e(\sigma)\right\}(\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta} \\ &= \sum_{\nu=1}^{\infty} [\Phi_\nu(x,y) - \varphi_\nu(x,y)] + E(x,y), \qquad (4.2) \end{aligned}$$

[36] P. Dienes, *The Taylor Series*, (Oxford University Press, New York, 1931).

[37] L. Ahlfors, *Complex Analysis* (McGraw-Hill Book Company, Inc., New York, 1953).

where $E(x, y)$ is an entire GASPT function, and $\varphi(x, y)$ is a GASPT polynomial.

If K is an odd integer, then the integrals

$$\Phi_\nu(x, y) = \alpha_K \int_{\mathfrak{L}} P_\nu\left(\frac{1}{\sigma - b_\nu}\right)(\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}, \quad (4.3)$$

may be evaluated easily using the calculus of residues if we replace $\mathfrak{L}$ by the unit circle and divide (4.3) by 2. For instance, if

$$P_\nu\left(\frac{1}{\sigma - b_\nu}\right) = \sum_{\mu=1}^{m_\nu} \frac{M_{\mu\nu}}{(\sigma - b_\nu)^\mu}, \quad (4.4)$$

then

$$\begin{aligned}\Phi_\nu(x, y) &= \frac{\alpha_K}{2} \sum_{\mu=1}^{m} M_{\mu\nu} \int_{|\zeta|=1} \frac{(\zeta - \zeta^{-1})^{K-1}}{(\sigma - b_\nu)^\mu} \frac{d\zeta}{\zeta} \\ &= \frac{\pi i}{2} \alpha_K \sum_{\mu=1}^{m_\nu} M_{\mu\nu} \frac{(-1)^\mu}{\mu!} \left(\frac{2}{iy}\right)^K \frac{\partial^\mu}{\partial x^\mu} \\ &\quad \times [(x - b_\nu)^2 + y^2]^{\frac{1}{2}K}. \quad (4.5)\end{aligned}$$

Consequently, if K is an odd integer then $u(x, y)$ has the representation

$$\begin{aligned}u(x, y) = \alpha_K i\pi \sum_{\nu=1}^{\infty} \Bigg\{ \Bigg[\sum_{\mu=1}^{m_\nu} M_{\mu\nu} \frac{(-1)^\mu}{\mu!} \left(\frac{2}{iy}\right)^K \frac{\partial^\mu}{\partial x^\mu} \\ \times ((x - b_\nu)^2 + y^2)^{\frac{1}{2}K} \Bigg] - \varphi_\nu(x, y) \Bigg\} + E(x, y). \quad (4.6)\end{aligned}$$

It may be shown, that this infinite-series representation converges in every compact set in the z plane which does not contain singularities of the $\Phi_\nu(x, y)$.[38]

It is possible to obtain bounds for GASPT-function elements with meromorphic $\mathfrak{a}_K$ associates. This problem has already been considered by Bergman[15] for harmonic functions in three variables, and been discussed for the case of GASPT by the author.[7]

Let us assume that the $\mathfrak{a}_K$ associate of $u(x, y)$ may be written as $f(\sigma) \equiv h_1(\sigma)/h_2(\sigma)$, where the $h_K(\sigma)$ $(K = 1, 2)$ are entire functions of order λ. [That is, if $M_K(r)$ is the maximum modulus of $h_K(\sigma)$ $(K = 1, 2)$ on $|\sigma| = r$ we have $\lambda = \lambda_K \overline{\lim}_{r\to\infty}$ $(\log \log M_K(r)/\log r)$.] It is possible for us to obtain a lower bound for the minimum modulus of an entire function from a theorem of Borel's.[35] For instance, if $m(r)$ is the minimum modulus of $h(\sigma)$ and λ the order, then $m(r) > e^{-r^{\lambda+\epsilon}}$ on circles of arbitrarily large radius for those regions $\mathfrak{R}$ excluding the circles $|\sigma - \sigma_n| \leq |\sigma_n|^{-h}$, where $h > \lambda$.

We may use these results to obtain bounds for GASPT-function elements in their domains of association as follows. Suppose, that the poles of $f(\sigma)$ are located at the set of points $\{b_\nu\}_{\nu=1}^{\infty}$, and that no $b_\nu = 0$. In this case $u(x, y)$ is regular at $z = 0$, and its domain of association is the z plane less the segments $E\{z \mid x = \operatorname{Re} b_\nu, y^2 \geq |\operatorname{Im} b_\nu|; \nu = 1, 2, \cdots\}$. After Bergman,[15] we then consider the z plane minus strips of thickness $2\,|b_\nu|^{-h}$ covering the above segments; that is, we are interested in the domain

$$\begin{aligned}\mathfrak{B} \equiv E\Bigg\{ z \;\Bigg|\; \sum_{\nu=1}^{\infty} [\operatorname{Re} b_\nu + |b_\nu|^{-h} \leq x \\ \leq \operatorname{Re} b_{\nu+1} - |b_{\nu+1}|^{-h}, y^2 < \infty] \Bigg\} \\ UE\Bigg\{ z \;\Bigg|\; \sum_{\nu=1}^{\infty} [\operatorname{Re} b_\nu - (b_\nu)^{-h} \leq x \leq \operatorname{Re} b_\nu \\ + |b_\nu|^{-h}, y^2 < [1b_\nu|^{-2h} - (x - b_\nu)^2|] \Bigg\}. \quad (4.7)\end{aligned}$$

It is clear that if $z \in \mathfrak{B}$ then $\sigma \in \mathfrak{R}$; consequently, we may prove the result.

Theorem 4.1. *Let $u(x, y)$ be a* GASPT *function with a meromorphic associate $f(\sigma)$. Furthermore, let $f(\sigma)$ be representable as a quotient of entire functions of order λ, $h(\sigma)/g(\sigma)$, such that $h(\sigma)$ is not of maximum type. Then, in the subdomain $\mathfrak{B}$ of the domain of association for $u(x, y)$, and for $|z| = r$ sufficiently large one, has the inequality*

$$|u(x, y)| \leq \frac{\pi^{\frac{1}{2}}}{2^{K-1}} \frac{e^{a' r^{\lambda+\epsilon'}}}{\Gamma[\frac{1}{2}(K + 1)]\Gamma(\frac{1}{2}K)}, \quad (4.8)$$

where $a' = a + 1 + \epsilon$, and ϵ, $\epsilon' > 0$ are arbitrarily small.

From the integral representation for $u(x, y)$ we have for r sufficiently large

$$\begin{aligned}|u(x, y)| &\leq 2^{K-1} |\alpha_K| \int_0^{\pi} |f(x + iy\cos\alpha)| \sin^{K-1}\alpha \, d\alpha \\ &\leq |\alpha_K| \frac{\pi\Gamma(K)}{\Gamma[\frac{1}{2}(K + 1)]^2} \frac{e^{(a+\epsilon)r^\lambda}}{e^{-r^{\lambda+\epsilon'}}} \\ &= \frac{\pi^{\frac{1}{2}}}{2^{K-1}} \frac{e^{r^{\lambda+\epsilon'}[(a+\epsilon)r^{-\epsilon'}+1]}}{\Gamma[\frac{1}{2}(K + 1)]\Gamma(\frac{1}{2}K)} \\ &\leq \frac{\pi^{\frac{1}{2}}}{2^{K-1}} \frac{e^{a' r^{\lambda+\epsilon'}}}{\Gamma[\frac{1}{2}(K + 1)]\Gamma(\frac{1}{2}K)},\end{aligned}$$

where $a < \infty$ since $h(\sigma)$ is not of maximum type, and ϵ, $\epsilon' > 0$ are arbitrarily small.

As Bergman[15] has shown in the case of harmonic functions in three variables, one may obtain an interesting class of meromorphic GASPT functions by considering

[38] See Bergman[15] and Gilbert[7] for a discussion of this proof.

$$f(\sigma) = \frac{\Gamma'(\sigma+\alpha)}{\Gamma(\sigma+\alpha)} = -\gamma + \sum_{n=0}^{\infty}\left(\frac{1}{n+1} - \frac{1}{n+\sigma+\alpha}\right), \quad \alpha \neq 0, \tag{4.9}$$

where γ is Euler's constant and $\Gamma(\sigma)$ is the gamma function. From the relations

$$\frac{\Gamma'(\sigma)}{\Gamma(\sigma)} = \log\sigma - \frac{1}{2\sigma} - \frac{1}{2}\sum_{\mu=3}^{\infty}(\mu-3)\mu^{-1} \times \sum_{\nu=1}^{\infty}(\sigma-\nu)^{-\mu}, \quad \operatorname{Re}\sigma > 0, \tag{4.10}$$

and

$$\frac{\Gamma'(m\sigma)}{\Gamma(m\sigma)} = \frac{1}{m}\sum_{\nu=0}^{m-1}\frac{\Gamma'(\sigma+\nu/m)}{\Gamma(\sigma+\nu/m)} + \log m, \quad m = 2, 3, 4\cdots, \tag{4.11}$$

we may obtain the following transplanted relations concerning the functions $\phi_K(z;\ \alpha) \equiv \mathfrak{A}_K[\Gamma'(\sigma-\alpha)/\Gamma(\sigma-\alpha)]$:

$$\phi_K(z;\alpha) = \mathfrak{L}_K(z;\alpha) - \tfrac{1}{2}\mathfrak{F}_K^{(1)}(z;\alpha) - \frac{1}{2}\sum_{\mu=3}^{\infty}(\mu-2)\mu^{-1}\sum_{\nu=1}^{\infty}\mathfrak{F}_K^{(\mu)}(z+\nu;\alpha),$$

and

$$\phi_K(mz;m\alpha) = \frac{1}{m}\sum_{\nu=0}^{m-1}\phi_K\left(z+\frac{\nu}{m};\alpha\right) + \frac{i\pi^{\frac{1}{2}}2^{1-K}}{\Gamma[\frac{1}{2}(K+1)]\Gamma(\frac{1}{2}K)}\log m,$$

where

$$\mathfrak{L}_K(z;\alpha) = \alpha_K\int_{\mathfrak{L}}\log(\sigma+\alpha)(\zeta-\zeta^{-1})^{K-1}\frac{d\zeta}{\zeta},$$

$$\mathfrak{F}_K^{(\mu)}(z+\nu;\alpha) \equiv \alpha_K\int_{\mathfrak{L}}(\sigma+\nu+\alpha)^{-\mu}(\zeta-\zeta^{-1})^{K-1}\frac{d\zeta}{\zeta};$$

$$\mathfrak{F}_K^{(\mu)}(z+\nu;\alpha) = \sum_{n=0}^{\infty}\binom{\mu+n-1}{\mu-1}(-1)^n(\nu+\alpha)^{-n-\mu} \times \frac{n!}{\Gamma(K+n)}r^nC_n^{\frac{1}{2}K}(\xi)$$

for $|z| < |\alpha|$, and

$$\mathfrak{F}_K^{(\mu)}(z+\nu;\alpha) = -\frac{\pi^{\frac{1}{2}}2^{\frac{1}{2}-\frac{1}{2}K}}{\Gamma(\frac{1}{2}K)}(1-\xi^2)^{\frac{1}{4}(1-K)} \cdot \sum_{n=0}^{\infty}\binom{n+\mu-1}{n-1}(\nu+\alpha)^nr^{-n-\mu}P_{n+\mu+\frac{1}{2}-\frac{1}{2}K}^{(\frac{1}{2}-\frac{1}{2}K)}(\xi),$$

for $|z| > |\alpha|$, where the $P_\beta^\alpha(\xi)$ are associated Legendre functions.

V. GASPT FUNCTIONS WITH ENTIRE ASSOCIATES

We shall call a GASPT function *entire* if it is without singularities in the finite z plane. Because of Theorems (3.6) and (3.8), we realize that a GASPT function is entire if and only if its $\mathfrak{A}_K$ associate is entire; consequently, we have the following result.

Theorem 5.1. *A* GASPT *function, regular about the origin, and defined by a series development*

$$u(x,y) = \sum_{n=0}^{\infty}a_n(x^2+y^2)^{\frac{1}{2}n}C_n^{\frac{1}{2}K}\left[\frac{x}{(x^2+y^2)^{\frac{1}{2}}}\right], \tag{5.1}$$

is an entire GASPT *function if and only if* $\lim_{n\to\infty}|a_n|^{1/n} = 0$.

It is possible also to characterize entire GASPT functions by their order and type in a similar manner as is done for entire analytic functions. With this in mind we shall establish the following:

Theorem 5.2. *Let* $u(x, y)$ *be a* GASPT *function with a series development* (5.1), *and let*

$$\overline{\lim_{n\to\infty}}\, n^{1/\lambda}|a_n|^{\frac{1}{n}} = (\lambda e a)^{1/\lambda}, \tag{5.2}$$

then

$$e^{r^{\lambda-\epsilon}} < |u(x,y)| < e^{r^{\lambda+\epsilon}}, \tag{5.3}$$

for $|z| = r$ *sufficiently large and all* $\epsilon > 0$. *Furthermore, we may distinguish three cases,* $a = 0$, $0 < a < \infty$, $a = \infty$ *as* $u(x, y)$ *being of minimum, normal, and maximum type, respectively.*

This therorem is essentially a transplant of the one concerning analytic functions.[36] First, we note that the $\mathfrak{A}_K$ associate of $u(x, y)$ is

$$f(\sigma) = \sum_{n=0}^{\infty}a_n\frac{\Gamma(K+n)}{n!}\sigma^n,$$

and then once we have determined the order and type of $f(\sigma)$ we bound $|u(x, y)|$ above and below by using $\mathfrak{A}_K[f]$ and its inverse $\mathfrak{A}_K^{-1}[u]$. The inverse operator $\mathfrak{A}_K^{-1}[u]$ may be rewritten as[5]

$$f(\sigma) = \mathfrak{A}_K^{-1}[u] = \left(\frac{i}{2}\right)^K\frac{\beta_K}{i}(r^2-\sigma^2) \times \int_{+r}^{-r}\frac{V(z,\bar z)(z-\bar z)^K}{[(z-\sigma)(\bar z-\sigma)]^{\frac{1}{2}K+1}}\frac{dz}{z}, \tag{5.4}$$

where

$$V(z,\bar z) \equiv u\left(\frac{z+\bar z}{2}, \frac{z-\bar z}{2i}\right), \qquad r^2 = z\bar z.$$

From (5.4) we have for $|u(x, y)| \leq M(R)$ (for $|z| \leq R$) that

$$|f(\sigma)| \leq 2\pi\, |\beta_K|\, M(R)R^K[(R^2 + \rho^2)/(R - \rho)^{K+2}] \quad (\text{for } |\sigma| \leq \rho < R), \tag{5.5}$$

and for $\rho \leq \frac{1}{2}R$, say, we have

$$|f(\sigma)| \leq 5K\Gamma(\tfrac{1}{2}K)^2 2^{2K-1} M(R). \tag{5.6}$$

Expression (5.6) tells us if $u(x, y)$ is of finite order and normal type, then $|f(\sigma)|$ also is bounded above in this manner. If the order and type of $f(\sigma)$ are λ and a ($a \neq 0, \infty$), respectively, we must have[36]

$$\overline{\lim_{n\to\infty}} \left(n^{1/\lambda} \cdot \left| an \frac{\Gamma(K+n)}{n!} \right|^{\frac{1}{n}} \right) = (\lambda e a)^{1/\lambda}; \tag{5.7}$$

but since

$$\overline{\lim_{n\to\infty}} \left(n^{1/\lambda} \left| an \frac{\Gamma(K+n)}{n!} \right|^{\frac{1}{n}} \right) = \overline{\lim_{n\to\infty}}\, n^{1/\lambda}\, |a_n|^{\frac{1}{n}}, \tag{5.8}$$

we have that both $|u(x, y)|$ and $|f(\sigma)|$ are bounded above for R sufficiently large by $e^{aR^{\lambda-\epsilon}}$ for $|z| \leq R$, $|\sigma| \leq R$, respectively, and any $\epsilon > 0$. Because of (5.6), however, unless $|u(x, y)|$ also is bounded below by $e^{aR^{\lambda-\epsilon}}$, $|f(\sigma)|$ cannot be of order λ and type a. Our theorem follows from this.

Corollary 5.3. *Let $u(x, y)$ be an entire* GASPT *function of finite order, then $\mathfrak{M}(r) = \max_{|z|\leq r} |u(x, y)|$ is bounded by an inequality of the form*

$$\exp R^{\lambda-\epsilon} < \mathfrak{M}(r) < \exp R^{\lambda+\epsilon}. \tag{5.9}$$

Theorem 5.4. *Let $u(x, y)$ be regular in the disk $|z| \leq R$, $u(0, 0) = 0$, and let $\mathfrak{M}(r)$ be as above. Then, for all $|r| \leq \frac{1}{2}R$, $\mathfrak{M}(r)$ is bounded by*[39]

$$\mathfrak{M}(r) \leq 5\pi^{\frac{1}{2}}[\Gamma(\tfrac{1}{2}(K+2))/\Gamma(\tfrac{1}{2}(K+1))]r. \tag{5.10}$$

VI. A RESIDUE CALCULUS FOR GASPT-FUNCTION ELEMENTS

The GASPT equation may be rewritten in the form

$$\frac{\partial}{\partial x}\left(y^K \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y}\left(y^K \frac{\partial u}{\partial y} \right) = 0, \tag{6.1}$$

and in this form it implies the existence of a stream function $v(x, y)$, which satisfies the Stokes–Beltrami equations[2,6]

$$y^K \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad y^K \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{6.2}$$

It is possible to introduce an integral operator $\mathfrak{A}_K^*[f]$ which generates the stream function $v(x, y)$. For instance, if $u(x, y)$ is a GASPT function, then $f(\sigma) = \mathfrak{A}_K^{-1}[u]$, and

$$v(x, y) = \mathfrak{A}_K^*[f] \equiv \tfrac{1}{2}iy^K \mathfrak{A}_K[(\zeta + \zeta^{-1})f(\sigma)]. \tag{6.3}$$

It is convenient in what follows that the functions $u(x, y)$, $v(x, y)$ be real. This may be done by choosing $f(\sigma)$ so that it is real on the real axis; in the remainder of this section we shall assume that $u(x, y)$, $v(x, y)$ are real.

We now introduce a complex combination of the real potential and stream functions,

$$w(x, y) = u(x, y) + iy^{-K}v(x, y) \equiv \mathfrak{A}_K[f(\sigma)\{1 + \tfrac{1}{2}(\zeta + \zeta^{-1})\}] \equiv A_K[f(\sigma)] = -\tfrac{1}{2}\mathfrak{A}_K[f(\sigma)(\zeta - 1)^2\zeta^{-1}], \tag{6.4}$$

consider its domain of association, and the various function elements related to this domain of association. To simplify our discussion we consider the case where $f(\sigma) = h(\sigma)/(\sigma - \alpha)$, and $h(\sigma)$ is entire. Then, in order to understand the connection between the different function elements we must consider the singularity manifold for $f(\sigma)(\zeta - \zeta^{-1})^{K+1}$, which we represent as

$$\mathfrak{z}^2 \equiv E(\zeta_\nu = (i/y)\{(x - \alpha) + (-1)^\nu \times [(x - \alpha)^2 + y^2]^{\frac{1}{2}}\}; \nu = 1, 2). \tag{6.5}$$

Now, unless $(x - \alpha)^2 + y^2 \equiv (z - \alpha)(\bar{z} - \alpha) = 0$, the singularity manifold has two branches, which is particularly interesting in terms of the Theorems (3.6) and (3.8).

Let us assume that the representation (6.4) is defined in the neighborhood of an initial point z^0. It is possible for us to extend this initial domain of definition by continuing $w(x, y)$ along a contour γ, starting at z^0, providing that no point of γ corresponds to a singularity of the integrand on the path of integration. We recall, however, that σ depends on both z and ζ, and hence as we continue $w(x, y)$ along γ in the z plane, the singularities of the integrand move in the ζ plane. For instance the points $\zeta = \zeta_\nu(z)$, ($\nu = 1, 2$) [Eq. (6.5)] move and may cross over the path of integration $\mathfrak{L}$. If this should happen, the integral will have a jump in value equal to a branch of $w(x, y)$. If $\zeta_\nu \in \mathfrak{L}$ ($\nu = 1. 2$), then the corresponding "singular" points in the z plane are given by

[39] For a more detailed discussion of this result see R. Gilbert, *Some inequalities for generalized axially symmetric potentials with entire and meromorphic associates*, Technical Note BN-315, March 1963, to appear in Duke Math. J.

$$x = \operatorname{Re}[\alpha] - \operatorname{Im}[\alpha][(|\zeta|^2 - 1)/(|\zeta|^2 + 1)] \tan \psi, \quad (6.6)$$

$$y = 2 \operatorname{Im}[\alpha][|\zeta|/(|\zeta|^2 + 1)] \sec \psi, \quad \psi = \arg[\zeta].$$

Equations (6.6) may be interpreted as a one-to-two mapping of the ζ plane onto the z plane, $z = T\zeta$. For instance, the arcs $\mathfrak{L}_k \equiv E\{\zeta \mid |\zeta| = 1, (k-1)\frac{1}{2}\pi \leq \arg \zeta \leq k\frac{1}{2}\pi\}$ $(k = 1, 2, 3, 4)$ map onto the half lines (when $\operatorname{Im}[\alpha] > 0$),

$$\Lambda_j = E\{z \mid x = \operatorname{Re}[\alpha]; (y/\operatorname{Im}[\alpha])^{(-1)^j} \leq (-1)^{j+1}\} \quad (j = 1, 2), \quad (6.7)$$

in the following manner:

$$\mathfrak{L}_1 \to \Lambda_1, \quad \mathfrak{L}_2 \to \Lambda_2, \quad \mathfrak{L}_3 \to \Lambda_2, \quad \mathfrak{L}_4 \to \Lambda_1;$$

the point at infinity of the z plane serves as a "double" branch point.

By a simple computation it can be shown that when z crosses Λ_1 from the right, $\zeta_1(z)$ leaves the unit disk by crossing $\mathfrak{L}_4$, whereas $\zeta_2(z)$ enters over $\mathfrak{L}_1$. On the other hand, as z crosses Λ_2 from the right, $\zeta_1(z)$ leaves the unit disk by passing over $\mathfrak{L}_2$, and $\zeta_2(z)$ enters over $\mathfrak{L}_3$. We are now able to consider the continuation of $w(x, y)$ along γ, which originates at a point $z^0 \notin \Lambda_1 \cup \Lambda_2$. If $\gamma \cap \{\Lambda_1 \cup \Lambda_2\} = 0$, then $w(x, y)$ may be continued to any point z which is the terminal point of γ. On the other hand if $\gamma \cap \{\Lambda_1 \cup \Lambda_2\} = z^1$, then $w(x, y)$ has a jump in value at z^1, as we cross over Λ_1 or Λ_2, which is equal to a branch of $w(x, y)$. Because of this property associated with the lines Λ_1, Λ_2, we refer to them as the *lines of separation* for $w(x, y)$.

To clarify this point we consider as an illustration the associate $f(\sigma) = h(\sigma)/(\sigma - \alpha)$, where $h(\sigma)$ is entire and nonzero at $\sigma = \alpha$. Then

$$w(x, y) = \alpha_K \int_{\mathfrak{L}+1}^{-1} \frac{h(\sigma)}{\sigma - \alpha} (\zeta - \zeta^{-1})^{K+1} \, d\zeta = \frac{2\alpha_K}{iy} \int_{\mathfrak{L}} \frac{h(\sigma)(\zeta - \zeta^{-1})^{K+1}}{[\zeta - \zeta_1(z)][\zeta - \zeta_2(z)]} \zeta \, d\zeta, \quad (6.8)$$

where $\zeta(z)$ is given by (6.5). When z crosses Λ_1 from the left, $w(x, y)$ goes through a jump in value equal to

$$[-4\pi/\Gamma(\tfrac{1}{2}K)^2]h(\alpha)([x - \alpha]^2 + y^2)^{\frac{1}{2}K - 1} \times (-1/2y)^K(y + i[x - \alpha]). \quad (6.9)$$

We are now in a position to consider integrals of the following type:

$$\int_{\mathcal{J}_s} (u \, dx - y^{-K} v \, dy) = \operatorname{Re}\left\{\int_{\mathcal{J}_s} w(x, y) \, dz\right\}, \quad (6.10)$$

where $\mathcal{J}_s \cap \{y \leq 0\} = 0$, and $\mathcal{J}_s$ is a smooth Jordan curve. It is possible to rewrite the integral (6.10) by making use of the identity,

$$u \, dx - y^{-K} u \, dy = \alpha_K \int_{+1}^{-1} f(\sigma)\left[dx + \frac{i}{2} dy(\zeta + \zeta^{-1})\right](\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta} = \alpha_K \int_{+1}^{-1} [f(\sigma) \, d\sigma](\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}, \quad (6.11)$$

where z is contained in a sufficiently small neighborhood of z^0, $\mathfrak{N}(z^0)$, such that $\mathfrak{N}(z^0) \cap [\Lambda_1 \cup \Lambda_2] = 0$. Then, if $\mathcal{J}_s$ does not intersect $\Lambda_1 \cup \Lambda_2$, one has clearly—since the integrand is absolutely integrable—that

$$\operatorname{Re}\left\{\int_{\mathcal{J}_s} w(x, y) \, dz\right\} = \alpha_K \int_{\mathcal{J}_s} \left[\int_{+1}^{-1} f(\sigma)\left[dx + \frac{i}{2} dy(\zeta + \zeta^{-1})\right] \times (\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}\right. = \alpha_K \int_{+1}^{-1} \frac{d\zeta}{\zeta} (\zeta - \zeta^{-1})^{K-1} \int_{C_s} f(\sigma) \, d\sigma, \quad (6.12)$$

where C_s is the image of $\mathcal{J}_s$ under the mapping $z \to \sigma$ (ζ fixed).

Theorem 6.1. *Let $w_0(x, y)$ be the* GASPT-*function element defined in a neighborhood of z_0 by means of the operator $A_K[f]$, where $f(\sigma) = h(\sigma)/(\sigma - \alpha)$, $h(\sigma)$ is entire, and $h(\alpha) \neq 0$. Furthermore, let $\mathcal{J}_s$ be a smooth Jordan curve, which originates at a point $z_0 \notin \Lambda_K$ $(K = 1, 2)$, and which intersects Λ_1 at just one point z_1. Then*

$$\operatorname{Re}\left\{\int_{\mathcal{J}_s} w(x, y) \, dz\right\} \equiv \int_{\mathcal{J}_s} (u \, dx - y^{-K} v \, dy) = \operatorname{Re}\left\{\int_{\mathcal{J}_s} w_0(x, y) \, dz\right\} + (-\tfrac{1}{2})^{K+1} \frac{4\pi h(\alpha)}{K\Gamma(\frac{1}{2}K)^2} \times \int_{\xi_1}^{\xi_s} \xi^{-1} \, d[(1 + \xi^2)^{\frac{1}{2}}]^K = 2\pi i \alpha_K h(\alpha) \times \int_{M_t} (\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta},$$

where $\xi = (z - \alpha)/y$, and

$$M_t = E\{\zeta \mid |\zeta| = 1; 0 \leq \arg[\zeta] \times \cos^{-1}[\operatorname{Im}(\alpha)/y]\}. \quad (6.13)$$

To establish this result we note that the integral

of the jump in value of $w(x, y)$ when z crosses Λ_1 from the left may be written as

$$\begin{aligned}&\frac{-4\pi}{\Gamma(\frac{1}{2}K)^2}\,h(\alpha)\int_{s_1}^{s_0}([x-\alpha]^2+y^2)^{\frac{1}{2}K-1}\\&\quad\times\left(\frac{-1}{2y}\right)^K([x-\alpha]\,dy-y\,dx)\\&=+\frac{\pi}{K}\frac{h(\alpha)}{\Gamma(\frac{1}{2}K)^2}\int_{s_1}^{s_0}\left(\frac{y}{x-\alpha}\right)d\left\{\left(1+\left[\frac{x-\alpha}{y}\right]^2\right)^{\frac{1}{2}K}\right\}\\&=\frac{\pi}{K}\frac{h(\alpha)}{\Gamma(\frac{1}{2}K)^2}\int_{\xi_1}^{\xi_0}\xi^{-1}\,d[(1+\xi^2)^{\frac{1}{2}K}].\end{aligned}\tag{6.14}$$

When K is an integer, the jump integrals (6.14) may be evaluated as[6,9]

$$\begin{aligned}(K\text{ odd})&=\pi\frac{h(\alpha)}{\Gamma(\frac{1}{2}K)^2}\left\{(\eta^2-1)^{\frac{1}{2}}\sum_{\nu=0}^{\frac{1}{2}(K-3)}(-1)^\nu\right.\\&\quad\times\frac{(K-2;-2;\nu)}{(K-1;-2;\nu+1)}\,\eta^{K-2-2\nu}+(-1)^{\frac{1}{2}(K-1)}\\&\quad\left.\times\log[\eta+(\eta^2-1)^{\frac{1}{2}}]\right\}_{\eta=(\xi_1^2+1)^{\frac{1}{2}}}^{\eta=(\xi_0^2+1)^{\frac{1}{2}}},\end{aligned}\tag{6.15}$$

and

$$\begin{aligned}(K\text{ even})&+\pi\frac{h(\alpha)}{\Gamma(\frac{1}{2}K)^2}\left\{(\eta^2-1)^{\frac{1}{2}}\sum_{\nu=0}^{\frac{1}{2}K-1}(-1)^\nu\right.\\&\quad\left.\times\frac{(K-2;-2;\nu)}{(K-1;-2;\nu+1)}\,\eta^{K-2-2\nu}\right\}_{\eta=(\xi_1^2+1)^{\frac{1}{2}}}^{\eta=(\xi_0^2+1)^{\frac{1}{2}}},\end{aligned}\tag{6.16}$$

where $(m;-d;\nu)=d^\nu\Gamma(m/d+1)/\Gamma(m/d-\nu+1)$.

Theorem 6.1 may be reinterpreted in terms of the composition on quasimultiplication[6,9] for GASPT functions, which we define as

$$\begin{aligned}u_3(x,y)&=u_1(x,y)*u_2(x,y)=\mathfrak{A}_K[f_1(\sigma)f_2(\sigma)],\\u_j(x,y)&=\mathfrak{A}_K[f_j(\sigma)],\qquad j=1,2,3.\end{aligned}\tag{6.17}$$

(We note that quasimultiplication is commutative, associative, and distributive over ordinary addition.)

Let

$$\begin{aligned}&N(\alpha,K)h(\alpha)\\&=\alpha_K\int_{+1}^{-1}(\zeta-\zeta^{-1})^{K-1}\frac{d\zeta}{\zeta}\int_{C_\sigma}\frac{h(\sigma)}{\sigma-\alpha}\,d\sigma,\end{aligned}\tag{6.18}$$

where

$$N(\alpha,K)=\int_{M_\Gamma}(\zeta-\zeta^{-1})^{K-1}\frac{d\zeta}{\zeta},$$

and $h(\sigma)$ is entire. Furthermore, let $\operatorname{Re}\{\bar{w}(x,y)dz\}\overset{\text{def}}{\equiv}\mathfrak{A}_K[f(\sigma)d\sigma]$, and

$$\begin{aligned}\frac{1}{\rho_K(z,\alpha)}&=\mathfrak{A}_K\left[\frac{1}{\sigma-\alpha}\right]\\&=\frac{\pi}{\Gamma(\frac{1}{2}K)^2}\left(\frac{-1}{2y}\right)^{K-1}\frac{1}{[(z-\alpha)(\bar{z}-\alpha)]^{1-\frac{1}{2}K}},\end{aligned}\tag{6.19}$$

then

$$\operatorname{Re}\{\bar{w}(x,y)\,dz\}*\left\{\frac{1}{\rho_K(z,\alpha)}\right\}=\mathfrak{A}_K\left[\frac{f(\sigma)\,d\sigma}{\sigma-\alpha}\right].\tag{6.20}$$

The introduction of the terms $\rho_K(z,\alpha)$ suggests an expression for GASPT functions, corresponding to $\mathfrak{A}_K$ associates having Laurent expansions about $\sigma=\alpha$. Suppose

$$\begin{aligned}g(\sigma)&=\frac{b_{m+1}}{(\sigma-\alpha)^{m+1}}+\frac{b_m}{(\sigma-\alpha)^m}+\cdots\\&\quad+\frac{b_1}{\sigma-\alpha}+\sum_{i=0}^{\infty}c_i(\sigma-\alpha)^i,\end{aligned}\tag{6.21}$$

and then for K odd we have

$$\begin{aligned}G(x,y)&=\mathfrak{A}_K[g(\sigma)]\\&=\sum_{i=1}^{m+1}b_i\mathfrak{A}_K[\sigma-\alpha)^{-i}]+\sum_{i=0}^{\infty}c_i\mathfrak{A}_K[(\sigma-\alpha)^i]\\&=\sum_{i=1}^{m+1}\frac{(-1)^ib_i}{i!\,\Gamma(\frac{1}{2}K)^2}\left(\frac{-1}{2y}\right)^{K-1}\frac{\partial^{i-1}}{\partial x^{i-1}}\{(z-\alpha)(\bar{z}-\alpha)\}^{\frac{1}{2}K-1}\\&\quad+\sum_{i=0}^{\infty}c_i[(z-\alpha)(\bar{z}-\alpha)]^{\frac{1}{2}i}C_i^{\frac{1}{2}K}\left\{\frac{x-\alpha}{[(z-\alpha)(\bar{z}-\alpha)]^{\frac{1}{2}}}\right\}.\end{aligned}\tag{6.22}$$

In general, however, we would write

$$\begin{aligned}G(x,y)&=\sum_{i=1}^{m+1}\left\{\frac{1}{\rho_K(z,\alpha)}*\right\}^i\\&\quad+\sum_{i=0}^{\infty}R(z,\alpha)^iC_i^{\frac{1}{2}K}\left[\frac{x-\alpha}{R(z,\alpha)}\right],\end{aligned}\tag{6.23}$$

where

$$\left\{\frac{1}{\rho_K}*\right\}^i=\frac{1}{\rho_K}*\frac{1}{\rho_K}*\cdots*\frac{1}{\rho_K}$$

(j terms), and $R(z,\alpha)^2=(z-\alpha)(\bar{z}-\alpha)$.

Theorem 6.1 suggests also the following analog of the argument principle in GASPT.

Theorem 6.2. *Let the A_K associate of $w(x, y)$ be $f'(\sigma)/f(\sigma)$, and suppose that $\mathcal{J}$ is a smooth, Jordan curve in the z plane, not passing through any zeros or poles of $f(z)$. Furthermore, let $f(z)$ have zeros at $\{a_\nu\}_{\nu=1}^n$ with multiplicities $\{r_\nu\}_{\nu=1}^m$ and poles at $\{b_\varrho\}_{\varrho=1}^n$ with multiplicities $\{s_\varrho\}_{\varrho=1}^n$, inside of $\mathcal{J}$. Then*

$$\frac{1}{4}\left(\frac{2}{iy}\right)^K\left\{\sum_{p=1}^{m} r_p \int_{\mathcal{J}} \{[x-a_p]\,dy - y\,dx\}([x-a_p]^2+y^2)^{\frac{1}{2}K-1} - \sum_{q=1}^{n} s_q \times \int_{\mathcal{J}} \{[x-b_q]\,dy - y\,dx\}([x-b_q]^2+y^2)^{\frac{1}{2}K-1}\right.$$

$$= \frac{1}{2\pi i}\int_{\substack{+1\\ |\zeta|=1}}^{-1} n\{f[(\sigma(\zeta)];0\}[\zeta-\zeta^{-1}]^{K-1}\frac{d\zeta}{\zeta}$$

$$= \sum_{p=1}^{m} r_p N(a_p;K) - \sum_{q=1}^{n} s_q N(b_q;K), \tag{6.24}$$

where $n\{\gamma;0\}$ is the winding number of γ with respect to 0, and $f[C_\sigma(\zeta)]$ is the image of C_σ by $f(\sigma)$ for fixed ζ. C_σ is the image of $\mathcal{J}$ under the mapping $z \to \sigma$ for fixed ζ.

Theorem 6.3. *Let the A_K associate of $w(x, y)$ be $g(\sigma)f'(\sigma)/f(\sigma)$, and suppose that $\mathcal{J}$ is a smooth Jordan curve in the z plane not passing through any of the zeros $\{a_p\}_{p=1}^m$ or poles $\{b_q\}_{q=1}^n$ of $f(\sigma)$. (The zeros and poles have the orders r_p, s_q, respectively.) Furthermore, let $g(z)$ be regular-analytic for z inside $\mathcal{J}$, then*

$$\int_{\mathcal{J}} \mathfrak{A}_K\left[\frac{f'}{f}\,g\,d\sigma\right] = \int_{\mathcal{J}} \mathfrak{A}_K\left[\frac{f'}{f}\right] * \mathfrak{A}_K[g\,d\sigma]$$

$$= \int_{\mathcal{J}} \left\{\sum_{p=1}^{m}\frac{r_p}{\rho_K(z;a_p)} - \sum_{q=1}^{n}\frac{s_q}{\rho_K(z;bq)}\right\} * \{\mathrm{Re}\,G(x,y)\,dz\}$$

$$= 2\pi i\alpha_K\left\{\sum_{p=1}^{m} r_p g(a_p)N(a_p;K) - \sum_{q=1}^{n} s_q g(b_q)N(b_q;K)\right\},$$

where

$$G(x,y) = -\tfrac{1}{2}\mathfrak{A}_K[g(\sigma)(\zeta-1)^2\zeta^{-1}]. \tag{6.25}$$

VII. POISSON'S EQUATION IN GASPT

In this section, which concludes our present discussion of function-theoretic methods in GASPT, we consider the nonhomogeneous equation

$$L^{(K)}[u(x,y)] = \rho(x,y). \tag{7.1}$$

For $K = 0$, this equation may be solved formally by replacing x, y by $z = x + iy$, $z^* = x - iy$ and integrating. That is,

$$\Delta^{(2)}u = 4\frac{\partial^2 u}{\partial z\,\partial z^*} = \rho\left(\frac{z+z^*}{2}, \frac{z-z^*}{2i}\right), \tag{7.2}$$

which may be integrated directly into

$$u = \int^{z^*}\int^{z^*} \rho\left(\frac{z+z^*}{2}, \frac{z-z^*}{2i}\right) dz\,dz^* + g(z) + h(z^*). \tag{7.3}$$

This method suggests that we consider the integral operator $\mathfrak{A}_K[f]$:

$$u(x,y) = \mathfrak{A}_K[f] = \alpha_K \int_{\substack{+1\\ |\zeta|=1}}^{-1} f(\sigma,\sigma^*)(\zeta-\zeta^{-1})^{K-1}\frac{d\zeta}{\zeta}, \tag{7.4}$$

where $f(\sigma, \sigma^*)$ is a holomorphic function of the two complex variables σ, σ^* in the polydisk $\{|\sigma| \le R\} \times \{|\sigma^*| \le R'\}$. σ is the same as before, and $\sigma^* = x - (iy/2)(\zeta + \zeta^{-1})$, z^0 is an initial point of definition for $u(x, y)$, and $|z - z^0| < \epsilon$ where $\epsilon > 0$ is sufficiently small, etc. Furthermore, it is evident that for $|\zeta| = 1$, and x, y real $\sigma^* = \bar{\sigma}$; if $\partial f/\partial\sigma = 0$, or $\partial f/\partial\sigma^* = 0$, then f is just a function of either σ^* or σ, respectively, and $\mathfrak{A}_K \to \mathfrak{A}_K$.

In an earlier work[8] we established the following theorem concerning the generality of solutions generated by $\mathfrak{A}_K[f]$.

Theorem 7.1. *In a sufficiently small neighborhood $\mathfrak{N}(z^0)$ of an initial point z^0, the representation*

$$u(x,y) = \mathfrak{A}_K[f] = \alpha_K \int_{\substack{+1\\ |\zeta|=1}}^{-1} f(\sigma,\sigma^*)(\zeta-\zeta^{-1})^{K-1}\frac{d\zeta}{\zeta}$$

yields the most general, analytic (in the real variables x, y) solution to $L^{(K)}[u] = \rho$, providing that f is contained in the class of analytic functions of two variables $\mathfrak{I}$ which satisfy the integral equation

$$\rho(x,y) = 4\int_{\substack{+1\\ |\zeta|=1}}^{-1} \frac{\partial^2 f}{\partial\sigma\,\partial\sigma^*}(\zeta-\zeta^{-1})^{K-1}\frac{d\zeta}{\zeta}. \tag{7.5}$$

In order to understand how this operator transforms analytic functions of two variables into solutions of (7.1) we consider the integrals

$$u_{nm}(x,y) = \mathfrak{A}_K[\sigma^n\sigma^{*m}] = \alpha_K\int_{\substack{+1\\ |\zeta|=1}}^{-1}\sigma^n\sigma^{*m}(\zeta-\zeta^{-1})^{K-1}\frac{d\zeta}{\zeta}, \tag{7.6}$$

which we evaluate to be

$$u_{nm}(x,y) = \alpha_K\frac{\pi i^K}{K}\,n!\,m!\,r^{n+m} \times \sum_{\nu=-n}^{+n}\sum_{\mu=-m}^{+m}\frac{(-1)^\nu}{(n+\nu)!\,(m+\mu)!} \times \frac{P_n^\nu(\xi)P_m^\mu(\xi)}{B(\frac{1}{2}(K+1+\mu+\nu),\frac{1}{2}(K+1-\mu-\nu))}, \tag{7.7}$$

$B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$ by using the integral relations for the associated Legendre functions

$$\frac{n!\,r^nP_n^m(\xi)}{^m(n+m)!} = \frac{1}{2\pi i}\int_{|\zeta|=1}\sigma^n\zeta^m\frac{d\zeta}{\zeta}, \tag{7.8}$$

and

$$\frac{i^m n!}{(n-m)!} r^n P_n^m(\xi) = \frac{1}{2\pi i} \int_{|\zeta|=1} \sigma^{*n} \zeta^m \frac{d\zeta}{\zeta}.$$

Consequently, the function $f(\sigma, \sigma^*)$ holomorphic about the origin,

$$f(\sigma, \sigma^*) = \sum_{n,m=0}^{\infty} a_{nm} \sigma^n \sigma^{*m}, \tag{7.9}$$

are mapped by $\mathfrak{A}_K[f]$ onto the solutions

$$u(x, y) = \sum_{n,m=0}^{\infty} a_{nm} u_{nm}(x, y) \tag{7.10}$$

of the nonhomogeneous GASPT equation that are regular in a neighborhood of the origin.

In order to generate a particular solution of $L^K[u] = \rho$, where

$$\rho(x, y) = \sum_{n,m=0}^{\infty} \rho_{nm} x^n y^m, \qquad |z| < R, \tag{7.11}$$

we consider the integral equation

$$\int_{\substack{+1 \\ |\zeta|=1}}^{-1} g(\sigma, \sigma^*)(\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta} = \rho(x, y), \tag{7.12}$$

where $g(\sigma, \sigma^*) = 4\alpha_K(\partial^2 f/\partial\sigma\partial\sigma^*)$. We may solve formally for the Taylor coefficients g_{MN} of $g(\sigma, \sigma^*)$ as follows:

$$\rho_{nm} = \frac{\partial^{n+m}\rho(0, 0)}{\partial x^n \, \partial y^m}$$

$$= \int_{\substack{+1 \\ |\zeta|=1}}^{-1} \frac{\partial^{n+m} g(\sigma, \sigma^*)}{\partial x^n \, \partial y^m}\bigg|_{s=0} (\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}$$

$$= \left(\frac{i}{2}\right)^m \int_{\substack{+1 \\ |\zeta|=1}}^{-1} \sum_{\mu=0}^{m} \binom{m}{\mu}\binom{\mu}{\nu}(-1)^\mu \frac{\partial^{n+m} g(0, 0)}{\partial\sigma^{\nu+\mu} \, \partial\sigma^{*(m+n)-(\nu+\mu)}}$$

$$\times (\zeta + \zeta^{-1})^m (\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}$$

$$= \left(\frac{i}{2}\right)^m \sum_{\mu=0}^{m} \sum_{\nu=0}^{n} \binom{m}{\mu}\binom{n}{\nu}(-1)^\mu g_{\nu+\mu,(m+n)-(\nu+\mu)}$$

$$\times \int_{\substack{+1 \\ |\zeta|=1}}^{-1} (\zeta + \zeta^{-1})^m (\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}, \tag{7.13}$$

$$a_{nm} = \beta_{mK} \times \sum_{\mu=0}^{m} \sum_{\nu=0}^{n} (-1)^\mu \binom{m}{\mu}\binom{n}{\nu} g_{\nu+\mu,(m+n)-(\nu+\mu)}, \tag{7.14}$$

where

$$\beta_{mK} = \left(\frac{-1}{2}\right)^m K\pi i^K \times \sum_{\mu=0}^{m} \frac{(-1)^\mu \binom{m}{\mu}}{B(\frac{1}{2}(K+1-n)-\nu, \frac{1}{2}(K+1+n)-\nu)}. \tag{7.15}$$

There are two special cases for which we may obtain $g(\sigma, \sigma^*)$ immediately, when ρ is a function of either x or y alone. When $\rho(x, y) = p(x)$, we may choose $g(\sigma, \sigma^*)$ to have the form $G(\sigma + \sigma^*) = G(2x)$. Then

$$g(\sigma, \sigma^*) = G(2x) = \frac{p(x)}{\beta_{0K}}, \tag{7.16}$$

where

$$\beta_{0K} = \frac{i\pi}{KB(\frac{1}{2}(K+1), \frac{1}{2}(K+1))},$$

and

$$f(\sigma, \sigma^*) = \frac{p(x)}{4\alpha_K \beta_{0K}} \sigma\sigma^*. \tag{7.17}$$

A particular solution to (7.1) is then given by

$$u(x, y) = \frac{p(x)}{4\beta_{0K}} \int_{\substack{+1 \\ |\zeta|=1}}^{-1} \sigma\sigma^*(\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}$$

$$= \frac{\pi}{K} r^2 i^K p(x) \sum_{\nu=-1}^{+1} \sum_{\mu=-1}^{+1} \frac{(-1)^\mu P_1^\nu(\xi) P_1^\mu(\xi)}{(\nu+1)!\,(\mu+1)!\,B(\frac{1}{2}(K+1+\mu+\nu), \frac{1}{2}(K+1-\mu-\nu))}. \tag{7.18}$$

When $\rho(x, y) = p(y)$, a particular solution may be obtained by representing $g(\sigma, \sigma^*)$ by $G(\sigma - \sigma^*) = G(iy[\zeta + \zeta^{-1}])$, where $G(\sigma - \sigma^*)$ must satisfy the integral equation

$$\int_{\substack{+1 \\ |\zeta|=1}}^{-1} G(iy[\zeta + \zeta^{-1}])(\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta} = p(y). \tag{7.19}$$

Now, if $G(z) = \sum_{\nu=0}^{\infty} g_\nu z^\nu$, and $p(y) = \sum_{\nu=0}^{\infty} p_\nu y^\nu$, then we may compute the coefficients g_ν formally as follows:

$$\int_{\substack{+1 \\ |\zeta|=1}}^{-1} \sum_{n=0}^{\infty} g_n \left\{\sum_{\nu=0}^{n} (iy)^n \binom{n}{\nu} \zeta^{2\nu-n}\right\} (\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}$$

$$= \sum_{n=0}^{\infty} (iy)^n g_n \left\{\sum_{\nu=0}^{n} \binom{n}{\nu} \int_{\substack{+1 \\ |\zeta|=1}}^{-1} \zeta^{2\nu-n}(\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}\right.$$

$$= \pi K i^K \sum_{n=0}^{\infty} g_n (-y)^n \left\{\sum_{\nu=0}^{\infty} \binom{n}{\nu}(-1)^\nu \right.$$

$$\left. \left. B\left(\frac{K+1-n}{2} + \nu, \frac{K+1+n}{2} - \nu\right)^{-1}\right\}, \tag{7.20}$$

from which we have

$$g_n = \frac{(-1)^n p_n}{\pi K i^K \gamma_{Kn}}, \quad \text{where} \quad \gamma_{Kn} \equiv \sum_{\nu=0}^{n} \binom{n}{\nu} (-1)^\nu B\left(\frac{K+1-n}{2} + \nu, \frac{K+1+n}{2} - \nu\right)^{-1}, \tag{7.21}$$

providing $\gamma_{Kn} \neq 0$ for all n, and $K < 0$. In this case we may proceed as before and obtain a particular solution

$$u(x, y) = \frac{1}{4} \int_{\substack{+1 \\ |\zeta|=1}}^{-1} \sigma\sigma^* G(iy[\zeta + \zeta^{-1}]) \times (\zeta - \zeta^{-1})^{K-1} \frac{d\zeta}{\zeta}. \tag{7.22}$$

In closing we should like to mention several other papers, which are related to the present subject:

R. Gilbert, Proc. Am. Math. Soc. **13**, 229 (1962);
R. Gilbert, Contrib. Differential Eqs. **1**, 441 (1963);
J. B. Diaz, and G. S. S. Ludford, Canadian J. Math. **8**, 82 (1956);
R. Gilbert, Pacific J. Math. **13**, 79 (1963);
R. Gilbert and H. C. Howard, Technical Note BN-350, Univ. Maryland (1964);
R. Gilbert and H. C. Howard, Technical Note BN-352, Univ. Maryland (1964);
E. Kreyzig, Archiv. Math. **14**, 193 (1963);
J. Mitchell, Math. Z. **82**, 314 (1963);
A. White, Ann. Polonici Math. **10**, 81 (1961).

Fluctuations in Multiple Capture Processes

ALBERT G. PETSCHEK
University of California, Los Alamos Scientific Laboratory, Los Alamos, New Mexico
(Received 17 December 1963)

The fluctuation in the number of many-fold captures by an element exposed to a high neutron flux is computed. It is found that the fluctuations are identical to those that would be obtained by selecting nuclei at random from an infinite supply with the appropriate average composition.

IT is known that, in a nuclear weapon, uranium can be subjected to fluxes of neutrons sufficient to allow it to capture neutrons repeatedly.[1] Inasmuch as the transuranium elements resulting from this process are easily detectable and it may in future tests be possible to mine substantially all the debris from the explosion, it is of interest to compute the statistical fluctuations in the number of plural captures. One might suppose that an anomalously high number of nth captures would lead to an anomalously high number of $(n + 1)$th captures, etc., and that the fluctuations would therefore build up; but he would be wrong. It is shown below that the distribution of n-fold captures among N nuclei exposed to a flux is in fact just that which would result from choosing N nuclei at random from a supply having the appropriate average distribution.

Consider first an infinite box containing balls of colors labeled $0, 1, \cdots, n, \cdots m$ in the proportions p_n. Then it is well known for $m = 1$ and true in general that the probability of having exactly x_0 balls of color 0, x_1 of color 1, etc., among N balls chosen at random is the term containing $p_0^{x_0} p_1^{x_1} \cdots$ in the binomial expansion of $(\sum p_n)^N$. It is convenient to introduce a generating function g by means of

$$g(y_0, y_1, \cdots, y_m) = (\sum p_n y_n)^N \qquad (1)$$

in terms of which it is easy to evaluate the moments of the distribution. For example the mean number of n-colored beads is $\langle x_n \rangle = \partial g/\partial y_n$ taken at constant y_k, $k \neq n$ and evaluated at $y_l = 1$, $l = 0, 1, \cdots, m$; the standard deviation in x_n can be found from $\langle x_n(x_n - 1)\rangle = \partial^2 g/\partial y_n^2$, and the correlation between x_n and x_k from $\langle x_n x_k \rangle = \partial^2 g/\partial y_n\, \partial y_k$.

It is now shown that the generating function for the distribution of multiple captures is identical in form to g. Let $P(x_0, x_1, \cdots x_m; z)$ be the probability that there are exactly x_n n-fold captures after an exposure z. Because it takes nothing from the understanding of the problem and simplifies the expressions for the average number of captures immensely,[2] the fission cross sections are taken to be zero and the capture cross sections are assumed equal except for the mth. In that case it is convenient to define $z = \int \eta\sigma v\, dt$ where η is the neutron density, σ the capture cross section, v the velocity, and t the time. To make the problem finite, take the mth capture cross section to be zero or, alternatively, refuse to notice the difference between an m-fold capture and any higher capture. In that case

$$\frac{\partial}{\partial z} P(x_0 x_1 \cdots x_m; z)$$
$$= -(N - x_m)P(x_0 x_1 \cdots x_m; z) + \sum_n (x_n + 1)$$
$$\times P(x_0 x_1 \cdots, (x_n + 1), (x_{n+1} - 1), \cdots x_m; z), \qquad (2)$$

where the first term represents the decrease in P because of a capture which changes x_n to $x_n - 1$ and x_{n+1} to $x_{n+1} + 1$ while the sum represents the increase due to a capture which changes the nth argument of P from $x_n + 1$ to x_n and the $(n + 1)$th from $x_{n+1} - 1$ to x_{n+1}. Now let

$$G(y_0, y_1, \cdots y_m; z)$$
$$= \sum_{x_0, x_1, \cdots} y_0^{x_0} y_1^{x_1} \cdots y_m^{x_m} P(x_0 x_1 \cdots x_m; z). \qquad (3)$$

Then $\partial G/\partial y_n = x_n G/y_n$ and therefore the following equation for G is obtained by multiplying (2) by the product of the $y_k^{x_k}$, summing, promoting the index x_n, and demoting x_{n+1} in the rightmost term:

$$\frac{\partial}{\partial z} G = -NG + y_m \frac{\partial}{\partial y_m} G + \sum y_{n+1} \frac{\partial}{\partial y_n} G. \qquad (4)$$

The characteristics of this differential equation are given by

$$dy_m/y_m = -dz; \quad dy_n = -y_{n+1}\, dz, \quad n \neq m, \qquad (5)$$

[1] A. Ghiorso, *et al.*, Phys. Rev. 99, 1048 (1955); P. R. Fields, *et al.*, Phys. Rev. 102, 180 (1956).

[2] J. J. Devaney, A. G. Petschek, and M. T. Menzel, Los Alamos Rept. LAMS 2226 (not otherwise published).

whose solution is

$$y_m = A_m e^{-z}, \quad y_{m-1} = A_m e^{-z} + A_{m-1}, \cdots, \tag{6}$$
$$y_0 = A_m e^{-z} \pm A_{m-1}[z^{m-1}/(m-1)!] \mp \cdots + \tfrac{1}{2}A_2 z^2 - A_1 z + A_0,$$

where the A_n are constants of integration. Along the characteristics G satisfies

$$dG/dz = -NG \tag{7}$$

and if the initial condition on P is taken to be $P(N, 0 \cdots 0; 0) = 1$ then

$$G(y_0, y_1, \cdots y_m; 0) = y_0^N \tag{8}$$

where the y_0 on the right must of course be found by following the appropriate characteristic to $z = 0$ which makes it equal to $A_m + A_0$. The solution of (7) with (8) is

$$G = (A_m + A_0)^N e^{-Nz}. \tag{9}$$

Equation (6) allows one to replace A_0 and A_m in (9) by the y's, giving

$$G = (\sum p_n y_n)^N \tag{10}$$

with

$$p_n = \frac{z^n}{n!} e^{-z} \qquad n < m, \qquad p_m = \sum_m^\infty \frac{z^k}{k!} e^{-z}. \tag{11}$$

G is therefore exactly of the form (1). The p_n given by (11) are the probabilities of n-fold capture and can also be found from simpler calculations.[2]

The contention made in the first paragraph, that the correlations among the x_n are no different from those resulting from the random selection of nuclei from a reservoir, is therefore established.

ACKNOWLEDGMENTS

I am indebted to Dr. G. I. Bell, who has applied similar techniques to more complicated problems[3] for interesting discussion.

This work was done under the auspices of the U. S. Atomic Energy Commission.

[3] G. I. Bell, Ann. Phys. **21**, 243 (1963).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 5, NUMBER 7 JULY 1964

Errata: Excitation Spectrum of an Impurity in the Hard-Sphere Bose Gas

Toshio Soda
The Enrico Fermi Institute for Nuclear Studies, The University of Chicago, Chicago, Illinois
(Received 18 February 1964)
[J. Math. Phys. 5, 142 (1964)]

Equation (1) on p. 142 should read

$$H_A = \sum_p p^2 a_p^+ a_p + \tfrac{1}{2} \sum_{p,p',q} V_q a_{p+q}^+ a_{p'}^+ a_{p'+q}' a_p. \quad (1)$$

Equation (4) on p. 143 should read

$$\alpha_k = u_k a_k + v_k a_{-k}^+. \quad (4)$$

Equation (7) on p. 143 should read

$$H_B = \sum_p p^2 a_p^+ a_p + \tfrac{1}{2} \sum_{p,p',q} V_q a_{p+q}^+ a_{p'}^+ a_{p'+q} a_p + \sum_p p^2 b_p^+ b_p + \sum_{p,p',q} V_q a_{p+q}^+ b_{p'}^+ b_{p'+q} a_p. \quad (7)$$

Equation (24) on p. 145 should read

$$E_k = k^2 + E_k^{(2)} - E_g^{(2)} = k^2 - \frac{64}{45} \frac{(\rho a^3)^{\frac{1}{2}}}{\pi^{\frac{1}{2}}} k^2 + O[(\rho a^3) k^2, k^4]. \quad (24)$$