General theory of the metal-insulator transition

K. Moulopoulos and N. W. Ashcroft
Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853

(Received 15 October 1991)

We present a general theory of the metal-insulator transition based on the formal introduction of an auxiliary irrotational gauge field and the subsequent behavior of the system in response to it. Correspondingly general scaling relations are derived which determine the point of transition at a critical value of the controlling density in terms of a series expansion with respect to a coupling constant. Functional integral techniques are also used to show the relation of this transition to a density-fluctuation instability. Finally, an interpretation is given of the transition in terms either of condensation of gauge bosons or, alternatively, in terms of the appearance of an additional geometric phase, an interpretation that is valid for quite general condensed-matter systems.

I. INTRODUCTION

The detailed microscopic understanding of the metal-insulator transition[1] remains one of the most important problems in theoretical condensed-matter physics, and in spite of several decades of effort[2,3] a satisfactory many-body theory of this phenomenon is still lacking. The simplest *realistic* system to consider is a canonical, neutral, many-body system of electrons and nuclei with mutual Coulomb interactions. We will limit discussion to the ground state of such a system, and imagine variation in a single quantity, the thermodynamic density, as the sole parameter controlling the onset of the transition. An immediate simplification is the consideration of only a single charged component but placed in a uniform *rigid* background of opposite sign. In this paper, we propose a general theory of the transition for the one-component problem (the extension to a two-component system is based on the general treatment of Appendix A and will be given later[4]), and we will also present calculations that can in principle determine the density of its occurrence.

To characterize the phase at fixed value of the controlling parameter, whether metallic or insulating, we need to examine how the system responds to an applied electric field **E**, i.e., whether it is conducting (the corresponding response current density $\mathbf{J}\neq 0$) or not. Experimentally **E** can be established through the application of a time-dependent vector potential $\mathbf{A}(t)$ and/or a space-dependent scalar potential $A_0(\mathbf{r})$ to the system. It is convenient to begin the discussion by first setting $A_0=0$. Moreover, we can imagine the remaining vector potential $\mathbf{A}(t)$ to be turned on slowly, from $t=-\infty$. For a nonrelativistic system a linear-response determination of both the real and imaginary parts of the conductivity in the static limit then shows[5,6] that the parameter t, although physically important for the *establishment* of **A**, is actually no longer *formally* important. We may therefore draw conclusions about the response of the system merely from a consideration of the dependence of its energy on the magnitude of **A**, at least in the limit $A\rightarrow 0$. Accordingly, to begin the analysis we take the convenient limit of static **A**, which clearly corresponds to a vanishing **E**. This case also allows us to treat the system as being in equilibrium, with no dissipation permitted.

The basic formal element is therefore the introduction of an auxiliary vector potential initially considered external to the system (later we take it as part of the system). Its form also will be chosen so that it induces neither a magnetic nor an electric field; its presence then serves only as a "detector" of the transition. Experimentally, the establishment of this vector potential always proceeds in a finite time through some time-dependent magnetic flux outside the system, and this will generally lead to a finite current **J** through electromagnetic induction. The phase of our system, whether conducting or nonconducting, is characterized through the value of this current. The *metallic state* will be defined as a current-carrying state ($\mathbf{J}\neq 0$); otherwise, the complete lack of net induction (for which $\mathbf{J}=0$) will define an insulating state.

The arguments now to be presented are quite general, and proceed in stages. In the first stage, we introduce just a static vector potential **A**, but so chosen that $\nabla\times\mathbf{A}=0$. Following Kohn,[5] we stipulate that the system possesses a macroscopic but finite ring topology, and **A** is then to be associated with a magnetic flux permeating the region physically *inaccessible* to the particles (the hole of the ring). Alternatively, we could imagine a *simply connected* but initially finite space with a constant **A**, a situation that corresponds to the magnetic flux being at infinity. Here it must be understood that the possible establishment of a current must occur *first* (in the system with the ring topology), and only *later* is the thermodynamic limit to be taken ($N\rightarrow\infty$; $V\rightarrow\infty$; $N/V\rightarrow n$); that is, a part of the ring is subsequently allowed to increase so that its volume fills the whole of space. A criterion[5] for the establishment of the insulating phase is then the independence of the energy of the system on the value of A, at least in the limit $A\rightarrow 0$. A nonvanishing dependence on A would signal the transition to a metallic phase. Kohn[5] showed this for a metallic state in terms of a mutual incompatibility that exists between a gauge transformation that could remove the A dependence (it actually *does* in the case of an insulator) and the single-valuedness condition of the many-body wave function when we circumnavigate the ring (essentially the Aharonov-Bohm effect[7]). We will prove here a similar but more general criterion for the simply connected space, and by treating **A** as the *total* (external plus induced) vector potential. In the case of the metal we will show that the A dependence of the ground-state energy induces a current which, through its interaction with the vector potential, has *exactly* the value required to produce this A dependence of the energy. We also derive general scaling relations for the charged system in the presence of a vector potential and use them to determine the point of transition through a series expansion in the coupling constant e^2. At the lowest nontrivial order, a random-phase approximation approach yields the metal-insulator transition at $r_{s,0}=61.7$ for the electron-gas problem.

In the second stage of the argument, we return to the more general case in which the gauge field has a fourth component, namely the scalar potential A_0; once again, a spatially independent A_0 will first be used as a limiting case of zero electric field. A functional integral treatment for fermions and the associated introduction of a plasmon field ϕ then leads to a criterion of the metal-insulator transition that can be written entirely in closed form, as a summation over ϕ. We also provide an alternative but in fact simpler criterion by taking advantage of the similarity in appearance of ϕ and A_0 in the action functional and through the use of charge neutrality. In both criteria, our controlling parameter, the density, enters only implicitly. The latter criterion, however, provides a *direct* connection between the metal-insulator transition and the instability (in a mean-field sense) of the plasmon field.

In the third and final stage of the argument, we view the four-component gauge field as a part of the system itself, with no external contributions, and we quantize it. The resulting theory shows that the transition to the metallic phase already identified in stages 1 and 2 can actually be reinterpreted as condensation of Goldstone, and what are referred to as "ghost" bosons in certain regions in real-space time. We show how this result is connected with an alternative interpretation of the metallic side of the transition, in terms of the appearance of a geometric phase. This phase appears as a generalization of Berry's phase[8] and it is related to the adiabatic parallel transport of the center of mass of the physically accessible region of the system around the underlying ring. The appearance of this phase should actually be expected from Kohn's argument, in view of the known relation[8] between Berry's phase and the Aharonov-Bohm effect.

As can be seen from the above, we actually achieve a mapping between the macroscopic *ring problem* and the charged system in a *simply connected* space, as a result of which the holonomies of the first are associated with the metallic state in the second. And, as will be seen below, we are led as a consequence to new ideas and interpretations of the metal-insulator transition in addition to some new formal procedures for the location of the density of its occurrence in charged systems.

II. GLOBAL GAUGE SYMMETRY

Our criterion for the establishment of the metallic state, namely the A dependence of its energy, is a manifestation of global gauge symmetry breaking through the underlying Aharonov-Bohm effect[7] for the delocalized state. The very existence of this criterion can itself be traced to an insufficiency of the conventional form of gauge invariance[9] and it arises because of an additional "surface term" associated with the ring topology; it is crucially related to correct order of limits. This term is omitted in conventional derivations[10] but, as will be seen below, it is important, and fundamentally connected to the *experimental* definition of insulators and metals.

In more detail, introduction of a general external gauge field ($\mathbf{A}^{\text{in}}$, A_0^{in}) induces a current that can be written in closed form[10] in terms of current-current and current-charge-density-response functions χ_{JJ} and $\chi_{J\rho}$. A gauge transformation $\mathbf{A}^{\text{in}} \rightarrow \mathbf{A}^{\text{in}} + \nabla\Lambda$ and $A_0^{\text{in}} \rightarrow A_0^{\text{in}} - (1/c)(\partial\Lambda/\partial t)$ then changes the induced current by

$$-\omega_p^2 \frac{1}{c}(\nabla\Lambda)_i + \int dt' d\mathbf{r}' \chi_{J_iJ_j}(\mathbf{r},\mathbf{r}';t-t')\frac{1}{c}\nabla'_j\Lambda(\mathbf{r}'t') + \int dt' d\mathbf{r}' \chi_{J_{i\rho}}(\mathbf{r}\mathbf{r}';t-t')\frac{1}{c}\frac{\partial\Lambda(\mathbf{r}'t')}{\partial t'} \;.$$

Integration by parts with respect to time and space and use of Ward identities and causality cancels out all of these terms *except* for a surface term, namely

$$\int_{S'} d\mathbf{r}' \int dt' \chi_{J_iJ_j}(\mathbf{r}\mathbf{r}';t-t')\frac{1}{c}\Lambda(\mathbf{r}'t') \;.$$

This term is frequently discarded by making the assumption that Λ vanishes on the boundaries. However, for the ring topology, and especially for an experimentally motivated process in which the thermodynamic limit is taken by first making the inner radius very large (so that the system becomes a cylinder) and *then* taking the outer radius to infinity, the surface term is generally nonvanishing. This is especially so for those parts of S' that can be identified as what were formerly the cross sections of the ring. After proceeding to the thermodynamic limit, these parts are sections through which charge can "leak" to or from a compensating region at infinity along the conducting direction [see Fig. 1(b)].

The above transformation and argument are very general. In our case of zero magnetic field (and irrotational vector potential) we can take $\mathbf{A}^{\text{in}} = A_0^{\text{in}} = 0$, $\mathbf{A} = \nabla\Lambda$, and $A_0 = -(1/c)(\partial\Lambda/\partial t)$, i.e., quite generally we can view the imposition of the gauge field as a pure gauge transformation. Kohn[5] takes $\Lambda \propto \mathbf{A}\cdot\sum_i \mathbf{r}_i$ and we see immediately that if we have a conducting state, the above surface term is related to the physical transport of the center of mass of the simply connected component around the ring (or along the conducting direction), as will also be proposed in Sec. V.

If this essential finiteness of the (experimentally required) ring topology is *not* taken into account *before* taking the thermodynamic limit, then, in addition to the obvious experimental problem of how such a current could ever be established, a formal theoretical problem also arises: for $A_0 = 0$ the taking of the thermodynamic limit first would give[9]

$$J_i(\mathbf{r},\omega)=\int dr'[\chi_{J_iJ_j}(\mathbf{rr}';\omega)-\omega_p^2\delta(\mathbf{r}-\mathbf{r}')\delta_{ij}]\frac{A_j(\mathbf{r}',\omega)}{c}$$

with $\lim_{k\to 0}\chi_{JJ}(\mathbf{k},0)=\omega_p^2+0(k^2)$ for an isotropic time-reversal-symmetric medium. It always gives a vanishing current and a conducting state would not then even be *formally* possible. It indicates the importance of taking due theoretical account of the ring topology *before* the thermodynamic limit is subsequently taken, and links obviously to the experimental requirements for establishing currents.

III. SCALING RELATIONS

Imagine a single-component system of particles with mutual Coulomb interactions and immersed in a uniform inert background of opposite charge (in simply connected space). Also introduce a constant and static vector potential **A** but, as indicated above, so chosen that its presence produces no physical effects. It enters into the Hamiltonian through a "minimal substitution," i.e., $\mathbf{p}\to\mathbf{p}-(e/c)\mathbf{A}$ only. Here **A** is understood to be the *total* (external plus induced) vector potential. By taking derivatives with respect to various parameters of the Hamiltonian (a method that we have applied in the past[11] to a different problem and described for *both* a two- and a single-component system in Appendix A) we obtain the following expression which we emphasize is *exact* for the ground state and also independent of state symmetry:

$$\frac{d}{dr_s}\left[\frac{E}{N}\right]=-\frac{1}{r_s}\left[\frac{\langle T\rangle}{N}+\frac{E}{N}\right]+\frac{1}{r_s}\frac{V}{N}\frac{1}{c}\mathbf{A}\cdot\mathbf{J}\ , \quad (1)$$

where $\mathbf{J}=-(e/V)\langle\sum_i\{[\mathbf{p}_i-(e/c)A]/m\}\rangle$ for a single component. This expression shows that the case of an insulator ($\mathbf{J}=0$) is equivalent to the absence of **A** and in this case we recover both the known scaling relations and the virial theorem.[11] The charged system in the insulating phase therefore does *not* feel the presence of **A**, at least so far as the energy is concerned. This result is independent of the structural properties of the phase and is valid for arbitrary values of **A**. It is in agreement with Kohn's results,[5] although it is perhaps more general, since it is *not* a result of perturbation theory in A. We also see that in the case of a metal the **A** dependence of the energy arises exactly from the interaction of the established current with **A**. This result generalizes previous work based on noninteracting particles.[12]

Expression (1) leads to the following scaling relations:

$$\frac{E}{N}=\frac{f(mr_s;Ar_s)}{r_s}=\frac{g(e^2r_s;A^2r_s)}{r_s^2}\ , \quad (2)$$

which are derived in Appendix A together with their two-component analogs. These relations can be used to locate the point of the metal-insulator transition (for $A\to 0$) from the high-density behavior $r_s\to 0$ (for any A) but under conditions where, for example, we take $m\to\infty$ in such a way that mr_s remains constant. At which side of the transition we find ourselves depends on the value of this constant. Use of (2) also gives the following series expansion:

$$\frac{E}{N}=\frac{E(A=0)}{N}+\left\{\alpha+\beta\left[\frac{r_0}{a_0}\right]+\gamma\left[\frac{r_0}{a_0}\right]^2+0\left[\left[\frac{r_0}{a_0}\right]^3\right]\right\}\frac{e^2A^2}{mc^2}+\cdots\ , \quad (3)$$

where $a_0=\hbar^2/me^2$ and $2r_0$ is the mean distance between two electrons. To lowest order a random-phase approximation (RPA) calculation of the correlation energy gives $\alpha=\frac{1}{2}$ and $\beta=-\frac{1}{2}(4/9\pi)^{1/3}[(1-\ln 2)/\pi^2]r_s$. This is shown in Appendix B by going through the calculation of the RPA correlation energy but now by translating all wave vectors by $-e\,\mathbf{A}/c\hbar$. If we retain only these first two terms, the coefficient of A^2 vanishes at $r_{s,0}=61.7$. According to our criterion, this is the next correction to the "critical" value of $r_s=\infty$ of the noninteracting case (a metal at all densities, as expected), and it should be reasonably close to the metal-insulator transition for the case of very small coupling (i.e., for $e\simeq 0$). However, higher corrections are crucial in locating the point with precision, as can be seen from the fact that the A dependence does not identically vanish at lower densities in this approximation. We note at this point that the corresponding relativistic problem for a ring of finite circumference shows a metal-insulator transition *even* at the noninteracting level.[4] But in this case, it is a consequence of the equivalence of space and time coordinates that the spatial ring topology must be accompanied by a corresponding *time* ring topology and this *necessitates* a time-dependent gauge field or a nonzero electric field. In this case a "ring" corresponds to an increase of the magnetic flux at infinity by a flux quantum, and in this case dissipation must be treated as part of the problem. This can be done through an appropriate manifestation of the arrow of time, which can be viewed as the time-reversal symmetry breaking analogous to the spatial breaking of the right-left symmetry in the case of a current.

In the other important limit of very low densities, we expect a broken symmetry spatial state, namely a crystal. This can be expected either from the theory of freezing of a classical plasma in combination with the correspondence principle, or directly from the quantum theory of the Wigner crystal. (Alternatively a more formal justification is discussed in Sec. IV.) The first approximation to such a state is a harmonic solid with an energy

$$\frac{E(m,A)}{N}=\frac{\alpha(A)}{r_s}+\frac{\beta(A)}{r_s^{3/2}}\left[\frac{m_e}{m}\right]^{1/2}\text{Ry}\ .$$

For the Wigner crystal at zero gauge field the values of the constants are $\alpha(A=0)\simeq -1.79$ and $\beta(A=0)\simeq 2.65$. It is straightforward to show that this state must also be insulating and in fact it can be directly proved by the use of the "harmonic-oscillator virial theorem," which states

$$T(m)=U(m)-U(\infty)=\tfrac{1}{2}[E(m)-E(\infty)]\ , \quad (4)$$

where T and U are the ground-state kinetic and potential energy, respectively. Combination of (4) with (1) shows that the term proportional to $\mathbf{A}\cdot\mathbf{J}$ is identically zero, demonstrating therefore that by the same criterion as used above this state must be insulating. [We do not discuss the possibility of a *collective* current of the charge-density-wave (CDW) type (e.g., a sliding crystal), since this would require a finite electric-field threshold and considerations of pinning mechanisms not present in our fundamental Hamiltonian.] The presumption is that the metal-insulator transition accompanies a structural transition to an ordered state. Though the expectation is a transition to a crystalline state, there is no proof so far that a nondiffusive disordered state (such as a Coulomb glass) can be ruled out.

IV. FUNCTIONAL INTEGRATION

As indicated above, the next step of the argument is to introduce a static scalar potential A_0 and again take the limit of zero spatial variation. This new field component also leads to no physical consequences; rather, it is again a formal device whose purpose is to "detect" neutrality (or charge conservation), as we will see below. From Sec. II it can be shown that the case of fixed gauge choice $A_0=0$ that was used above is entirely equivalent to the more general case $A_0\neq 0$, and the criterion for the metal-insulator transition therefore still applies (this is true even with the inclusion of the surface term). Next we introduce a functional integral method,[13] which we recently used[14] to describe two-component pairing; the essential difference here is the addition of a four-component constant gauge field as part of the action. Again, it enters by the "minimal substitution," namely $\mathbf{p}\rightarrow\mathbf{p}-(e/c)\mathbf{A}$ and $i\hbar(\partial/\partial t)\rightarrow i\hbar(\partial/\partial t)-eA_0$. If we first start from the high-density region, we may apply a Hubbard-Stratonovich transformation[15] of the density type; this yields a description in terms of a plasmon field ϕ. From the above discussion the criterion for the metal-insulator transition will then be

$$\frac{\delta J_j}{\delta A_i(1)}=0=\frac{\delta}{\delta A_i(1)}\int d_2\frac{\delta S}{\delta A_j(2)}=\int d_2\frac{\delta^2 S}{\delta A_i(1)\delta A_j(2)}\,,\qquad i,j\neq 0\,.$$

An expansion around $\mathbf{A}\simeq 0$ gives an action functional of the form

$$S[\phi,\mathbf{A}]=S[\phi,\mathbf{A}=0]+\int dt\int d\mathbf{r}\frac{m}{2}\rho[\mathbf{r},t]\left[\frac{e\mathbf{A}}{mc}\right]^2,\tag{5}$$

which possesses a form familiar from recent discussions of mesoscopic systems[16,17] (where finite temperature generalizations are given), but which is appearing now with the "superfluid density" $\rho[\mathbf{r},t]$ being proportional to a renormalized density of conduction electrons in the metallic state. In fact, it is easy to show that in the case of noninteracting fermions of density n we would immediately have $\rho[\mathbf{r},t]=n$. The static and long-wavelength limit ρ of the Fourier transform of $\rho[\mathbf{r},t]$ can be given entirely in closed form[16] in terms of the propagator G of the electrons in the absence of $\mathbf{A}$, and its vanishing can therefore also serve as a criterion of the onset of the metal-insulator transition. However, the crucial difference in the physics of our problem from these approaches is also manifested formally: we do *not* minimize S with respect to A; rather, we exploit the properties of $J_i=\delta S/\delta A_i$, and any such minimization would only be formally equivalent to the case $J=0$ (which in the treatments of mesoscopic systems is associated with the Meissner effect,[18] and in fact leads to flux quantization).

There is, however, an alternative way of deriving $\delta^2\rho/\delta A_0^2$ (rather than ρ itself), and this proceeds by making use of neutrality. Let Q be the total electronic charge; then clearly for our *canonical* system

$$\frac{\delta Q}{\delta A_0(1)}=0=\frac{\delta}{\delta A_0(1)}\int d_2\frac{\delta S}{\delta A_0(2)}=\int d_2\frac{\delta^2 S}{\delta A_0(1)\delta A_0(2)}\,.\tag{6}$$

But because of the manner[13] in which ϕ enters in S we can show that

$$\frac{\delta^2 S}{\delta A_0(1)\delta A_0(2)}=\frac{\delta^2 S}{\delta\phi(1)\delta\phi(2)}-v_c^{-1}(12)\,.$$

Since $\delta^2 S/\delta\phi(1)\delta\phi(2)$ is just the inverse plasmon propagator,[13] we immediately expect a connection to the *instability* of the plasmon phase described by the divergence of the plasmon susceptibility at $q\rightarrow 0$. Indeed, application of (6) on the action (5) yields the final result

$$\left.\frac{\partial^2\rho}{\partial A_0^2}\right|_{A_0,\mathbf{A}\rightarrow 0}\propto\sum_{\mathbf{p},\omega}G(\mathbf{p},\omega;\mathbf{A}=A_0=0)\times G(\mathbf{p},\omega;\mathbf{A}=\mathbf{A}_0=0)\,,\tag{7}$$

where G is the propagator of electrons moving in the full plasmon field $\phi(\mathbf{p},\omega)$. Its form can be concisely written as $G^{-1}=G_0^{-1}-e\phi$, in terms of the Green's function G_0 for noninteracting electrons. Expression (7) can now serve as a closed form of an alternative criterion for the location of the metal-insulator transition: When the left-hand side is nonzero we should have a metallic phase, and when it vanishes we should have an insulating phase; its dependence on the controlling parameter (the density) is entirely implicit.

Equation (7) shows that the metal-insulator transition occurs at the density where the static and long-wavelength limit of the proper polarizability $\Pi^*(\mathbf{q},\omega)$ [i.e., the right-hand side of (7)] vanishes. By the compressibility sum rule,[19] it is also the point where the compressibility vanishes. This is in agreement with, and in fact generalizes the recent result[20] for 1D electron liquids, namely that the current-carrying effective mass [which is inversely proportional to the coefficient of A^2 in (3)] is inversely proportional to the compressibility. The above result is also consistent with very general dielectric and conductivity arguments; we know, for example, that $\epsilon(\mathbf{q}=0,\omega)=1+(4\pi i/\omega)\sigma(\mathbf{q}=0,\omega)$, and also $\epsilon(\mathbf{q},\omega)=1-(4\pi e^2/q^2)\Pi^*(\mathbf{q},\omega)$, so either the divergence of $\epsilon(0,0)$ [through a nonzero $\Pi^*(0,0)$ in the metallic phase] or the finiteness of $\sigma(0,0)$ in the static limit, directly relate a dielectric catastrophe of the low-density phase to the metal-insulator transition and to the nonvanishing of (7).

We have also shown[4] that such a dielectric catastrophe criterion for a simple low-density phase of hydrogen atoms (the Mott problem, which, however, is quite different from the problem of the metal-insulator transition in the Wigner crystal) yields the metal-insulator transition at a value $r_s \sim 1.7$, which is close to Mott's value for a Thomas-Fermi potential (which is 1.72 according to accurate numerical work[21]). Finally, we will report[4] that through the use of Ward identities a relation similar to (7) can be established for the general problem with time- and space-dependent gauge fields. The vanishing of (7) then actually coincides with the vanishing of ρ itself, thereby placing the connection between the metal-insulator transition and the dielectric catastrophe on firmer grounds. It appears that this transition can also be connected to the possibility of *translational* symmetry breaking along the conducting path. In the case of such a symmetry breaking it is indeed possible to show[4] that the projection of Π^* along the conducting direction vanishes, in agreement with the above discussion.

The above arguments are easily generalized to two components (for example, the ionic crystal case); specifically, it can be shown[4] that ionic fluctuation is crucial in precisely fixing the point where the transition occurs. This fluctuation appears in the definition of Π^*, which now includes both components. Finally, a finite-temperature generalization is readily available in the functional integral language, so that the metal-insulator transition can actually be studied by functional many-body techniques for nonzero temperatures as well.

V. GAUGE BOSONS; BERRY'S PHASE

Lastly, we comment that a treatment of the four-component gauge potential as a dynamical variable on a quite equal footing with the charged particles in simply connected space, combined with the subsequent quantization of the *total* system through the canonical method, appears to offer another interpretation of the metal-insulator transition. In order to quantize canonically, a field conjugate to A_0 must be introduced from the beginning in the Lagrangian density.[22] Its quantization then produces Goldstone and "ghost" bosons, but they never actually become observable under ordinary circumstances. In fact, these bosons merely control the "escaping" of charge from the physical region[23] to what, in the Gupta-Bleuler theory,[24] are referred to as "unobservable parts" of the Hilbert space. These parts are actually the compensating parts of the rest of the ring [Fig. 1(b)], which are formally at infinity after the thermodynamic limit is taken. But when the Goldstone and "ghost" bosons "condense" in the sense of the standard Bose transformation method,[25] a condensation which corresponds to a displacement of the gauge boson fields by c number functions $f(x)$ and of the classical gauge fields by $\alpha_\mu(x)$, then we have two distinct cases to consider. First, if $f(x)$ should be completely regular, then the condensation is equivalent to a pure gauge transformation, and in such a case any change of the gauge potential can be simply "gauged away" without the introduction of any physical effect. We suggest that this is exactly the case of the insulator as envisioned by Kohn,[5] an interpretation that is strengthened by the fact that in this situation the current *is* zero because of an exact cancellation,[22] namely $(e/c)\alpha_\mu(x)-\hbar\partial_\mu f(x)=0$. If, however, $f(x)$ should possess singularities, and here we might imagine that it is *not* single valued around the underlying ring, these cancellations are then incomplete. It is this situation that leads to a finite current, and it is also for this case that the condensation of the gauge bosons is known to produce *macroscopic phenomena.* Such phenomena have been discussed in the past,[23] and always appear to be associated with the "escaping of charge" either to infinity or to boundaries. What we propose here is that the establishment of a current, when we pass to the metallic phase, is also a macroscopic phenomenon (related to the "escaping" of charge associated specifically with the underlying ring topology) and is controlled by the condensation of the same gauge bosons. This now establishes a mapping between the simply connected space and the experimentally motivated ring topology. Correspondingly, the gauge bosons in the simply connected space are mapped onto the Goldstone bosons associated with the symmetry breaking accompanying the imposition of ring periodicity on the system [see Fig. 1(a)]. Then the condensation of the gauge bosons is mapped onto the condensation of the Goldstone bosons. The physical picture above therefore views the conducting state as "distortion" of the electron system, in exactly the same way that the distortion of a lattice can be described by condensation of acoustic phonons[25] with a singular displacement $f(x)$. [A regular $f(x)$ would just produce sound waves.] The singularity of $f(x)$ is closely related to the self-consistent appearance, in the conducting state, of the magnetic flux through the physically inaccessible region. The current is then, for the corresponding simply connected problem, precisely the macroscopic object necessary to reconcile the gauge symmetry breaking with the gauge invariance of the Heisenberg equations for the fundamental electron fields.

Finally, it can be shown[23] that the additional $f(x)$ associated with the Bose condensation induces a corresponding additional phase on the electron field operators. Again, for $f(x)$ regular this phase can be gauged away in the case of the insulator, but for $f(x)$ singular it is nontrivial in the case of the metal. This nontrivial phase is the analog of the Burgers vector[25] for the "distorted" electron system, but with the important difference that in the case of the metal-insulator transition the topological constant is not quantized; actually, it can be shown[25] to be equal to the underlying Aharonov-Bohm phase, a result that ties in nicely with Kohn's original arguments. An additional element, however, is that the phase possesses a quite simple geometric origin having to do with the slow parallel transport of the center of mass of the system around the underlying ring. This can therefore be seen as a generalization of the well-known Berry's phase,[8] and has been discussed recently[26] in the context of transport of a single particle along a ring; we see here its natural generalization to the many-body system and to the metal-insulator transition (see also Sec. II).

There are other recent studies that lead to results entirely consistent with our propositions in this section. An example is the gauge symmetry breaking due to nonintegrable phases by a mechanism due to Hosotani,[27] but the detailed connection requires additional investigation. We are also examining further the connection of the above propositions to recent theories that relate a current-type Berry's phase with time-reversal symmetry breaking[28] and with motion of phase singularities.[29] The first establishes a source for dissipation, and the second views the passage to the insulating state as the pinning of singularities on the physical system (rather than on the physically inaccessible region). Because at the points of phase singularities the amplitude of the wave function must vanish, this viewpoint is completely consistent with Kohn's notion of disconnectedness of the many-body wave function[5] in the insulating state.

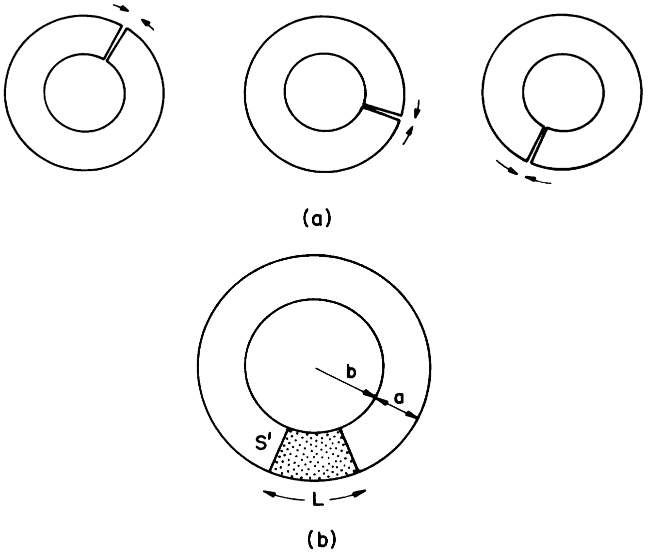

FIG. 1. (a) Equivalent impositions of the ring periodicity; all give the same physical system, i.e., one living on a torus but irrespective of where the "cut" is placed, and in consequence there are always gapless (Goldstone) bosons associated with this degeneracy. This will be true for *any* problem with ring topology. The singular condensation of these bosons produces a distortion and hence a displacement (generally time dependent) around the ring. (b) After the taking of the thermodynamic limit ($b \rightarrow \infty$ and *then* $a \rightarrow \infty$ and $L \rightarrow \infty$) and after excerpting a simply connected (shaded) region, these Goldstone bosons map onto "gauge bosons." The corresponding singular condensation of gauge bosons produces the current in the simply connected space. The mapping between the condensation of bosons in a ring and in a simply connected space leads to an alternative interpretation of conduction in the simply connected space in terms of Berry's phase, the latter being associated with the physical transport of the center of mass of the shaded region around the corresponding ring.

VI. CONCLUSION

By way of conclusion, a major result is that we have established a mapping between a charged system in simply connected space, and the same system derived from a macroscopic ring topology by the proper taking of a thermodynamic limit. The transition to the metallic state can be viewed as the breaking of global gauge symmetry through the appearance of a nonintegrable phase. In this way, a connection between the metallic state and the holonomies of the system is established.

Viewing the ring periodicity as an ordered state (in the same sense that the periodicity in a crystal is an ordered state) the subsequent breaking of this order, i.e., a distortion, proceeds through the condensation of the Goldstone bosons associated with the symmetry breaking that actually *generated* the order. In the case of the crystal, singular condensation of phonons produces dislocations; in the case of the simply connected region (that was mapped to the ring topology), singular condensation of gauge bosons produces current (see Fig. 1). The connection of the metal-insulator transition with the condensation of gauge bosons (especially the Goldstone ones that are the phasons relevant to superconductivity) may hide some profound and even deeper relation between the metal-insulator transition and superconductivity, a relation that may be relevant to the phenomenon of the high-T_c superconductivity. Finally, the identification of the metallic state with the holonomies of the corresponding charged many-body system will undoubtedly offer formal advantages and calculational power (in the precise location of the metal-insulator transition) in the future, since the subject of quantum gauge field theories in multiply connected space is currently under intense investigation.

ACKNOWLEDGMENT

This work has been supported by the National Science Foundation under Grant No. DMR-8715590.

APPENDIX A

Consider the ground state of a neutral system of N positive charges (e) with mass m_p and N negative charges ($-e$) with mass m_e (considered as point particles) in volume V and in the presence of a static and homogeneous gauge field $\mathbf{A}=A\hat{\mathbf{x}}$ ($\hat{\mathbf{x}}$ is supposed to be along a possibly conducting direction). Omitting relativistic corrections the Hamiltonian of this system is

$$H=\sum_{i=1}^{N}\frac{[\mathbf{p}_{i,e}-(e/c)\mathbf{A}]^2}{2m_e}+\sum_{i=1}^{N}\frac{[\mathbf{p}_{i,p}+(e/c)\mathbf{A}]^2}{2m_p}+\tfrac{1}{2}\sum_{i}\sum_{j\neq i}\frac{e^2}{|\mathbf{r}_{i,e}-\mathbf{r}_{j,e}|}+\tfrac{1}{2}\sum_{i}\sum_{j\neq i}\frac{e^2}{|\mathbf{r}_{i,p}-\mathbf{r}_{j,p}|}-\sum_{i}\sum_{j}\frac{e^2}{|\mathbf{r}_{i,e}-\mathbf{r}_{j,p}|}\ . \qquad \text{(A1)}$$

We rescale all variables using

$$V=N\tfrac{4}{3}\pi r_0^3,\quad r_s=\frac{r_0}{a},\quad a=\frac{\hbar^2}{\overline{m}e^2}=\frac{M}{m_p}a_0\ ,$$

$$M=m_e+m_p,\quad \bar{m}=\frac{m_e m_p}{M}$$

and introduce $\bar{r}$, $\bar{p}$, etc., according to the definitions

$$V=\bar{V}r_0^3,\quad \mathbf{r}_e=r_0\bar{\mathbf{r}}_e,\quad \mathbf{r}_p=r_0\bar{\mathbf{r}}_p\ ,$$

$$\mathbf{p}_e=\hbar\bar{\mathbf{p}}_e/r_0,\quad \mathbf{p}_p=\hbar\bar{\mathbf{p}}_p/r_0\ .$$

Equation (8) then reads

$$H=\frac{e^2}{2a}\frac{1}{r_s^2}\left[\frac{m_p}{M}\sum_{i=1}^{N}\left[\bar{\mathbf{p}}_{i,e}-\frac{ear_s}{\hbar c}A\hat{\mathbf{x}}\right]^2+\frac{m_e}{M}\sum_{i=1}^{N}\left[\bar{\mathbf{p}}_{i,p}+\frac{ear_s}{\hbar c}A\hat{\mathbf{x}}\right]^2\right.$$
$$\left.+r_s\sum_i\sum_{j\neq i}\frac{1}{|\bar{\mathbf{r}}_{i,e}-\bar{\mathbf{r}}_{j,e}|}+r_s\sum_i\sum_{j\neq i}\frac{1}{|\bar{\mathbf{r}}_{i,p}-\bar{\mathbf{r}}_{j,p}|}-2r_s\sum_i\sum_j\frac{1}{|\bar{\mathbf{r}}_{i,e}-\bar{\mathbf{r}}_{j,p}|}\right]\equiv T+U\ , \tag{A2}$$

where T and U are the total kinetic- and potential-energy operators, respectively. We now take derivatives of (A1) or (A2) with respect to various parameters, and at the same time apply the Hellmann-Feynman theorem;[11] for example, differentiation with respect to r_s gives

$$\frac{\partial\langle\Psi|H|\Psi\rangle}{\partial r_s}=\left\langle\Psi\left|\frac{\partial H}{\partial r_s}\right|\Psi\right\rangle\ , \tag{A3}$$

(where $H|\Psi\rangle=E|\Psi\rangle$). It is straightforward to show that if (A3) is applied to (A2) and combined with the definition of the total current

$$\mathbf{J}=\frac{e}{V}\left\langle\sum_{i=1}^{N}\left[\left[\frac{\mathbf{p}_{i,p}+(e/c)\mathbf{A}}{m_p}\right]-\left[\frac{\mathbf{p}_{i,e}-(e/c)\mathbf{A}}{m_e}\right]\right]\right\rangle$$

it finally gives (with $E=\langle T\rangle+\langle U\rangle$)

$$\frac{d}{dr_s}\left[\frac{E}{N}\right]=-\frac{1}{r_s}\left[\frac{\langle T\rangle}{N}+\frac{E}{N}\right]+\frac{1}{r_s}\frac{V}{N}\frac{1}{c}\mathbf{A}\cdot\mathbf{J} \tag{A4}$$

which has exactly the same form as Eq. (1) of the text.

On the other hand, differentiation of (A1) with respect to each mass independently leads to

$$\langle T\rangle=-m_e\frac{\partial E}{\partial m_e}-m_p\frac{\partial E}{\partial m_p}\ , \tag{A5}$$

and differentiation with respect to e leads to

$$\langle U\rangle=\frac{1}{2}\left[e\frac{\partial E}{\partial e}-\frac{V}{c}\mathbf{A}\cdot\mathbf{J}\right]\ . \tag{A6}$$

A final differentiation with respect to A leads to

$$J=\frac{c}{V}\frac{\partial E}{\partial A}\ . \tag{A7}$$

Combination of (A4), (A5), and (A7) then yields

$$r_s\frac{\partial E}{\partial r_s}-m_e\frac{\partial E}{\partial m_e}-m_p\frac{\partial E}{\partial m_p}-A\frac{\partial E}{\partial A}=-E(r_s,m_e,m_p;A)\ . \tag{A8}$$

Similarly, combination of (A4), (A6), and (A7) yields

$$r_s\frac{\partial E}{\partial r_s}-\tfrac{1}{2}e\frac{\partial E}{\partial e}-\tfrac{1}{2}A\frac{\partial E}{\partial A}=-2E(r_s,e;A)\ . \tag{A9}$$

Equations (A8) and (A9) are *exact* partial differential equations for the full quantum-mechanical internal energy ($E\equiv\langle H\rangle$). The general solutions are easily found to possess the scaling forms

$$\frac{E}{N}=\frac{f(m_e r_s,m_p r_s;Ar_s)}{r_s} \tag{A10}$$

and

$$\frac{E}{N}=\frac{g(e^2 r_s;A^2 r_s)}{r_s^2}\ , \tag{A11}$$

which are *exact* properties of the ground-state energy of the two-component Coulomb system. These scaling relations are independent of the phase of the nuclei, and are valid to all orders with respect to the gauge field A.

The same method can be applied to the standard one-component Coulomb system of point particles in a rigid uniform compensating background of opposite sign. In this case the Hamiltonian consists only of the first and third terms of (A1). Invoking exactly the same procedure as above leads again to Eq. (1) of the text, and also yields the partial differential equations

$$r_s\frac{\partial E}{\partial r_s}-m\frac{\partial E}{\partial m}-A\frac{\partial E}{\partial A}=-E(r_s,m;A)\ , \tag{A12}$$

as well as Eq. (A9), with the general solutions being in the form (2) of the main text.

It has therefore been shown in this appendix that Eq. (1) has the same form for both the two- and one-component systems, although the quantities entering (such as $\langle T\rangle$ or $\mathbf{J}$) have different values (and origins) in the two cases. As a result, Eq. (1) provides a criterion for the metal-insulator transition for *both* cases, in terms of the relation between the establishment of the insulating phase and the complete lack of dependence of the energy E on the gauge field A. The generality of (1) shows that, in the two-component case, the criterion neither requires further information on the phase of the nuclei nor does it provide it; similarly, in the one-component case, it does not require any special treatment of the rigid background. However, in the latter case, a translational symmetry-broken state can be expected, as discussed independently in Secs. III and IV.

APPENDIX B

Equation (1) of the text indicates that the ratio of the change in energy (that distinguishes between a metal and an insulator) to the value of $\mathbf{J}\cdot\mathbf{A}$ is an extensive quantity. Although this is shown explicitly in Appendix A for a simply connected space, we also show it here by considering a multiply connected region (e.g., a torus) and then taking the thermodynamic limit in the way described in the text. At the same time, we explicitly calculate here the values of α and β of Eq. (3) of the text.

Consider electrons on a torus with vanishing electric and magnetic fields but with a nonzero flux Φ existing only in the hole. Since on the physical region we have $\nabla\cdot\mathbf{A}=\nabla\times\mathbf{A}=0$, we must have $\mathbf{A}=A\hat{\phi}$ and also

$$\Phi=AL=A2\pi r\ . \tag{B1}$$

For noninteracting electrons this gives a Hamiltonian

$$H=\sum_i\left[\frac{(p_{r,i}^2+p_{z,i}^2)}{2m}+\frac{\hbar^2}{2mr^2}\left[n_i-\frac{\Phi}{\Phi_0}\right]^2\right] \tag{B2}$$

with $\Phi_0=hc/e$ and $n_i=$integer. This appears to show that the energy, if written in terms of the enclosed flux Φ, would go as $L^{(d-2)}$ and is therefore not extensive (in our case $d=3$, but it could as well be $d=2$ if we had considered a cylinder, or $d=1$ if we had considered a ring[6]). However, if we write (B2) in terms of A, by using (B1), and then take the thermodynamic limit precisely as described in the text to produce a simply connected physical region but with a constant $\mathbf{A}$ along, say, direction $\hat{\mathbf{x}}$ (what was formerly direction $\hat{\phi}$), then we can easily see that the ratio of the energy to A^2 is indeed an extensive number [essentially because of the cancellation of r^2 in (B2)]. This conclusion can also be reached by Eq. (A4) of Appendix A, where no Φ but only an $\mathbf{A}$ is considered (actually in that case of simply connected space the corresponding Φ would be infinite after the thermodynamic limit is taken). The important point here is the use of the right conjugate variables for the description of a thermodynamic system: the variable conjugate to $\mathbf{J}$ for our infinite system is $\mathbf{A}$ and *not* an enclosed flux.

In more detail, this mapping between the torus and the simply connected space produces wave vectors $k_x=2\pi/L(n-\Phi/\Phi_0)$, $k_y=(2\pi/L_y)n_y$, and $k_z=(2\pi/L_z)n_z$, i.e., the ground state of noninteracting electrons is a Fermi sphere but displaced along the $\hat{\mathbf{x}}$ direction by $-2\pi(A/\Phi_0)$. The k space density of levels is therefore $V/(2\pi)^3$, which gives $k_F^3=3\pi^2(N/V)$ for the radius of the displaced Fermi sphere. A subsequent calculation of the energy will then give

$$E=2\frac{V}{(2\pi)^3}\int_{\mathrm{DFS}}d^3k\frac{\hbar^2k^2}{2m}\ ,$$

where DFS stands for displaced Fermi surface. If we write $\mathbf{k}=\mathbf{k}'+(2\pi A/\Phi_0)\hat{\mathbf{x}}$, then this finally gives

$$\frac{E}{N}=\frac{2.21}{r_s^2}\ \mathrm{Ry}+\frac{e^2}{2mc^2}A^2\ , \tag{B3}$$

and once again we note that the coefficient of A^2 *is* extensive.

Finally, we can proceed in the same spirit to the case of interacting fermions. Hence we can include exchange through the standard Hartree-Fock approximation, and correlation through a RPA, merely shifting the x component of all wave vectors by $-2\pi A/\Phi_0$. It is shown in the text, by use of the scaling relations of Appendix A, that if we wish the energy for small A, we can consider fermions with $m\rightarrow\infty$ and take the limit $r_s\rightarrow 0$ in such a way that the product mr_s remains a constant. On the metallic side, where this constant is small, the above procedure leads to the following results.

(1) There is no A dependence of the exchange energy, i.e., $E_{\mathrm{ex}}=-0.916/r_s$ Ry.

(2) The correlation energy has the form

$$E_{\mathrm{corr}}=\frac{2}{\pi^2}(1-\ln 2)\ln\left|\frac{\lambda-\delta^2}{q_c^2-\delta^2}\right|\ \mathrm{Ry}\ , \tag{B4}$$

where $\lambda=2(4/9\pi)^{1/3}\tilde{r}_s$, q_c is a dimensionless parameter, and $\delta=(2\pi A/\Phi_0)/k_F$. Here $\tilde{r}_s=(m/m_e)r_s$ with $m\rightarrow\infty$ and $r_s\rightarrow 0$ again in such a way that mr_s is finite, and also in such a way that $\lambda\ll q_c^2\ll 1$. Equation (B4) has the standard RPA behavior[30] in the absence of the gauge field ($\delta=0$) and for the particular choice $m=m_e$. Finally, an expansion of (B4) around $\delta\sim 0$ leads to

$$\frac{E_{\mathrm{corr}}}{N}=\frac{E^{\mathrm{RPA}}(A=0)}{N}-\frac{e^2}{2mc^2}A^2\left[\frac{4}{9\pi}\right]^{1/3}\frac{(1-\ln 2)}{\pi^2}\tilde{r}_s\ . \tag{B5}$$

The results (B3) and (B5) determine the values of α and β which appear in the text under Eq. (3) for the weak-coupling region. In turn, these values imply a critical point $r_{s,0}$ for the metal-insulator transition as is discussed in the text.

[1] N. F. Mott, *Metal-Insulator Transitions,* 2nd ed. (Taylor and Francis, London, 1991).

[2] *The Metallic and Nonmetallic States of Matter,* edited by P. P. Edwards and C. N. R. Rao (Taylor and Francis, London, 1985).

[3] S. V. Vonsovskii and M. I. Katsnel'son, Usp. Fiz. Nauk **158**, 723 (1989) [Sov. Phys. Usp. **32**, 720 (1989)].

[4] K. Moulopoulos and N. W. Aschcroft (unpublished).

[5] W. Kohn, Phys. Rev. **133A**, 171 (1964).

[6] B. S. Shastry and B. Sutherland, Phys. Rev. Lett. **65**, 243 (1990).

[7] W. Ehrenberg and R. E. Siday, Proc. Phys. Soc. London Sect. B **62**, 8 (1949); Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).

[8] M. V. Berry, Proc. R. Soc. London, Ser. A **392**, 45 (1984).

[9] P. C. Martin, *Many-Body Physics* (Gordon and Breach, New York, 1968).

[10] L. P. Kadanoff and P. C. Martin, Phys. Rev. **124**, 677 (1961).

[11] K. Moulopoulos and N. W. Ashcroft, Phys. Rev. B **41**, 6500 (1990).

[12] E. L. Feinberg, Usp. Fiz. Nauk **78**, 53 (1962) [Sov. Phys. Usp. **5**, 753 (1963)].

[13] H. Kleinert, Fortschr. Phys. **26**, 565 (1978).

[14] K. Moulopoulos and N. W. Ashcroft, Phys. Rev. Lett. **66**, 2915 (1991).

[15] R. L. Stratonovich, Dokl. Akad. Nauk. SSSR **115**, 1097 (1957) [Sov. Phys. Dokl. **2**, 416 (1958)]; J. Hubbard, Phys. Rev. Lett. **3**, 77 (1959).

[16] U. Eckern, G. Schön, and V. Ambegaokar, Phys. Rev. B **30**, 6419 (1984).

[17] G. Schön and A. D. Zaikin, Phys. Rep. **198**, 237 (1990).

[18] N. Byers and C. N. Yang, Phys. Rev. Lett. **7**, 46 (1961).

[19] D. Pines and P. Nozieres, *The Theory of Quantum Liquids* (Benjamin, New York, 1966).

[20] N. Kawakami and S. Yang, Phys. Rev. Lett. **65**, 3063 (1990).

[21] F. J. Rogers *et al.*, Phys. Rev. A **1**, 1577 (1970).

[22] H. Matsumoto, G. Semenoff, H. Umezawa, and M. Tachiki, Fortschr. Phys. **28**, 67 (1980).

[23] H. Umezawa, H. Matsumoto, and M. Tachiki, *Thermo Field Dynamics and Condensed States* (North-Holland, Amsterdam, 1982).

[24] S. N. Gupta, Proc. Phys. Soc. London Sect. A **63**, 681 (1950); K. Bleuler, Helv. Phys. Acta **23**, 567 (1950).

[25] M. Wadati, Phys. Rev. D **18**, 520 (1978); see also W. Kohn and D. Sherrington, Rev. Mod. Phys. **42**, 1 (1970), Appendix II.

[26] J. Q. Liang, Phys. Lett. A **142**, 11 (1989).

[27] Y. Hosotani, Ann. Phys. (N.Y.) **190**, 233 (1989); see also M. Burgess and D. J. Toms, *ibid.* **210**, 438 (1991).

[28] J. Ihm, Phys. Rev. Lett. **67**, 251 (1991).

[29] J. Riess, Phys. Rev. B **38**, 3133 (1988).

[30] See, for example, Eq. (12.60) of A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971).