REVIEWS OF MODERN PHYSICS VOLUME 40, NUMBER 4 OCTOBER 1968

SESSION IV—ELECTRONS IN WIDE BANDS

Possible Anomalies at a Semimetal–Semiconductor Transistion

B. I. HALPERIN, T. M. RICE
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

An indirect band-gap semiconductor may be converted to a semimetal, or vice versa, by application of pressure. At low temperature, an excitonic phase or some other anomaly must occur in the neighborhood of the transition pressure. We also discuss the direct-band-gap case and the case where a band gap is zero by symmetry, as in gray tin.

I. INTRODUCTION

According to the one-electron picture of solids, a perfect crystal with an even number of electrons per unit cell is classified as either an insulator or a metal depending on its electronic band structure. If the minimum energy of the conduction band lies above the maximum of the valence band by a finite energy gap $E_G > 0$, then the crystal is called an insulator (or semiconductor if E_G is small). At absolute zero temperature, the valence band of a semiconductor is filled and the conduction band empty; there are no free carriers and the dc conductivity is zero. If there is an overlap between the valence and the conduction band ($E_G < 0$), then the crystal is considered a metal or a semimetal. The ground state of the crystal will have a certain number of electrons in the conduction band and an equal number of holes in the valence band. At 0°K the material will have a nonzero conductivity whose magnitude will be inversely proportional to the density of scattering centers in the crystal, a density which we assume to be very small.

At finite temperatures both semiconductor and semimetal will have free carriers, and will have finite conductivities proportional to the number of carriers and inversely proportional to the scattering rates due to impurities and/or phonons.

The energy gap E_G, in the simple one-electron model, will be a function of external parameters such as the lattice constants, and may be made to vary under applied pressure or uniaxial strain. (The gap will also depend somewhat on the temperature because of renormalizations in the one-electron potential due to changes in the occupation numbers of the states.) We may easily conceive of a case where E_G could be varied through zero under pressure. The phase diagram for such a model is indicated in Fig. 1. The dotted curve is the locus of the points where the temperature renormalized gap $E_G(T)$ is equal to zero; it divides the nominally semiconducting region (region 1) from the nominally semimetallic region (region 2). Of course, there are no discontinuities in the electrical conductivity or any other physical properties along the curve $E_G(T) = 0$. There is only a singular point ($E_G = 0$, $T = 0$) where there is a discontinuity in the behavior of the conductivity.

When electron–electron interactions are taken into account, the simple phase diagram of Fig. 1 will no longer be valid for a semimetal–semiconductor transition in a pure material. At high temperatures, it should still be possible to go continuously from region 1 to region 2. However, we believe that there exists a temperature $T_M > 0$ below which one can go from region 1 to region 2 only by passing through one or more phase transitions. Thus there must be some anomalous regions in the phase diagram at low temperatures, in the vicinity of the point $E_G = 0$. It is the purpose of this paper to explain briefly why we believe that there must be an anomaly in this region, and to speculate on some possible kinds of anomalies that might occur. We also briefly discuss the prospects for experimental observation of these anomalies. We restrict ourselves to materials having wide valence and conduction bands compared to the strength of the electron–electron interaction.

II. ARGUMENTS FOR THE EXISTENCE OF AN ANOMALY

The first argument that there must be an anomaly at the semimetal–semiconductor transition, was advanced by Mott.[1] He considered a semimetal with a very small number of electrons and holes and noticed the following: When the number of carriers is sufficiently small, the Coulomb interaction between electrons and holes is only weakly screened. An unscreened Coulomb interaction will always lead to a bound state between an electron and a hole. Therefore, argued Mott, when the density of carriers was low enough,

---

[1] N. F. Mott, Phil. Mag. 6, 287 (1961).

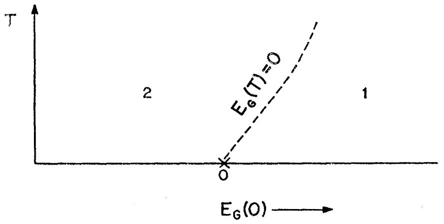

FIG. 1. Phase diagram for the semimetal–semiconductor transition in the simple one-electron band picture. The abscissa of this plot measures the density of the crystal and is labeled by $E_G(0°\mathrm{K})$, the value of the indirect energy gap at 0°K for a crystal of the given density. Region 1 is nominally semiconducting, while region 2 is norminally semimetallic. The dividing line between the two regions, $E_G(T)=0$, does not represent a phase transition at finite temperatures; the only singular point is $E_G=T=0$.

the electrons and holes would spontaneously form nonconducting bound states, and the semimetal would become an insulator. Of course, such a transition could only occur at temperatures smaller than the electron–hole binding energy.

A second argument was advanced by Knox,[2] who considered the case of an indirect-band-gap semiconductor. The minimum energy necessary to create an exciton in a semiconductor is equal to E_G-E_B, where E_B, the binding energy of an exciton, is determined by the effective masses of the electron and hole, and by the dielectric constant of the material:

$$E_B=\mu/m\epsilon^2 \quad \text{(rydbergs)}, \tag{1}$$

where μ is the reduced effective mass of the electron and hole bands, ϵ is the dielectric constant, and m is the bare mass of an electron. The dielectric constants and effective masses may depend sensitively on the direct energy gap, but not on the indirect energy gap. Thus we may make E_G arbitrarily small while E_B remains finite. When E_G becomes less than E_B, the energy necessary to create an exciton becomes negative, and the ordinary ground state of the crystal becomes unstable with respect to the spontaneous formation of excitons.

Before we turn to speculations on the kinds of anomalies which may arise, let us note that a crucial point in the above arguments has been that the dielectric constant remains finite in the normal insulating ground state as the semimetal transition is approached. Certainly this is the case in the Hartree–Fock approximation for a pure indirect-gap semiconductor with vanishing energy gap, and no arguments have been advanced that suggest this should not be the case when terms beyond the Hartree–Fock are included. Nevertheless, a conclusion of this nature should always be regarded with some caution.

In contrast to the case of a pure semiconductor, Mott and Davis[3] have argued that the dielectric constant does diverge in certain *disordered* insulators when their parameters are varied towards the critical values for nonzero electrical conductivity at 0°K. In particular, Mott and Davis considered the conductivity of a doped semiconductor (with fixed energy gap) as a function of impurity concentration, and found a phase diagram in which there is a singularity only at 0°K, as is the case in Fig. 1. Below a certain critical doping, the semiconductor has zero conductivity at 0°K. Above the critical impurity concentration, the material has finite electrical conductivity at 0°K. At finite temperatures, however, there is no singularity in the electrical conductivity or in any other physical property as a function of impurity concentration.

We believe that a semimetal–semiconductor transition with no anomaly at finite temperatures can also occur when an indirect energy gap in an impure semiconductor is varied through zero, if the scattering is sufficiently great. In this case the simple phase diagram of Fig. 1 is again applicable, except that the energy gap E_G can no longer be precisely defined. Very roughly, the criterion for suppression of the finite temperature phase transitions should be $\hbar/\tau \geq E_B$, where τ is the scattering lifetime.[4,5] In addition, if the semiconductor is sufficiently doped with charged impurities so that it is degenerate and has finite conductivity at 0°K, in the region $E_G \gg 0$, then there need be no anomaly in the dc conductivity even at 0°K, when the energy gap varies from positive to negative.[4]

III. SIMPLE FIRST-ORDER TRANSITION

Let us return, now, to the case of the pure crystal, and consider some of the anomalies that might occur at the semimetal–semiconductor transition. Conceptually, the simplest possible anomaly would be to have a direct first-order transition, at low temperatures, between the semiconductor, region 1, and the semimetal, region 2. This is the kind of transition which is suggested

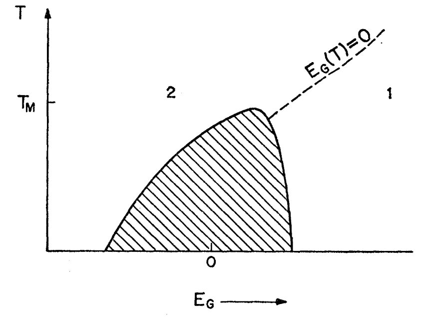

FIG. 2. Phase diagram with direct first-order transition, at low temperatures, from semimetal to semiconductor. Region 1 is the undistorted semiconductor, region 2 the undistorted semimetal. The shaded region is an area in which no stable phase exists. The semimetal–semiconductor transition below T_m is accompanied by a discrete change in lattice parameter and a jump from one side of the shaded region to the other.

---

[2] R. S. Knox, Solid State Phys. Suppl. **5**, 100 (1963).

[3] N. F. Mott and E. A. Davis, Phil. Mag. **17**, 1269 (1968).

[4] B. I. Halperin and T. M. Rice in *Solid State Physics*, F. Seitz, D. Turnbull, and H. Ehrenreich, Eds. (Academic Press Inc., New York, 1968), Vol. 21.

[5] J. Zittartz, Phys. Rev. **164**, 575 (1967).

by the original article of Mott.[1] The phase diagram in this situation is indicated in Fig. 2. The shaded region represents a range of temperature and lattice constant for which no stable state exists. A crystal subjected to varying applied pressure, at a constant temperature below T_M, would jump directly from the left-hand side of the shaded region to the right-hand side at the transition pressure. Above T_M, the crystal would pass directly from the semiconductor to the semimetal without any kind of phase transition.

The phase diagram of Fig. 2 would be the simplest way to avoid the excitonic instability of Knox and Mott, but it is not the kind of anomaly which is most naturally suggested by these instabilities. Furthermore, although a "large" first-order transition of this nature is very difficult to rule out in general, there are certainly no calculations which suggest that this kind of transition must occur at the semimetal–semiconductor boundary.

IV. THE EXCITONIC STATE

A. Description

The type of anomaly which seems most likely to occur is that there is a new distorted phase of the crystal, known as the "excitonic state," which occurs in a region of the temperature–energy gap plane surrounding the transition point ($E_G=0$, $T=0$). The theory of the excitonic phase, in the Hartree–Fock approximation, was first developed by des Cloizeaux[6] and by Keldysh and Kopaev.[7] Further details of the state have been worked out by Kozlov and Maximov,[8] by Baklanov and Chaplik,[9] by Kopaev,[10] by Jérome, Rice, and Kohn,[11–13] by Zittartz,[5,14–16] and by the present authors.[4] Physically, the excitonic phase is characterized either by an antiferromagnetic order or by a lattice distortion and charge-density oscillation, either case resulting in a change in the crystal periodicity.[6] (We restrict our discussion, for the present to the case of an indirect-band-gap material, reserving the direct-band-gap case until the last section.)

Theoretical study of the excitonic state is simplified by the fact that the interband dielectric constants in semiconductors and semimetals, although finite, are generally large compared to 1. This means that exciton-binding energies are much smaller than typical atomic energies and the exciton radii are large compared to the atomic spacings. It is then an excellent first approximation to treat the electrons and holes in an effective mass approximation, and to consider only the long range attractive part of the Coulomb interaction between electrons and holes. Let us simplify our problem still further by considering a two-band situation with a single nondegenerate conduction-band minimum and a single nondegenerate valence maximum. From time reversal invariance arguments, it follows that the band extrema must either lie at the zone center ($\mathbf{k}=0$) or at a point on the zone boundary ($\mathbf{k}=\mathbf{w}$) such that $\mathbf{w}$ differs from $-\mathbf{w}$ by a reciprocal lattice vector $\mathbf{G}$; i.e., $\mathbf{w}=\frac{1}{2}\mathbf{G}$. We assume that the valence maximum is at $\mathbf{k}=0$ and that the conduction minimum is at $\mathbf{k}=\mathbf{w}$.

Let us focus our attention on the crystal at 0°K, in the semiconducting region with E_G close to E_B. For $E_G>E_B$, no excitons are present in the ground state of the crystal, and the nondistorted state is stable. For $E_G<E_B$, excitons are present. In this region, treating the crystal in a Hartree–Fock approximation is roughly equivalent to treating the excitons as a weakly repulsive Bose gas.[17] Excitons form until the repulsive potential cancels the negative energy (E_G-E_B) associated with the creation of a single exciton. Furthermore, and most important, the excitons present will form a Bose condensate in the exciton state of minimum energy, namely the state with wave vector $\mathbf{w}$. Let $A_{\mathbf{w}}^{\dagger}$ be the creation operator for an exciton with wave vector $\mathbf{w}$. This operator can be written as

$$A_{\mathbf{w}}^{\dagger}=\sum_{\mathbf{k},\sigma,\sigma'} f_{\sigma\sigma'}(\mathbf{k})\, b_{\mathbf{k}+\mathbf{w},\sigma}^{\dagger} a_{\mathbf{k},\sigma'}, \tag{2}$$

where $f_{\sigma\sigma'}(\mathbf{k})$ is an "envelope" wavefunction, peaked near $\mathbf{k}=0$, and where $a_{\mathbf{q},\sigma}^{\dagger}$ and $b_{\mathbf{q},\sigma}^{\dagger}$ are the creation operators for an electron of wave vector $\mathbf{q}$ and spin σ in the valence and conduction bands, respectively. A Bose condensate of excitons means that $\langle A_{\mathbf{w}}^{\dagger}\rangle\neq 0$, and hence we must also have

$$\langle b_{\mathbf{k}+\mathbf{w},\sigma}^{\dagger} a_{\mathbf{k},\sigma'}\rangle\neq 0 \tag{3}$$

for a suitable choice of σ and σ'. A quantity such as (3) must vanish in a periodic lattice unless $\mathbf{w}$ is the multiple of a reciprocal lattice vector of the lattice. The excitonic state must therefore have a period in real space which is just double the period of the undistorted lattice.[6]

In the Hartree–Fock picture, the one-electron states of the new distorted crystal are made up of linear combinations of states of wave vectors $\mathbf{k}$ and $\mathbf{k}+\mathbf{w}$ from the valence and conduction bands, respectively, of the old nondistorted crystal. When the self-consistent

---

[6] J. des Cloizeaux, J. Phys. Chem. Solids **26**, 259 (1965).

[7] L. V. Keldysh and Yu. V. Kopaev, Fiz. Tverd. Tela **6**, 2791 (1964) [English transl.: Soviet Phys.—Solid State **6**, 2219 (1965)].

[8] A. N. Kozlov and L. A. Maksimov, Zh. Eksperim. i Teor. Fiz. **48**, 1184 (1965) [English transl.: Soviet Phys.—JETP **21**, 790 (1965)].

[9] E. V. Baklanov and A. V. Chaplik, Fiz. Tverd: Tela **7**, 2768 (1965) [English transl.: Soviet Phys.—Solid State **7**, 2240 (1966)].

[10] Yu. V. Kopaev, Fiz. Tverd. Tela **8**, 223 (1966) [English transl.: Soviet Phys.—Solid State **8**, 175 (1966)].

[11] D. Jérome, T. M. Rice, and W. Kohn, Phys. Rev. **158**, 462 (1967).

[12] W. Kohn, in *Physics of Solids at High Pressures*, C. T. Tomizuka and R. M. Emrick, Eds. (Academic Press Inc., New York, 1965).

[13] W. Kohn, Phys. Rev. Letters **19**, 439 (1967).

[14] J. Zittartz, Phys. Rev. **162**, 752 (1967).

[15] J. Zittartz, Phys. Rev. **165**, 605 (1968).

[16] J. Zittartz, Phys. Rev. **165**, 612 (1968).

[17] Cf. the discussion of L. V. Keldysh and A. N. Kozlov, Zh. Eksperim. i Teor. Fiz. Pisma **5**, 238 (1967) [English transl.: JETP Letters **5**, 190 (1967)].

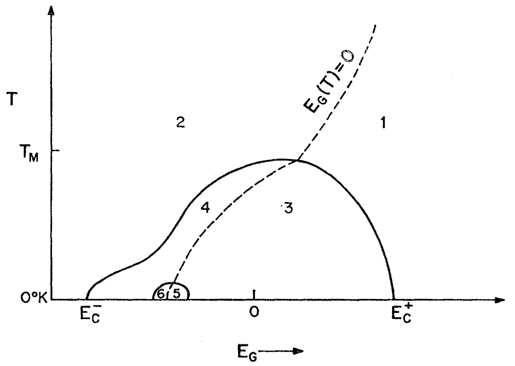

FIG. 3. Phase diagram for semimetal-to-semiconductor transition assuming second-order phase transitions to the excitonic phase. The abscissa E_G is a function of crystal density, and is the value of the indirect energy gap in the *nondistorted* ground state for the given density. The stable phase is nondistorted in regions 1 and 2, a simple excitonic distorted state in regions 3 and 4, a doubly distorted state in regions 5 and 6, etc. Regions 1, 3, and 5 are nominally semimetallic, regions 2, 4, 6 semiconducting. The dividing line (dotted curve) is the place where the renormalized energy gap $E_G(T)$ goes to zero, and does not represent any singularities in the physical properties of the system.

Hartree–Fock potential is calculated for the distorted state, the exchange potential has terms with periodicity $2\pi/w$; in order for the distorted state to have lower energy than the nondistorted state, the exchange potential must be strong enough to produce the required admixture of valence and conduction band states.

The interpretation of the Hartree–Fock excitonic state in terms of a dilute Bose exciton gas is only valid for E_G close to E_B. As E_G decreases through zero, the average distance between electrons or holes becomes smaller than the radius of the exciton. The Coulomb interaction becomes more and more strongly screened as E_G becomes more and more negative, and significant mixing of valence and conduction band states becomes restricted to states near the Fermi surface of the nondistorted semimetal. In this region the Hartree–Fock theory of the excitonic state is mathematically similar to the BCS theory of superconductivity.[7] If the electron and hole Fermi surfaces are identical in size and shape, as occurs in a simple two-band model with isotropic effective masses, then the nondistorted semimetal ground state is unstable for arbitrarily weak electron–hole attraction, just as the normal Fermi surface of a metal is unstable to the formation of Cooper pairs, if there is an arbitrarily weak attraction between spin-up and spin-down electrons.

Because of the relative simplicity of the calculations, many studies of the excitonic state have been carried out in the limit of the isotropic two-band semimetal with weak effective electron–hole interaction. The same model has also been used by Fedders and Martin[18] in their study of the antiferromagnetism of chromium.

For a real semimetal, the electron and hole Fermi surfaces are never identical in shape. There is therefore a minimum value $E_C^- < 0$ of the gap E_G below which the screened Coulomb interaction is too weak to support an excitonic distorted state, even at 0°K.[10] Quantitative calculations of the effects of anisotropy on the value of E_C^- have been undertaken by Zittartz.[14] For moderate anisotropy, the magnitude of E_C^- is comparable to the value of E_B.

B. Nature of the Distortion

The nature of the distortion in the excitonic state depends on whether the expectation value $\langle A_w^\dagger \rangle$ is real or imaginary, and also on whether the macroscopically occupied exciton state is a singlet or a triplet. In a model without spin–orbit coupling, the four possibilities are:

(1) Singlet with real phase—characterized by a charge-density oscillation.

(2) Triplet with real phase—characterized by an antiferromagnetic spin density oscillation.[19]

(3) Singlet with imaginary phase—characterized by transverse currents which change sign from one unit cell to the next (i.e., orbital antiferromagnetism).

(4) Triplet with imaginary phase—characterized by transverse spin currents. [In the presence of spin–orbit coupling, states (2) and (3) would be mixed, as would be states (1) and (4).]

The energies of these four states are degenerate if one includes in the electron–hole interaction only the dominant term, the long-range Coulomb attraction. The energies are split, however, by terms in the Hamiltonian representing interband scattering when the electron and hole are on the same lattice site. The energy splittings are proportional to the ratio of the unit cell volume to the volume of an exciton: these splittings are therefore very small for loosely bound excitons. The magnitude of the spin- or charge-density wave or current is also very small.

By treating the splitting terms as a small perturbation in the Hartree–Fock theory, we have found that the lowest energy state, in the simplest models, is always the spin-density wave state (2).[4] If coupling to the phonons is included in the model, however, and if the coupling to the phonons is sufficiently strong, then the charge-density oscillation can have a lower energy. The charge-density oscillation is, of course, coupled to a lattice distortion which doubles the lattice period, whereas in the antiferromagnetic state the lat-

---

[18] P. A. Fedders and P. C. Martin, Phys. Rev. **143**, 245 (1966).

[19] The idea that, under some circumstances, the energy of the Hartree–Fock ground state of a crystal could be lowered by the introduction of an antiferromagnetic spin oscillation was suggested by J. C. Slater, Phys. Rev. **82**, 538 (1951). Later, A. W. Overhauser [Phys. Rev. Letters **4**, 462 (1960)] proposed that the nondistorted ground state of a metal will always be unstable in favor of a state with a spin-density wave. Although it is in fact true in the strict Hartree–Fock approximation that the normal metallic ground state always has this instability, this is not true when the exchange interaction is screened by the dielectric function of the electron gas. It is believed by most authors that the nondistorted state will be unstable only for special kinds of band structures (see, for example, Ref. 18).

tice period remains the same and only the electronic period is doubled. In models with more than one valence-maximum and/or more than one conduction-minimum, the number of possible states is much greater than in the two-band model. Some relatively simple cases have been discussed in Ref. 4. In the absence of strong phonon coupling, the lowest excitonic states again seem to involve spin-density waves. The period of the oscillations is roughly 2π divided by the distance in $\mathbf{k}$ space between some valence maximum and some conduction maximum, and is no longer simply twice the period of the undistorted lattice.

The two possibilities of a lattice distortion or of antiferromagnetism have also been proposed in connection with the metal insulator transitions in the vanadium oxides.[20,21] There, however, the electron–hole interaction is strong and any excitons are tightly bound. (In fact, a metal–insulator transition in a model with an odd number of electrons per unit cell can only occur if the electron–electron interaction is large compared to the width of the electron energy band.) For a tightly bound exciton, the singlet–triplet splitting will not be small but may be comparable to the exciton binding energy itself. The splitting between a spin-density wave state and a lattice distortion may also be relatively large in the model of a weakly interacting isotropic two-band semiconductor.[22]

C. Phase Diagram

At 0°K, the excitonic state has lower energy than the nondistorted state for a range of densities given by

$$E_G^- < E_G < E_G^+ = E_B. \tag{4}$$

For any density in this range there is a finite temperature T_c, above which the nondistorted state again becomes stable. Calculations of T_c at various points in the range (4) have been attempted by a number of authors.[8,10,14] In all these calculations, the transition was found to be second order.

The maximum value of T_c should occur for $E_G \approx 0$, or perhaps better, when $E_G(T_c) \approx 0$. Let us write the value of T_c at $E_G = 0$ in the form

$$kT_c \,|_{E_G(T_c)=0} = CE_B, \tag{5}$$

where C is a constant which depends on the dimensionless parameters of the model, i.e., on the ratio of electron to hole effective masses, on the degree of anisotropy, and on the number of electron and hole pockets in the Brillouin zone. Zittartz[14] has calculated C by means of an approximate evaluation of the Hartree–Fock equations for a number of two-band cases. For isotropic effective masses, with $m_e = m_h$, he finds

$$C \approx 1.4. \tag{6}$$

Zittartz has neglected, however, the change in the effective electron–hole interaction at finite temperatures which arises from free-carrier screening. When screening is taken into account, there will probably be a significant reduction in the value of T_c.

When m_h is much greater than m_e, an approximate evaluation of the (unscreened) Hartree–Fock equations gives[23]

$$C \approx \tfrac{1}{4}[\log(m_h/m_e)]^{-1}. \tag{7}$$

On the other hand, if screening is included in the Hartree–Fock approximation, we find roughly

$$C \approx \tfrac{1}{5}[\log(m_h/m_e)]^{-3}, \tag{8}$$

in the limit of $m_h \gg m_e$. Reasoning beyond the Hartree–Fock type of approximation leads to still a different answer. We may estimate the transition temperature to the excitonic phase by calculating the Bose condensation temperature for a gas of bosons with the translational excitonic mass, $M = m_e + m_h$, at the density equal to the density of excitons in the excitonic phase at $E_G = 0$, a density roughly proportional to the inverse cube of the exciton radius. Since this last quantity depends on the reduced mass rather than the total mass, we find

$$C \propto \mu/M \approx m_e/m_h. \tag{9}$$

For $T > T_c$ but much less than E_B, the electrons and holes may be strongly correlated, leading to a significant reduction in the effective number of free carriers, even though no actual transition has taken place. Such effects are, of course, outside the Hartree–Fock approximation.

As we shall discuss in Sec. 5, the excitonic phase is probably replaced by another kind of transition if the mass ratio is too big ($m_h/m_e \gtrsim 100$). Thus, the differences between Eqs. (7), (8), and (9) cannot be very great in practice.

We have already indicated that impurity scattering tends to suppress the transition to the excitonic phase. Zittartz[5] has calculated the reduction in transition temperatures and increase in E_G^- when a finite concentration of impurities is added to an isotropic two-band semimetal.

The phase diagram for the pure material which has emerged from the various studies of the excitonic phase, using Hartree–Fock and related approximations, is given in Fig. 3. The abcissa E_G is here a measure of the atomic volume of the sample: it is the value of the energy gap at 0°K in a hypothetical undistorted crystal of the given density.[24] In regions 1 and 2 the undis-

---

[20] D. Adler and H. Brooks, Phys. Rev. **155**, 826 (1967).

[21] D. Adler in *Solid State Physics*, F. Seitz, D. Turnbull, and H. Ehrenreich, Eds. (Academic Press Inc., New York, 1968), Vol. 21.

[22] Yu. V. Kopaev, Fiz. Tverd. Tela **8**, 2633 (1966) [English transl.: Soviet Phys.—Solid State **8**, 2106 (1967)].

[23] Zittartz obtains the result $C \approx 0.7$ in the limit $m_h \gg m_e$. He has neglected, however, the temperature dependence of the Fermi level, which occurs for $m_e \neq m_h$.

[24] We use this convention for E_G in the remainder of this paper. The renormalized gap in the actual stable phase at a given density we denote by $E_G(T)$. The value E_G will coincide with $E_G(0°\mathrm{K})$ if and only if the stable state is nondistorted at the given density.

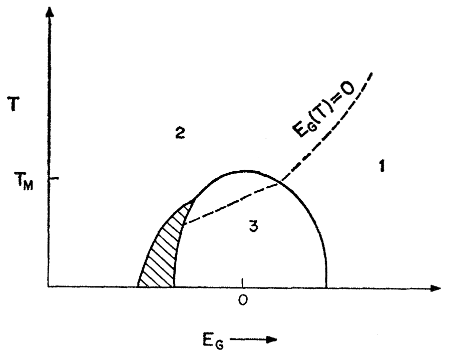

FIG. 4. Hypothetical phase diagram for a semimetal–semiconductor transition with an excitonic phase, under the assumption of a direct first-order phase transition from a nondistorted semimetal (region 2) to an excitonic semiconductor (region 3). Region 1 is the nondistorted semiconductor, while the shaded region is a region with no stable lattice.

torted crystal is stable. The crystal is characterized as a semiconductor in region 1 and a semimetal in region 2. The solid curve marks the location of a second-order transition to the excitonic state with a new crystal periodicity in regions 3 and 4. The new band structure, in the excitonic state can again be either "semiconducting" or "semimetallic" depending on whether its new, renormalized energy gap is positive or negative. The dotted curve in Fig. 3 is the locus of points where the renormalized energy gap $E_G(T)$ is equal to zero, and it thus divides the nominally semiconducting region from the nominally semimetallic region. Of course, no properties of the system are singular along this line at finite temperatures.

The dielectric constant in the excitonic semiconductor, region 3, will generally be larger than the dielectric constant in region 1, but it will remain finite as the renormalized energy gap approaches zero. Thus there will be a new exciton binding energy E_B' in the distorted phase which is smaller than E_B, but greater than zero. For temperatures much below E_B', we cannot pass directly from region 3 to region 4, but rather we must pass through a new, doubly distorted phase (region 5 and 6), with a lattice period different from those in regions 3 and 4 and regions 1 and 2. In fact, unless some place along the line there is a first-order transition, directly from a semiconducting state to a semimetal, we must find an infinite nested series of second-order transitions at the semimetal–semiconductor boundary in a perfect, pure crystal.[4,13,25] We cannot rule out the possibility, however, that there may *always* be a first-order semimetal-to-semiconductor transition at some point in this nested sequence of transitions. [A possible first-order transition, which might occur in a crystal with moderate anisotropy, which would be consistent with the occurrence of one excitonic state state (region 3), but not with an infinite sequence of nested transitions, is shown in Fig. 4.] Within the Hartree–Fock approximation it is relatively easy to find second-order transitions: one need only show that the nondistorted state becomes unstable with respect to an infinitesimal distortion of a certain kind. To show that such a transition is not actually first-order is much more difficult, however, since one must calculate the free-energy of all possible states with finite distortions. The situation is particularly complicated on the semimetal side, if one does not restrict oneself to an isotropic model, since one must then consider all possible distortions of the Fermi surface. In looking for a first-order transition it is important to include in the Hamiltonian not only the long-range electron–hole attraction, but also the terms involving the long-range electron–electron and hole–hole interactions.[26] [The latter terms are also the terms responsible for the finite-temperature renormalization of the one-electron energies, and of the energy gap $E_G(T)$ in the normal state due to changes in the occupation numbers of the states.] The present authors[4] have studied the transition at $E_C{}^+$ on the semiconducting end, and have calculated the energy of the system up to terms quartic in the order parameter. The sign of the quartic term was found to be positive, as is necessary for a second-order transition, but the coefficient of the quartic term was reduced by a factor of almost 5 when the electron–electron and hole–hole terms were included, as compared with the value when these terms were left out.

Another possible source of a first-order transition is the fact that in the excitonic state, screening of the Coulomb interaction is weaker, and the effective electron–hole interaction is stronger than in the nondistorted state. Kozlov and Maximov[8] and Baklanov and Chaplik[9] have considered this interaction renormalization for a two-band isotropic model in the semimetal limit ($E_G \ll 0$), and have concluded that the renormalization is too small to affect the nature of the transition. Similarly, there is very little change in the effective electron–hole interaction at the transition near $E_C{}^+$, on the semiconductor side. The possible effects of the change in the interaction on the transition for other cases does not seem to have been investigated.[27]

Even if Hartree–Fock calculations did always lead to a second-order transition, it is not clear that these calculations would be reliable. We can see, for example, that the Hartree–Fock approximation does not correctly describe the physical situation in the immediate

---

[25] In a real crystal, of course, there would not be an infinite series of transitions, because the transitions would be suppressed as soon as the exciton-binding energy fell below $\hbar/\tau$, where τ is the scattering lifetime due to impurities.

[26] Even if these terms are omitted, however, it is still possible to obtain a first-order transition for certain band structures [cf. J. C. Kimball and L. M. Falicov, Phys. Rev. Letters **20**, 1164 (1968)].

[27] Adler and Brooks (Ref. 17) found that the change in the screening of the electron–electron interaction was important in their narrow-band model of the metal insulator transition in the vanadium oxides, and could lead to a first-order transition in that model.

vicinity of E_C^+. Whereas the Hartree–Fock approximation treats the excitons as a dilute Bose gas with repulsive interactions, we know that in fact the interaction between two excitons of proper spin configuration is sufficiently attractive to lead to the formation of a bound "exciton molecule."[28] At the very least, the value of E_C^+ will be modified by this effect and will no longer be given by simple extrapolation of the exciton binding energy E_B appropriate to the semiconductor for $E_G \gg E_C^+$.[29] The exciton–exciton attraction might also lead, however, to a first-order transition of the gas liquid type in which there is a discrete jump in the exciton density.[30]

Further evidence which raises the possibility of a first-order transition in the excitonic state is the existence of a first-order transition at the Néel temperature of chromium.[31] For models with complicated band structures, there are also opportunities for first-order transitions within the excitonic phase, associated with changes in the period or polarization of the spin density wave.[32]

D. Detecting the Excitonic State

The best method of detecting the excitonic state, or any other anomalous state at the semimetal–semiconductor transition, is by monitoring the electrical resistivity as a function of pressure and temperature. Jérome, Rice, and Kohn[11] have calculated the electrical conductivity σ of the excitonic state for an isotropic two-band model, in the limit of a very small density of scattering centers. They find that the electrical conductivity is continuous at the (second-order) transition between the distorted and nondistorted states, but that the temperature derivative $d\sigma/dT$ is infinite in the excitonic phase just below the transition. Zittartz[15] has calculated the electrical conductivity for a finite density of impurities, and finds that there is only a finite discontinuity in $d\sigma/dT$ at the transition. Of course, if there is actually a first-order transition of some kind, there will be a finite jump in the conductivity at the transition. In any case, there should be an observable resistance anomaly at the transition, with the resistance increasing as one enters the excitonic phase. An anomaly of this type has been seen in the conductivity of chromium at its Néel point.[33]

If an anomaly in the resistivity is found consistent with the phase diagram for the excitonic state, then one may attempt to verify the existence of a lattice distortion or spin-density wave by using x-ray and/or neutron diffraction. If the exciton radius is more than a few lattice constants, however, the magnitude of the distortion will be too small to be detected by this means. A spin-density wave state may be detectable by its effect on nuclear magnetic resonance.[34]

Despite the formal analogies between the excitonic state and the theories of superconductivity and of superfluidity, the excitonic state exhibits no "super" transport properties.[4,10] In particular, the thermal conductivity has been calculated by Zittartz,[16] in a two-band isotropic model, and has been verified to be quite finite.

V. "CRYSTALLIZED EXCITONIUM"

Regardless of whether the transition to the excitonic state is first or second order, there are some values of the model parameters (i.e., of the ratio of electron to hole mass, number of electron and hole pockets, and the degree of anisotropy) for which the simple excitonic phase does not occur at all at the semimetal–semiconductor transition.

Consider a two-band model in which the hole mass is very much heavier than the electron mass, and in which there are small deviations from isotropy. In this case, for E_G close to E_C^+ the excitons do not form a Bose fluid, but condense to form a periodic array of exciton molecules precisely analogous to the array of protons and electrons in ordinary solid hydrogen. (When the hole mass is comparable to the electron mass, such a "solid" array would be unstable because of zero point motion.) This state is clearly insulating at 0°K, and is denoted by the symbol A in the phase diagram, Fig. 5.

If the unperturbed energy gap E_G is decreased further, the density of electrons and holes will increase, and a phase transition will occur to a state where there is a Wigner crystallization of the holes in a more or less uniform background of electrons.[35] This state, denoted by B in Fig. 5, will be conducting, and will be analogous to the metallic state of hydrogen which has been predicted to occur at pressures greater than 10^6 atm.[36] As E_G decreases still further, the density of holes will increase until the Wigner array of holes "melts," and the crystal becomes a normal semimetal (region 2) with well-defined Fermi surfaces of electrons and holes. In principle, if the electron and hole

---

[28] J. R. Haynes, Phys. Rev. Letters **17**, 860 (1966).

[29] Actually, it was shown in Ref. 4 that even in the Hartree–Fock approximation, there will be a deviation of E_C^+ from the extrapolated value of E_B, when the singlet–triplet splitting terms, beyond the dominant term approximation, are taken into account. That deviation, however, will be proportional to the cube of the ratio of the lattice constant to the exciton radius, and hence will be nebligible in the limit of loosely bound excitons.

[30] A more detailed discussion of the possible effects of terms beyond the Hartree–Fock on our picture of the transition to the excitonic phase, may be found in Ref. 4.

[31] A. Arrott, S. A. Werner, and H. Kendrick, Phys. Rev. Letters **14**, 1022 (1965).

[32] C. Herring in *Magnetism*, G. Rado and H. Suhl, Eds. (Academic Press Inc., New York, 1966), Vol. IV.

[33] See, for example, D. B. McWhan and T. M. Rice, Phys. Rev. Letters **19**, 846 (1967).

[34] A spin-density wave commensurate with the lattice will produce a splitting of the NMR line, while a noncommensurate spin-density wave will produce a broadening of the line, when a magnetic field is applied parallel to the direction of spin polarization.

[35] The possibility of such a state was suggested to us by C. Herring.

[36] E. Wigner and H. J. Huntington, J. Chem. Phys. **3**, 764 (1935).

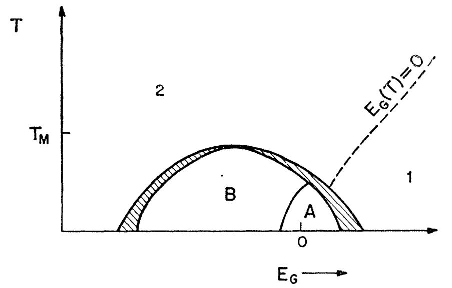

FIG. 5. Hypothetical phase diagram for a semimetal–semiconductor transition in which the hole mass is very much greater than the electron mass ($m_h/m_e \gtrsim 100$). Regions 1 and 2 are the nondistorted semiconductor and semimetal. Region A is a distorted semiconductor characterized by a crystallized array of exciton molecules. Region B is a distorted semimetal characterized by a Wigner array of holes in a conducting background of electrons. The shaded areas are unstable regions of first-order phase transitions. The nature of the transition from A to B is not known.

band were perfectly isotropic, one could have an excitonic state at very low temperatures on the semimetal side of region B; even a very slight amount of anisotropy would destroy this state, however.

The transitions between states A and 1, and between B and 2 are expected to be first order, since they are analogous to a solid–liquid transition.[37] The nature of the transition between A and B is not known. The new states A and B are similar to an excitonic state in that they are characterized by a change in the crystal periodicity caused by a charge-density oscillation. The period of this oscillation in states A and B is determined by the distance between electrons or holes on the crystallized array, and will be of the order of the Bohr radius of an exciton. By contrast, the period of oscillation in the excitonic distorted state will always be of the order of twice the lattice constant. The two kinds of states are thus distinguishable theoretically only for loosely bound excitons.

The states A and B can exist only if the ratio between the hole and electron masses is greater than some critical value R_c. In order for the state B to exist, we expect that the screening length of the electron gas must be greater than the distance between neighboring holes on the Wigner array, when this latter distance is at the critical value for melting of the array.[38] In this manner we estimate $R_c \approx 100$. For this value of the mass ratio we find that the transition between states (B) and (2) occurs at $E_G \approx -50E_B$, whereas the transition between states (A) and (1) remains near $E_G = E_B$.

VI. CHOICE OF MATERIALS

In order to examine experimentally the possible anomalies at the semimetal–semiconductor transition, we must first find a material which undergoes such a transition under applied pressure or anisotropic strain. When studied at high temperatures or in impure samples, the material must exhibit a gradual change in properties from semimetallic to semiconductorlike, or vice versa, without any polymorphic transition. One would then hope to purify the sample sufficiently, and bring it to low enough temperature to observe the transition to the excitonic phase, or any other anomaly that may occur.

Jérome, Rice, and Kohn[11] have suggested two classes of materials as possible candidates. These are the group V semimetals (Bi, Sb, As) and the divalent fcc metals (Ca, Sr, Yb). In the first of these classes only bismuth has been studied extensively. The resistivity of bismuth as a function of temperature and pressure has been studied by Souers and Jura,[39] Jaggi,[40] and Balla and Brandt.[41] At pressures above 6 kbar, Bi has a negative temperature coefficient of resistance, and at a pressure of 25 kbar, the resistance ratio $R_{4.2^\circ\mathrm{K}}/R_{300^\circ\mathrm{K}}$ is 40. However the appearance of a negative temperature coefficient of resistance does not necessarily indicate semiconducting behavior since the temperature variation in numbers of carriers for semimetals with low Fermi temperatures can also cause the resistance to drop as the temperature is raised.[41] Balla and Brandt[41] have analyzed the residual resistance as a function of pressure and they find that effective number of carriers at 4.2°K varies as

$$n_0(P) \propto |E_G|^{3/2} = \alpha(1-P/P_c)^{3/2}, \qquad (6)$$

where $|E_G|$ is the energy overlap and P_c the critical pressure at which $E_G(P_c)=0$. They find in Bi, $P_c \approx 24.5$ kbar.[42] Recent de Haas–van Alphen studies on Bi, at pressures up to 15 kbar,[43] give a value for the critical pressure in good agreement with that found by Balla and Brandt. No anomalies have been observed in Bi in the neighborhood of the critical pressure. Jérome, Rice, and Kohn[11] have estimated the size of "excitonic phase" region in Bi and find a maximum transition temperature, $T_M \approx 0.05$°K. The width in pressure is very small ($\sim 10^{-2}$ kbar) and the permitted number of charged impurities is also very small ($\approx 10^{12}$ cm^{-3}).

---

[37] An argument that a solid–liquid transition is always first order was given by L. D. Landau, Phys. Z. Sowjetunion **11**, 545 (1937). This argument depends on the isotropy of the fluid state, and can be applied to the crystallization of electrons and holes only if the anisotropy is not too great. The same argument indicates that if an Overhauser charge density wave could occur in a free electron gas, it would have to appear as a first-order transition. We are grateful to P. W. Anderson for pointing this out to us.

[38] We use the estimate $\Gamma_s \approx 25$ for the melting density of a Wigner lattice, given by H. M. Van Horn, Phys. Rev. **157**, 342 (1967).

[39] P. C. Souers and G. Jura, Science **143**, 567 (1964).

[40] R. Jaggi in *Proceedings of the Conference on Semiconductors*, 1964 (Dunod Cie., Paris, 1964).

[41] D. Balla and N. B. Brandt, Zh. Eksperim. i Teor. Fiz. **47**, 1653 (1964) [English transl.: Soviet Phys.—JETP **20**, 1111 (1965)].

[42] The polymorphic transition to the "metal" phase Bi III occurs at 45 kbar at liquid-helium temperatures.

[43] E. S. Itskevich and L. M. Fisher, Zh. Eksperim. i Teor. Fiz. **53**, 98 (1967) [English transl.: Soviet Phys.—JETP **26**, 66 (1968)].

Clearly these values impose conditions that are outside the range of experimental observation at present.

It is also possible to cause a semimetal to semiconductor transition in Bi by adding Sb.[44] In view of the effects of impurity scattering discussed above, however, we do not expect to find any anomalies at finite temperature associated with this transition.

The resistances of the fcc alkaline earth metals have been studied extensively under pressure. The resistance at room temperature rises continuously with pressure, in all three cases, until a polymorphic transition occurs to a bcc phase. In Ca, the pressures required to make large changes in the resistivity are very large (≈300 kbar),[45] and simultaneous high pressure and low temperature measurements have not been taken. McWhan[46] has studied the resistance of Sr as a function of temperature down to 4.2°K at all pressures up to 35 kbar, where Sr has a polymorphoric transition to the bcc phase. He finds a negative temperature coefficient of resistance in the vicinity of 32 kbar in some samples, but the samples with largest resistance ratios at 1 atm did not show a negative temperature coefficient of resistance. None of the samples was very pure, and it is not clear what would happen in pure Sr.

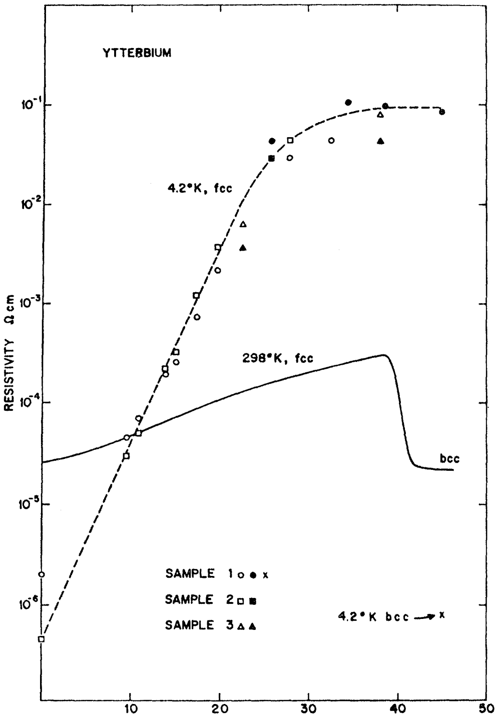

FIG. 6. The resistivity of Yb as a function of pressure at 4.2° and 2.98°K, from the data of McWhan (Ref. 46). The fcc–bcc transition occurs in the vicinity of 40 kbar. Before this transition is reached, the resistivity at 4.2°K becomes several orders of magnitude higher than that at 298°K.

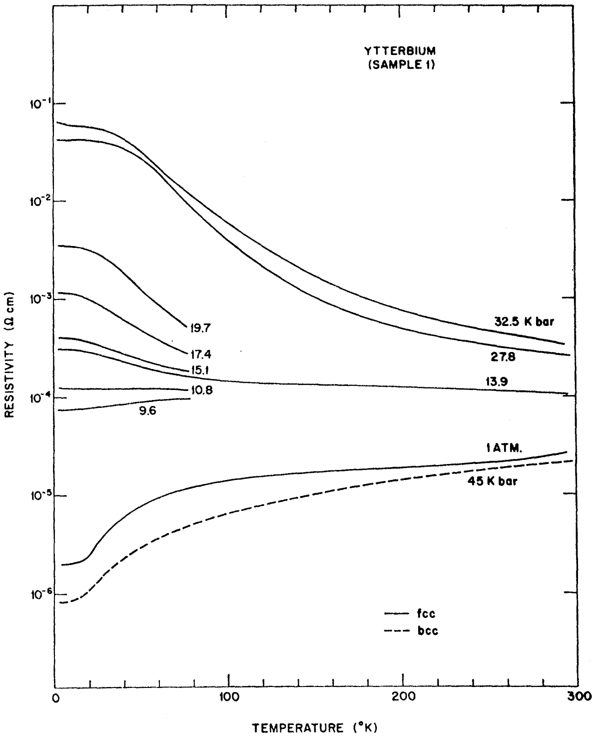

FIG. 7. The resistivity of Yb as a function of temperature at selected pressures as monitored by McWhan (Ref. 46). The dotted curve is in the high-pressure bcc phase and the solid curves are in the fcc phase. All the data refer to sample 1 of Fig. 2.

The behavior of the resistance of Yb under pressure has been studied down to 77°K by Souers and Jura,[47] and recently from 300° to 4.2°K by McWhan.[46] In Fig. 6, the resistivity of Yb at 300° and at 4.2°K, as measured by McWhan, is shown as a function of pressure. At pressures in the range 30–40 kbar, the resistivity at low temperatures reaches a value of 0.1 Ω cm, a value typical of a highly doped semiconductor.[48] This resistivity is 10^5 times the residual resistance at 1 atm and 300 times the value at 300°K at 30 kbar. In this same sample, at 1 atm, $R_{4.2°\mathrm{K}}/R_{300°\mathrm{K}}$ was $\approx\frac{1}{50}$.

In Fig. 7 the resistance is plotted as a function of temperature at selected pressures. The temperature coefficient of resistance becomes negative around 11 kbar. As we have seen from the example of bismuth, however, a negative temperature coefficient of resistance is not a guarantee of a semiconductor. It is not clear, at present, where (if at all) the semimetal–semiconductor transition would occur for pure Yb at 0°K. No anomalies have been observed in the resistivity

---

[44] D. Shoenberg and M. Z. Uddin, Proc. Roy. Soc. (London) **A156**, 687 (1936); E. S. Itskevich and L. M. Fisher, Zh. Eksperim. i Teor. Fiz. Pis'ma **6**, 748 (1967) [English transl.: JETP Letters **6**, 219 (1967)].

[45] H. G. Drickhamer, in *Solid State Physics*, F. Seitz and D. Turnbull, Eds. (Academic Press Inc., New York, 1965), Vol. 17.

[46] D. B. McWhan, T. M. Rice, and P. H. Schmidt (to be published).

[47] P. C. Souers and G. Jura, Science **140**, 481 (1963).

[48] H. Fritzsche, J. Phys. Chem. Solids **6**, 69 (1959).

which could be associated with the semimetal–semiconductor transition.

Interpretation of resistance measurements is always complicated by the difficulty of separating changes in the carrier concentration from changes in mobilities. Measurements of the de Haas–van Alphen effect, and/or Hall effect under pressure would be very helpful in clarifying the situation.

If further measurements confirm the existence of a semimetal–semiconductor transition in Yb, it may be possible to overcome the difficulties of sample purity, of pressure control and uniformity, and of low temperatures, and observe an excitonic phase. Jérome, Rice, and Kohn[11] have estimated an exciton binding energy of 13°K for Yb, using a dielectric constant estimated from the optical data of Müller,[49] taken at 1 atm. However, there are large uncertainties due to lack of knowledge of the band structure, its pressure dependence, and the pressure dependence of the dielectric constant, so that this estimate may be off by an order of magnitude.

To the best of the authors' knowledge, no band-structure calculations have ever been carried out for Yb. Vasvari, Animalu, and Heine[50] have carried out band-structure calculations for Ca and Sr, however, and we expect Yb to be qualitatively similar. They find that in the fcc structure, these materials, without spin-orbit coupling, do not become semiconductors but rather remain semimetals because of a line of degeneracy between the conduction and valence bands. The inclusion of spin–orbit coupling lifts this degeneracy, and for sufficiently strong spin–orbit coupling, leads to a semiconductor. We expect the spin–orbit gap to be ~0.1 eV in Yb, whereas it is much smaller in Ca and Sr.

Another possible candidate for an excitonic state is iodine,[45] which appears to have a continuous metal to insulator transition in the neighborhood of 160 kbar, but little is known of the band structure and other relevant properties of this material.

VII. DIRECT-BAND-GAP CASE

Thus far we have considered only the case of a semiconductor with an indirect band gap. A semiconductor can also have a *direct* band gap go to zero under pressure provided that the gap occurs at a symmetry point in the zone, and the wave functions of the conduction and valence bands have different symmetries at this point.[51] Three cases may be distinguished:

(1) The valence and conduction bands belong to different irreducible representations such that the interband matrix element of the momentum operator vanishes at the symmetry point. In this case the dielectric constant and the effective masses remain finite as the energy gap goes to zero, and an excitonic distorted state is predicted at low temperatures, just as in the indirect gap case.

(2) The valence and conduction bands belong to different irreducible representations such that the interband matrix element of the momentum operator does not vanish at the symmetry point. In the case of an *s*-like and a *p*-like band with cubic symmetry, it is easy to show that the effective masses go to zero as the energy gap goes to zero, and the exciton binding energy remains smaller than the gap. No excitonic phase occurs here, and there need not be any singularity at all in the physical properties at any finite temperature. We believe this conclusion also holds for other symmetries.

(3) The valence and conduction band states are different members of the same irreducible representation at the symmetry point: here the degeneracy of the valence band maximum and conduction minimum are required by symmetry and are not just an accidental degeneracy at one particular pressure. A band structure of this form is in fact believed to occur in gray tin.[52,53] The situation in this case is very complicated, and it is not clear at present whether or not there should be a transition at low temperatures to a distorted phase.[54]

Let us consider explicitly the case of gray tin, whose band structure is shown schematically in Fig. 8. In the Hartree approximation, the effective masses are finite

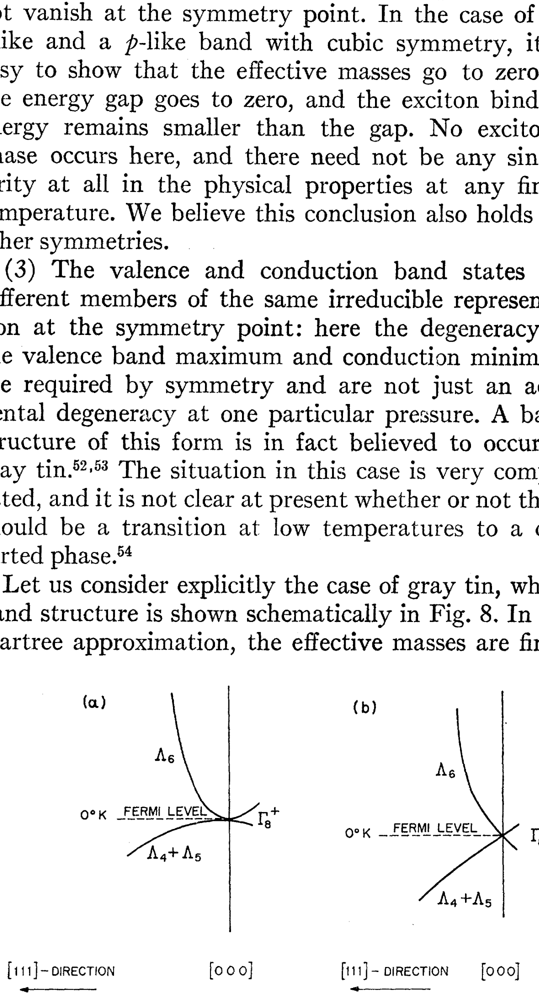

FIG. 8. Schematic band structure of (nondistorted) grey tin near zone center at 0°K in (a) the Hartree approximation and (b) the (unscreened) Hartree–Fock approximation. The section shown is along a [111] direction, but sections along other directions are qualitatively similar. The exchange contribution changes the energy bands from quadratic in k to linear in k, in the limit $k\rightarrow 0$.

---

[49] W. E. Müller, Phys. Letters **17**, 82 (1965); and Phys. Kondens. Materie **6**, 243 (1967).

[50] B. Vasvari, A. O. E. Animalu, and V. Heine, Phys. Rev. **154**, 535 (1967).

[51] If the valence and conduction states have the same symmetry then there is zero probability for the energy levels actually to cross.

[52] S. Groves and W. Paul, Phys. Rev. Letters **11**, 194 (1963); and in *Proceedings of the 7th International Conference on Semiconductors, Paris*, 1964 (Dunod Cie., Paris, 1964), p. 41.

[53] The gray tin problem was drawn to the authors' attention by W. Paul.

[54] The complications occur because the interband matrix element of the density operator

$$\langle \mathbf{k}, \text{valence band} \mid \rho_{\mathbf{q}} \mid \mathbf{k}+\mathbf{q}, \text{conduction band} \rangle$$

does not go to zero, as $\mathbf{q}\rightarrow 0$, if $\mathbf{k}$ remains of order q and is not parallel to $\mathbf{q}$.

at $\mathbf{k}=0$, and the band energies are quadratic in k near the origin. When we compute the exchange contribution to the Hartree–Fock energies, however, we find

$$E_{\mathbf{k}}-E_0 \approx \pm (3\pi e^2/32\epsilon)k + o(k^2), \qquad (10)$$

where the $(+)$ sign refers to the conduction band and the $(-)$ sign to the valence band.[55] In Eq. (10), ϵ is a "background dielectric constant," and we have assumed that the interactions between electrons in the valence or conduction bands behave like $e^2/\epsilon r$ for large separations r.

The linear term in the conduction band energy is equal to or larger than the quadratic term if

$$k \lesssim \tfrac{2}{3}(a_H)^{-1}, \qquad (11)$$

where a_H is the nominal Bohr radius for the conduction band:

$$a_H \equiv \hbar^2\epsilon/m_e e^2. \qquad (12)$$

If one forms a wave packet of large radius r from an electron and a hole, one finds in the Hartree–Fock approximation that both the kinetic energy (which is positive) and the potential energy (which is negative) are proportional to $e^2/\epsilon r$. Thus one cannot say without a more detailed calculation whether a bound state is formed, and whether the undistorted state is therefore unstable with respect to exciton formation at low temperatures.[56]

Further complications will arise if one goes beyond the simple Hartree–Fock approximation by including the effects of screening. The dielectric constant is infinite at 0°K, where one includes the effects of virtual transitions between the valence band and conduction band near $k=0$. If the 0°K static dielectric function $\epsilon(q)$ is calculated in the Random Phase Approximation,[57] one finds in the limit $q \rightarrow 0$,

$$\epsilon(q) \approx \epsilon + C'(\mu e^2/\hbar^2)\cdot q^{-1}, \qquad (13)$$

where $C' \sim \frac{3}{2}\pi$ for $m_h/m_e \gg 1$, as is the case for gray tin. In Eq. (13), ϵ is again the "background" dielectric constant, arising from all transitions other than those between the valence band and conduction band near $\mathbf{k}=0$. The divergent term in (13) becomes larger than the background term for $q \lesssim C'(a_H)^{-1}$.

In view of the above complications in the nondistorted phase, it is not clear whether or not an excitonic distorted phase will occur in pure gray tin at low temperatures. If an excitonic phase occurs, the resulting gap will remove the singularities at $\mathbf{k}=0$. Sherrington and Kohn[58] have examined some of the properties of the distorted phase that may occur.

Regardless of whether there is an actual phase transition, we would expect to see interesting many-body effects on the transport and other properties of gray tin at low temperatures in very pure samples. For example, at sufficiently low temperatures a linear energy term in the normal state would cause the number of free carriers to deviate markedly from the $T^{3/2}$ behavior predicted for parabolic bands in a pure crystal.[59]

ACKNOWLEDGMENT

We are grateful to D. B. McWhan for enlightening discussions, and for granting us permission to use his results prior to their publication.

Discussion of Halperin and Rice's Paper

J. M. Honig (Purdue University): If I understand your model correctly, you might anticipate some anomalies in thermal conductivity. Can you make a statement about this matter?

B. I. Halperin: Yes. It was proposed by some of the Russian authors that there would be an infinite thermal conductivity in this state. That is not correct despite the fact that there are analogies with superconductivity and superfluidity. There would be no supertransport properties of any kind in this state. In fact, Zittartz has explicitly calculated the thermal conductivity in a certain excitonic model and, of course, it turns out to be finite.

J. M. Honig: So you don't think that measurements of thermal conductivity represent a particularly good experimental tool for looking for such states?

B. I. Halperin: I wouldn't think so. I'm sure there would be some kind of discontinuity in the temperature derivatives of

---

[55] The details of this calculation are presented in another place. For simplicity we have neglected the possible cubic anisotropy of the energy bands in gray tin. The singular behavior of the Hartree–Fock energies near $k=0$ arises because of the sharp cutoff of the occupation numbers at the Fermi level and because of the long-range behavior of the Coulomb potential. In this respect it is similar to the singular behavior $(d\epsilon/dk=\infty)$ of the one-electron energies at the Fermi surface of a metal, in the Hartree–Fock approximation. The singularity in the metal case is generally considered to be spurious, because it is removed when the Coulomb exchange potential is screened by the dielectric function. If the exchange potential in gray tin is screened by the static dielectric function of Eq. (13), then the linear term in Eq. (10) disappears, but a singularity in the second derivative of E_k at the origin remains. At finite temperatures, the singularity in E_k is rounded off even in the Hartree–Fock approximation, because the occupation numbers are no longer discontinuous at the Fermi surface.

[56] The usual argument that a coulomb potential always has a bound state depends on the kinetic energy being proportional to $1/r^2$.

[57] In this approximation, the polarization "bubble" is calculated using the *Hartree* energies for the electron and hole, and electron–hole interactions ("ladder rungs") are ignored.

Note added in proof. This result has been obtained independently by L. Liu and D. Brust, Phys. Rev. Letters **20**, 65 (1968), and by D. Sherrington and W. Kohn (Ref. 58).

[58] D. Sherrington and W. Kohn, Rev. Mod. Phys. **40**, 767 (1968), following paper.

[59] In order for the linear term to be important in the conduction band, the number of electrons in the conduction band must be small enough so that the Fermi level is less than the nominal exciton binding energy E_B, calculated from Eq. (1). The value of ϵ is rather uncertain. Sherrington and Kohn estimate $E_B=5$°K, using the values $\mu=0.02m$, $\epsilon=24$. The electron concentration which gives a Fermi level of 5°K is about $10^{14}/\text{cm}^3$, a concentration which should be reached at a temperature $T\approx 1$°K in pure gray tin. Screening of the electron–electron interaction may reduce observable many-body effects still further. Because the holes in gray tin are much heavier than the electrons, the quadratic energy term is much smaller in the valence band than in the conduction band. A linear term in the valence band could therefore be important at temperatures larger than E_B. It is relatively difficult to study the heavy carriers experimentally, however, and it is not clear whether there would be any observable effects. The effects of impurity scattering on the linear energy term have not been studied, but it seems most likely that the linear term will be destroyed by very small concentrations of scatterers.

the thermal conductivity at the phase boundary, but I would imagine that it is harder to do that sort of measurement than an electrical resistance measurement.

R. W. Keyes (IBM): Electrical noise should be a useful tool for looking at these things. When one is at the border of some kind of transition, there is an additional possibility for fluctuations and one might expect that noise voltages would be increased.

B. I. Halperin: That might be a possibility.

M. A. Lampert (Princeton University): It seems to me that your discussion is incomplete and if so your phase diagrams would also be incomplete on the following basis. The excitons are not interacting yet, in fact, two excitons are unstable against the formation of an excitonic molecule. It is particularly easy if the hole mass is somewhat bigger than the electron mass, by the analog of two hydrogen atoms forming a hydrogen molecule. You have left that out of your discussion completely. I thought you were going to get to it when you came to the hydrogen atom metallic phase, but you didn't mention it there either. Could you comment on this?

B. I. Halperin: Actually this was discussed in a paper that Rice and I have written [B. I. Halperin and T. M. Rice in *Solid State Physics*, F. Seitz, D. Turnbull, and H. Ehrenreich, Eds. (Academic Press Inc., New York, 1968), Vol. 21] which will be coming out shortly. We actually considered this, and this is why I mumbled something about the phase diagram not being quite right near the point E_c^+. The interactions cannot be neglected. They can change the transition from second order to first order or they could have various other effects. They will certainly displace the position at which the transition will occur: it will not occur exactly at $E_G=E_B$. When E_G is substantially smaller than E_B, however, there is a rather large density of excitons. We believe that the excitonic molecules will then dissociate, because the binding energy of an exciton molecule is much smaller than the binding energy of an exciton, if the electron and hole masses are not too different. We believe that there will then be a Bose condensate of unpaired excitons, although we can't prove that rigorously.

G. J. Hyland (University of Liverpool): You draw the analogy between the phase transition in the system of noninteracting excitons and that which occurs in an ideal Bose gas; there is, however, a difference. You claim that in the case of excitons the transition is of second order—the condensation of an ideal Bose gas is, however, a first-order transition. Have you investigated fluctuations at all? As you know, infinite fluctuations are the precursors of a second-order phase transition.

B. I. Halperin: No. Any calculations that have been done have been restricted to Hartree–Fock. We have used our knowledge of the way in which the Hartree–Fock calculations relate to actuality in other transitions; however, in forming our qualitative picture of the transition to the excitonic state. I don't understand your comment that the Bose transition is a first-order transition in the noninteracting case. If you have a repulsive gas it will certainly be a second-order transition.

W. Kohn (University of California, San Diego): Since we are talking about Bose transitions, and it is quite useful in some ways to compare this phenomenon with Bose transitions, I do want to make a remark. We have here formally a situation which is analogous to a Bose transition, but I want to emphasize that the analogy is *formal* and it would be a mistake to look for anything like the superfluid properties that one finds in a real Bose transition. It is perfectly all right to use the terminology "Bose condensate of excitons," if you happen to like that, but this is the same kind of Bose condensate that, let's say, a NaCl crystal is, which you can also look at, if you wish, from the same standpoint. So, if you think of this as a Bose condensate, you should have in the back of your mind "like NaCl."

R. W. Keyes: In an anisotropic semiconductor where the minima of the conduction band are at the zone edge, for example, one has really a multivalley semiconductor. In the case of a bound state like a donor or an exciton, the lowest state of the exciton would be more complicated than simply one state from a valley and, in particular, would be a mixture of states from many valleys. That entirely changes the effect that perturbations can have on such a state because, in addition to just perturbing the energy of the valley, it can change the mixture of the states between the different valleys. Is that taken into account in your calculations, and what sort of effects does it have on the excitonic insulator?

B. I. Halperin: The perturbations are just the splitting term that we talked about. In the simple two-band model where you have a single conduction minimum at the zone boundary, it is not a multivalley situation. In the case where you have more than one minimum, it is a multivalley situation. You have to take the splitting terms into account to decide which of the many possible kinds of states will be lowest, whether it be a single spin-density wave or a double spin-density wave, or various other possibilities. Some of these possibilities have been considered by Rice and myself in our article in *Solid State Physics*.

D. S. Lieberman (University of Illinois, Urbana): The idea of using a generalized phase diagram of temperature and other thermodynamic parameters is a very attractive one and it may be profitable to reexamine the works of Willard Gibbs and others on the relationships among the various derivatives of the free energy and their implications, the conditions on the coexistence of phases, the "phase rule," etc., to see how far one can really extend the concepts to such systems as the author proposes. Relative to this diagram and others which you have sketched, I'd like some clarification about what you mean about a "phase separation," "miscibility gap" and something about "shifting over." Does it go to a two-phase mixture with regions consisting only of semimetal or semiconductor or does the whole thing go in some way throughout the volume? The second question relates to the possible instability and fluctuations implied by your diagram. This kind of diagram with a miscibility gap is associated with the phenomenon called *spinodal decomposition* which is related to a change in sign of the second derivative of the free energy with respect to composition. This has been discussed and reviewed most completely and clearly by Cahn [J. W. Cahn, Trans. Am. Inst. Min. Metal. Engrs. **242**, 166 (1968)]. Ben-Israel [D. H. Ben-Israel, abstract in J. Metals **67A** (Jan. 1968)] has reported an extension of Cahn's model to more generalized variables. I was wondering whether such instabilities and fluctuations in other parameters have been considered because they might produce an interesting kind of structure and related properties or perhaps aid in the interpretation of some of the phenomena.

B. I. Halperin: In answer to the first question, as to whether you would get a jump or a two-phase region: that depends on the boundary conditions. At a fixed pressure you would get a jump. If you confined your sample in some way, so that its total volume was fixed, then you would get a two-phase or a domain structure.

D. S. Lieberman: A total semiconductor or total semimetal in the whole specimen?

B. I. Halperin: Yes. Under fixed pressure the whole piece would jump at once from a semiconductor to a semimetal. This is assuming there would be no strains in the samples. If you have a large sample and surface effects weren't too important, I believe that's what would happen.

We have not considered the possibility raised by your second question. Perhaps that might be an interesting one to look at.