Theory of Semiconductor-To-Metal Transitions*

DAVID ADLER†
Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts
and
Center for Materials Science and Engineering, Massachusetts Institute of Technology,
AND
HARVEY BROOKS
Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts

(Received 20 September 1966)

This paper discusses a general model for a semiconductor-to-metal transition, in which the energy gap between the valence and conduction bands decreases linearly with the number of electrons excited across the gap. It is shown that this model results in a rapid disappearance of the forbidden gap with rising temperature according to either a first-order or a second-order phase transition, depending on the magnitude of the relative change in gap with the number of excited carriers. Two possible physical models are treated in detail. In one the energy gap results from the splitting of the first Brillouin zone by an antiferromagnetic exchange interaction, and in the other it results from a crystalline-structure distortion to lower symmetry. The latter model is considered in detail in terms of the pairing of ions in a one-dimensional crystal. With these models, using plausible values of the parameters, the explicit relationship between energy gap and free-carrier concentration is estimated. The thermodynamic theory is worked out for the limiting cases of band width large and small compared to the zero-temperature gap. In the narrow-band limit it is found that the parameters of the model are such as to give a second-order transition for the antiferromagnetic case and a first-order transition in the crystalline-distortion model. Using these models, the transition temperature can be evaluated explicitly in terms of the zero-temperature gap. A number of results relating experimentally measurable quantities such as the pressure coefficient of the transition temperature and the energy gap can be derived.

I. INTRODUCTION

THE transition-metal oxides provide a striking example of the inadequacy of simple band theory when an attempt is made to predict the electrical transport properties of crystals. Most of these oxides are insulators,[1] despite the apparent presence of a partially filled $3d$ band. Where the magnetic structure of these materials has been determined, they are all antiferromagnetic; however, they are insulating both below and above the Néel temperature.

The many attempts to explain the electrical properties of these oxides can be divided into three categories depending on whether the lack of conductivity is due to antiferromagnetism, to electron-electron interactions alone, or to electron-phonon interactions possibly combined with electronic interactions. Historically, the last class was the earliest suggestion. DeBoer and Verwey,[1] who pioneered the experimental work on these materials, assumed that a high potential barrier exists between any two transition-metal ions in the crystal. If the lifetime of an excited state in which an electron can tunnel through the barrier is much shorter than the tunneling time, the electrons can be considered to be localized. A mechanism for such a barrier emerged from the work of Landau[2] and of Gurney and Mott,[3] who showed that in polar crystals it is possible for an electron localized around an ion to be trapped in the potential well which results from the lattice polarization due to the presence of the electron. Yamashita and Kurosawa[4] worked out the theory of this model in detail, and were able to demonstrate the localized self-trapped state provided a self-consistent solution. Holstein[5] considered the case where the coupling between the electrons and the optical phonons is sufficiently strong that polaron states are appropriate. When the electronic band width is small, as must be the situation for $3d$ bands of transition metal oxides, the polarons are "small"; in this case, Holstein found that for temperatures greater than about half the Debye temperature, the quasiparticle band width has effectively shrunk to zero, and the polarons are essentially localized, so that such conduction as occurs will be by a diffusive hopping between adjacent sites rather than by the correlated motion described in band theory. Toyazawa[6] showed that the electrons could also be trapped by means of the short-range interactions with acoustical phonons, provided the coupling was sufficiently strong.

In all of the above-mentioned theories, except in the polaron band regime, conductivity occurs only by means of electronic hopping from site to site. Verwey[7] attributed the activation energy to ionization of the cations. Heikes and Johnson[8] noted that the measured

* Research supported by the Advanced Research Projects Agency.

† Present address: Center for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts.

[1] J. H. De Boer and E. J. W. Verwey, Proc. Phys. Soc. (London) **A49**, 59 (1937).

[2] L. D. Landau, Physik. Z. Sowjetunion **3**, 664 (1933).

[3] R. W. Gurney and N. F. Mott, Proc. Phys. Soc. (London) **A49**, 32 (1937).

[4] J. Yamashita and T. Kurosawa, J. Phys. Soc. Japan **15**, 802, 1211 (1960).

[5] T. Holstein, Ann. Phys. (N. Y.) **8**, 325, 343 (1959).

[6] Y. Toyazawa, Progr. Theoret. Phys. (Kyoto) **26**, 29 (1961).

[7] E. J. W. Verwey, P. J. Haaijman, F. C. Romeijn, and G. W. van Costerhout, Philips Res. Rept. **5**, 173 (1950).

[8] R. R. Heikes and W. D. Johnston, J. Chem. Phys. **26**, 582 (1957).

impurity ionization energies are too small to account for the observed activation energies in these materials. They ascribed the conduction to a diffusion of electrons from ion to ion, thus associating the activation energy with the mobility rather than the carrier concentration. In a quantitative version of this theory, Yamashita and Kurosawa[4] showed that the activation energy arises from Landau trapping.[2]

Mott[9] suggested that the localization of electrons need not arise from lattice interactions, but was more likely to be due to the Coulomb repulsion between two $3d$ electrons of opposite spin on the same ion, a situation analogous to the low-density crystallization of an electron gas, first noted by Wigner.[10] Mott proposed that a critical distance between the transition metal ions exists, above which Heitler-London theory is a more reasonable point of departure than Bloch theory. Anderson[11] pointed out that if the correlation energy was much larger than the band width, it would outweigh the decrease in average kinetic energy resulting from band formation, and the electrons would remain localized. Kohn[12] considered an array of hydrogenic atoms around a circular wire and showed that for sufficiently large separation and number of atoms the low-lying many-electron states are nonconducting.

Recently, much work has been carried out based on a quantitative description of Mott's original hypothesis, which has come to be known as the short-range, one-band model. The problem is to investigate the system described by the Hamiltonian

$$H=\sum_k \epsilon_k(n_{k\uparrow}+n_{k\downarrow})+I\sum n_{i\uparrow}n_{i\downarrow},$$

where $n_{k\sigma}$ is the number operator for an electron in the state k, σ, and I is the average intraionic Coulomb repulsion. In this model, interionic interactions are neglected. Hubbard[13] showed that when I is greater than a critical multiple of the band width an s band is split into two subbands or a d band into 10 subbands. Gutzwiller[14] and Kemeny[15] have also investigated the same Hamiltonian and have obtained results essentially similar to those of Hubbard. In these models, the splitting into two bands may be visualized as arising directly from interaction of electrons on a common ion.

A different approach to the same problem was suggested by Slater,[16] who used a two-band model. In this theory, it is antiferromagnetic ordering which leads to the insulating nature of the ground state; the doubly periodic exchange potential splits the first Brillouin zone in half and introduces an energy separation at the surface between the new subzones. A quantitative version of the Slater two-band model has been given by Des Cloiseaux.[17] This model requires that the exchange splitting be sufficiently large compared to the band width so that an energy gap opens up along the whole face of the new zone boundary. However, it is not clear why the insulating property should not disappear above the Néel temperature.

The lack of electrical conductivity is not the only enigma provided by the transition metal oxides. Another subclass consists of several oxides of titanium and vanadium, which are semiconducting at low temperatures, but undergo a transition to a metallic state above a critical temperature T_0. The electrical properties of these oxides were thoroughly studied by Morin,[18] who found that all of the lower oxides of titanium and vanadium with the exception of TiO exhibit this behavior.

Most of the theories presented to explain the lack of conductivity in the insulating oxides can be extended to give an insulator-to-metal transition. The theories which depend on electron-phonon coupling are least satisfactory in this respect. The polaron model of Holstein[5] results in a transition from a polaron band regime to a hopping regime as the temperature is raised, but the high-temperature state is not that of a metal. The acoustical phonon self-trapping model of Toyazawa[6] exhibits a sharp transition from a band picture to a localized picture as the strength of the interaction is increased, but there is reason to believe that the electron-phonon coupling constant (which depends inversely on the lattice parameter) decreases rather than increases with temperature.

The theories in which only electron-electron interactions are responsible for the nonconductivity exhibit a transition from an insulating to a metallic state as the correlation energy decreases. The original proposal of Mott[9] contained qualitative reasons why the transition should be a sharp one, both as a function of temperature and as a function of electronic density.

Kohn[12] quantitatively investigated Mott's theory by considering a ferromagnetic simple cubic array of hydrogen atoms. If the potential between a spin-up hole and a spin-down electron is a Coulomb attraction, an insulating spin-wave state is lowest; however, for a delta-function interaction, a critical strength exists, below which only a continuum of states, characteristic of a metal, is present.

Hubbard,[13] using the short-range, one-band model, found that at a critical ratio of the band width E_b to the intraionic Coulomb energy I, the energy gap due to electronic correlations has shrunk to zero, and an insulator-to-metal transition occurs. This may be re-

[9] N. F. Mott, Proc. Phys. Soc. (London) **A62**, 416 (1949); Nuovo Cimento Suppl. **7**, 312 (1958); Phil. Mag. **6**, 287 (1961).
[10] E. Wigner, Trans. Faraday Soc. **34**, 678 (1938).
[11] P. W. Anderson, in *Solid State Physics*, edited by F. Seitz and D. Turnbull (Academic Press Inc., New York, 1963), Vol. 14.
[12] W. Kohn, Phys. Rev. **133**, A171 (1964).
[13] J. Hubbard, Proc. Roy. Soc. (London) **A276**, 238 (1963); **A281**, 401 (1964).
[14] M. C. Gutzwiller, Phys. Rev. **137**, A1726 (1965).
[15] G. Kemeny, Ann. Phys. (N. Y.) **32**, 69, 404 (1965).
[16] J. C. Slater, Phys. Rev. **82**, 538 (1951).
[17] J. Des Cloiseaux, J. Phys. Radium **20**, 606 (1960); **20**, 751 (1960).
[18] F. J. Morin, Phys. Rev. Letters **3**, 34 (1959).

garded as another quantitative treatment of the Mott transition.[9] Hubbard finds that the energy gap shrinks continuously as the ratio of E_b to I increases, so that slightly below the critical value the material has an infinitesimal band gap, while just above the transition point the density of states at the Fermi surface is negligibly small, so that the material is a poor metal. This result is contrary to Mott's hypothesis of a sharp increase in the number of free carriers, although it is possible that Hubbard's neglect of interactions of electrons on different ions camouflaged the nature of the transition. Kemeny,[15] for the case of an extremely tightly bound simple cubic hydrogenic lattice at zero temperature, found that all electron-hole pairs making up the crystallized low-density state disassociate sharply into a metallic state above a critical density, in agreement with Mott's conjecture.

Morin[18] attempted to explain the conductivity discontinuities by adapting the two-band model of Slater.[16] If the materials were semiconducting because of a band splitting arising from antiferromagnetism, then a transition to a metallic state would be expected at the Néel temperature. Callaway[19] made the Slater-Morin theory a little more quantitative by studying the energy band structure of a body-centered cubic antiferromagnet. Considering only the first Fourier component of the exchange potential, a major simplication, he found that an insulating state exists whenever an interaction parameter (proportional to the effective mass, to the strength of the exchange potential, and to the square of the lattice constant) is sufficiently large. Neither Callaway nor Morin discussed the nature of the semiconductor-to-metal transition beyond noting that the band gap should disappear at the Néel temperature.

There are difficulties with the Slater-Morin theory, aside from the fact that it has never been quantitatively applied to the oxides of titanium and vanadium. Firstly, the existence of antiferromagnetism has been demonestratd only in Ti_2O_3,[20] and even in that material the antiferromagnetic moment is extremely small. Secondly, no model for the structure of the degenerate $3d$ bands has been presented which explains how Ti_2O_3, V_2O_3, or VO_2 can be insulating at $T=0$ even with the antiferromagnetic splitting. It is conceivable that the large amount of short-range order present up to two or three times the Néel temperature is sufficient to maintain the effective double periodicity felt by the slowly moving $3d$ electrons, and thus maintain an energy gap.

A somewhat different explanation of the transition was given by Goodenough,[21] who suggested that, because of direct cation-cation interactions, all would-be conduction electrons could be trapped in homopolar bonds at low temperatures. Goodenough applied his hypothesis to the oxides of titanium and vanadium, and was able to account for many of the previously unexplained symmetry changes. However, the theory does not lend itself readily to quantitative investigation.

In the present work, we shall present a model for semiconductor-to-metal transitions which can be applied to the vandium and titantium oxides and can be tested experimentally. We shall use an itinerant electron picture, and assume that the nonconducting state of these materials is that of a normal semiconductor, a filled valence band being separated from an empty conduction band by an energy gap. In Sec. II, we shall show how such a gap can arise in the transition metal oxides from antiferromagnetism or from a crystalline structure distortion to lower symmetry. It will be demonstrated that in these two cases the energy gap will shrink as carriers are excited across it, and the decrease in gap will be quantitatively estimated. In Sec. III, the theory of conductivity will be worked out in two limits, the effective-mass approximation and the limit of narrow bands. We shall demonstrate the existence of a semiconductor-to-metal transition, which can be either first or second order, and we shall calculate the transition temperature in terms of observable quantities.

II. DEPENDENCE OF ENERGY GAP ON CARRIER CONCENTRATION

A. General Hypothesis

Consider an intrinsic semiconductor for which the top of the valence band is separated from the bottom of the conduction band by an energy gap E_g. In general, E_g depends on the concentration of carriers in the conduction band n and on the temperature T:

$$E_g=E_g(n,T).$$

At low temperatures, the concentration of carriers is also small, and we can write

$$E_g=E_{g0}-\bar{\alpha}T-\bar{\beta}n,$$

where E_{g0} is the gap at $T=0$, $\bar{\alpha}\equiv-(\partial E_g/\partial T)_n$, and $\bar{\beta}\equiv-(\partial E_g/\partial n)_T$. Although the term linear in T is responsible for the major part of the decrease in band gap at very low temperatures, it does not contribute to the semiconductor-to-metal transition, and therefore will be dropped. Only a small error is introduced into the calculation by dropping it.[22] We are left with

$$E_g=E_{g0}-\bar{\beta}n, \tag{2.1}$$

which is our fundamental relation. The remainder of this section will be devoted to demonstrating the applicability of Eq. (2.1) in two particular situations, where the energy gap is due to antiferromagnetism and where

[19] J. Callaway, in *Proceedings of the International Conference on the Physics of Semiconductors, Exeter, 1962*, (The Institute of Physics and The Physical Society, London, 1962), p. 582.

[20] S. C. Abrahams, Phys. Rev. **130**, 2230 (1963).

[21] J. B. Goodenough, Phys. Rev. **117**, 1442 (1960); **120**, 67 (1960); *Magnetism and the Chemical Bond* (Interscience Publishers, New York, 1963).

[22] D. Adler, Gordon McKay Laboratory, Harvard University, Technical Report No. ARPA-12, 1964 (unpublished).

the gap arises from a crystalline structure distortion to lower symmetry. We shall evaluate $\bar{\beta}$ for both of these cases, and show that Eq. (2.1) remains valid as n becomes relatively large.

B. Thermodynamic Argument

In this section, we shall present a thermodynamic calculation of the change in energy gap of a semiconductor with the concentration of excited carriers. This will provide us with a general expression for $\bar{\beta}$ in Eq. (2.1). The first part of this argument follows closely a recent paper by Figielski.[23]

The differential form of the Gibbs free energy for a system where the number of particles may vary is

$$dG=-SdT+VdP+\sum_j \mu_j dN_j, \tag{2.2}$$

where N_j is the number of particles in the jth phase and μ_j is the chemical potential of the jth phase. Treating electrons in the valence and conduction bands as different phases and ignoring inner orbitals, we can write

$$\sum_j \mu_j dN_j=\mu_c dN_c+\mu_v dN_v. \tag{2.3}$$

For an intrinsic semiconductor, $N_v+N_c=$ constant, and (2.3) becomes

$$\sum_j \mu_j dN_j=(\mu_c-\mu_v)dN, \tag{2.4}$$

where N is the number of carriers.

Since $\partial^2G/\partial P\partial N=\partial^2G/\partial N\partial P$, (2.2) and (2.4) give

$$\left(\frac{\partial V}{\partial N}\right)_{P,T}=\frac{1}{V}\left(\frac{\partial V}{\partial n}\right)_{P,T}=\left(\frac{\partial \mu_c}{\partial P}\right)_{n,T}-\left(\frac{\partial \mu_c}{\partial P}\right)_{n,T}, \tag{2.5}$$

where $n\equiv N/V$ is the concentration of carriers.

It is clear that at $T=0$, $\mu_v=E_v$, and $\mu_c=E_c$, where E_v is the energy of the top of the valence band and E_c is the energy of the bottom of the conduction band. In general, $\mu_c-\mu_v$ is the change in free energy when an electron is removed from the valence band and placed in the conduction band. Thus, $\mu_c-\mu_v$ is the free-energy gap, which we shall call E_g. It is this free-energy gap, rather than the enthalpy gap, which determines the number of intrinsic free carriers when the densities of states are determined from measured physical quantities, such as effective masses. It is well known that the free-energy gap E_g differs from the enthalpy gap, which determines the activation energy for intrinsic conductivity, by terms of the order of kT.[24] For the remainder of this paper, we shall neglect this difference between free-energy gap and enthalpy gap. Writing (2.5) in terms of E_g:

$$\frac{1}{V}\left(\frac{\partial V}{\partial n}\right)_{P,T}=\left(\frac{\partial E_c}{\partial P}\right)_{n,T}-\left(\frac{\partial E_v}{\partial P}\right)_{n,T}=\left(\frac{\partial E_g}{\partial P}\right)_{n,T}. \tag{2.6}$$

Using the thermodynamic relations

$$\left(\frac{\partial E_g}{\partial n}\right)_{P,T}=\left(\frac{\partial E_g}{\partial V}\right)_{n,T}\left(\frac{\partial V}{\partial n}\right)_{P,T}+\left(\frac{\partial E_g}{\partial n}\right)_{V,T}$$

and

$$\left(\frac{\partial E_g}{\partial V}\right)_{n,T}=-\frac{1}{\kappa V}\left(\frac{\partial E_g}{\partial P}\right)_{n,T},$$

where κ is the isothermal compressibility, we find

$$\left(\frac{\partial E_g}{\partial n}\right)_{P,T}=-\frac{1}{\kappa V}\left(\frac{\partial E_g}{\partial P}\right)_{n,T}\left(\frac{\partial V}{\partial n}\right)_{P,T}+\left(\frac{\partial E_g}{\partial n}\right)_{V,T}. \tag{2.7}$$

Thus, (2.6) and (2.7) yield

$$\left(\frac{\partial E_g}{\partial n}\right)_{P,T}=-\frac{1}{\kappa}\left(\frac{\partial E_g}{\partial P}\right)^2_{n,T}+\left(\frac{\partial E_g}{\partial n}\right)_{V,T}. \tag{2.8}$$

We express the energy gap for varying carrier concentration and pressure as

$$E_g=E_{g0}-\bar{\beta}n-\bar{\gamma}P. \tag{2.9}$$

Thus,

$$(\partial E_g/\partial P)_{n,T}=-\bar{\gamma}. \tag{2.10}$$

Substituting (2.10) in (2.8)

$$\left(\frac{\partial E_g}{\partial n}\right)_{P,T}=-\frac{\bar{\gamma}^2}{\kappa}+\left(\frac{\partial E_g}{\partial n}\right)_{V,T}. \tag{2.11}$$

But also from (2.9), we see

$$(\partial E_g/\partial n)_{P,T}=-\bar{\beta}. \tag{2.12}$$

Combining (2.11) and (2.12)

$$\bar{\beta}=\frac{\bar{\gamma}^2}{\kappa}-\left(\frac{\partial E_g}{\partial n}\right)_{V,T}. \tag{2.13}$$

This is the general thermodynamic expression for $\bar{\beta}$. The first term on the right represents the contribution to $\bar{\beta}$ resulting from changes in the volume of the crystal. This term can be evaluated easily from the experimentally measurable quantities, $\bar{\gamma}$, κ, and is always positive. The other contribution to $\bar{\beta}$ is an explicit dependence of the gap on carrier concentration at constant volume, and can have either sign.

Relations analogous to (2.13) can be derived with uniaxial stress as the intrinsic variable. One such relation, appropriate to the case where stress is applied in the direction of the c axis can be written

$$\bar{\beta}=\frac{1}{\kappa_c}\left(\frac{\partial E_g}{\partial S_c}\right)^2_{n,T,S_a,S_b}-\left(\frac{\partial E_g}{\partial n}\right)_{c,T,S_a,S_b},$$

[23] T. Figielski, Phys. Status Solidi **3**, 1876 (1963).

[24] H. Brooks, in *Advances in Electronics and Electron Physics*, edited by L. Martin (Academic Press Inc., New York, 1955), p. 121.

where S_c is the applied stress and κ_c is the linear compressibility in the c direction.

These thermodynamic arguments show that whenever there is a pressure or stress dependence of the energy gap of a semiconductor, the gap must also depend on carrier concentration. Since the pressure coefficients of the gap in V_2O_3 and VO, two of the materials with which we are especially concerned, are anomalously large, we expect a relatively large decrease of gap with carrier concentration in these materials. But this argument demonstrates only that E_g depends on the number of excited carriers. It does not say anything about the validity of Eq. (2.1) as n becomes fairly large, nor does it indicate the microscopic reasons for such a variation in energy gap. Therefore, we now turn to specific models for which we can calculate expressions for E_g as a function of n. These calculations will provide us also with expressions for $\tilde{\beta}$ which can be tested experimentally.

C. Antiferromagnetism

Consider an antiferromagnetic crystal which can be described by Bloch wave functions. Assume that the crystal is an insulator at $T=0$ because of the splitting of the first Brillouin zone by the doubly periodic exchange potential. In other words, we have an empty conduction band which begins a distance E_g above the filled valence band, with E_g being a measure of the exchange energy. We assume that the exchange splitting occurs so that the conduction and valence band edges are at the same point on the zone, so that the gap and the splitting are the same. The lower band refers to wave functions whose amplitudes are large at the sublattice positions of the electron under consideration, whereas the upper band wave functions have large amplitudes at the positions of the sublattice of opposite spin. As the temperature is increased from $T=0$, the upper band becomes thermally populated. When an electron is excited across the energy gap, the net magnetization on either sublattice decreases, and thus the gap decreases with increasing concentration of carriers. Thus, a relationship like Eq. (2.1) can be expected to hold. In this section, we shall determine for how large a value of n Eq. (2.1) remains valid and also calculate the value of $\tilde{\beta}$. We shall employ a virtual crystal approximation. Consider metallic ions with spins ordered antiferromagnetically at $T=0$. Let an ion for which the associated spin is primarily down be called type A, a primarily spin-up ion type B. Consider the sublattice of ions for which the magnetization is negative when perfect order exists. In the vicinity of an A ion, an electron with spin up sees a potential V_A^+, whereas a spin-down electron sees V_A^-. Similarly, V_B^+ and V_B^- are the potentials in the vicinity of a B ion seen by a spin-up and a spin-down electron, respectively. It is clear that $V_A^+=V_B^-$ and $V_A^-=V_B^+$. For simplicity, we shall take the case of one $3d$ electron per cation. Since N is the total density of cations, the density on each sublattice is $N/2$.

The average potential seen by an electron with spin up on the sublattice under consideration is

$$V_+'=(2/N)(n_AV_A^++n_BV_B^+)$$
$$=(2/N)(n_AV_A^++n_BV_A^-), \quad (2.14)$$

where n_A is the number of A ions on the sublattice, n_B is the number of B ions. Similarly, for a spin-down electron,

$$V_-'=(2/N)(n_AV_A^-+n_BV_B^-)$$
$$=(2/N)(n_AV_A^-+n_BV_A^+). \quad (2.15)$$

The average exchange energy is then just the difference between (2.14) and (2.15), or

$$\langle V\rangle_{ex}=\frac{2}{N}(n_A-n_B)(V_A^+-V_A^-). \quad (2.16)$$

For perfect order, $n_A=N/2$, $n_B=0$, and

$$\langle V\rangle_{ex\ T=0}=V_A^+-V_A^-.$$

For complete disorder, $n_A=n_B=N/4$, and

$$\langle V\rangle_{ex}=0.$$

Since $n_A+n_B=N/2$, and the number of intrinsic carriers is just the number of ions with spin up on both sublattices, (2.16) can be written

$$\langle V\rangle_{ex}=\langle V\rangle_{ex\ T=0}(1-2n/N). \quad (2.17)$$

If we assume that the energy gap is proportional to the average exchange energy, in the spirit of the virtual-crystal approximation, then (2.17) becomes

$$E_g=E_{g0}(1-2n/N). \quad (2.18)$$

Thus,

$$\tilde{\beta}=2E_{g0}/N. \quad (2.19)$$

One-dimensional and three-dimensional models of antiferromagnetism have also been investigated in detail,[22] and the result (2.19) is verified to within terms of the order of the square of the ratio of the band width to the exchange energy. Thus, for the narrow $3d$ bands of the transition metal oxides, Eq. (2.1) will remain valid over a relatively large range of n. In computing (2.18) it is tacitly assumed that the lattice constant continuously adjusts itself to minimize the total energy, so that (2.18) is a constant-pressure result.

D. Crystalline Structure Distortion

The existence of antiferromagnetism is not a necessary condition for the applicability of (2.1). The relation can also be shown to be appropriate when an energy gap is caused by a crystal-structure distortion to lower symmetry. This type of gap can arise from an energy gain due to chemical binding—the lower band may be thought of as a bonding band, the upper an anti-

bonding band. Excitation of an electron across the energy gap decreases the gap because the excited electron no longer contributes to the chemical binding. Thus, the situation is analogous to the case of an antiferromagnetic crystal dealt with in Sec. C. However, the evaluation of $\bar{\beta}$ is more difficult when the gap is caused by crystalline distortion. In this section, we shall calculate expressions for $\bar{\beta}$, using a simple one-dimensional model, but employing as much as possible the physical properties of the vanadium oxides to which we expect the theory to apply.

Consider a one-dimensional crystal with two cations per unit cell at zero temperature. Once again, we shall examine the case of one $3d$ electron per cation in a nondegenerate band; the case of large concentrations of $3d$ electrons in degenerate bands is entirely analogous.

In accordance with these assumptions, we place ions at positions

$$x_{j1}=\left(j+\frac{1-2\epsilon}{4}\right)a,$$

$$x_{j2}=\left(j-\frac{1-2\epsilon}{4}\right)a.$$

Here ϵ is a parameter which ranges from 0 to $\frac{1}{2}$, and reflects the amount of distortion. The crystal is semiconducting due to the extra band gap brought about by the deviation from one cation per unit cell.

To begin with, the simplest interaction we can write down is a delta-function potential

$$V(x)=V_0\sum_j\left\{\delta\left[x-\left(j+\frac{1-2\epsilon}{4}\right)a\right]+\delta\left[x-\left(j-\frac{1-2\epsilon}{4}\right)a\right]\right\}.$$

This is essentially the situation in which the Coulomb interaction is very strongly screened, which is not too far from the case where an extremely high density of free electrons exists. Schrödinger's equation may be written

$$-\frac{\hbar^2}{2m}\psi''+V_0\sum_j[\delta(x-ja)+\delta(x-ja-b)]\psi=E\psi, \quad (2.20)$$

where $b\equiv(1-2\epsilon)a/2$.

The solution to (2.20) is[22]

$$\cos ka=\cos y+2z\frac{\sin y}{y}+2z^2\frac{\sin y(1/2-\epsilon)}{y}\frac{\sin y(1/2+\epsilon)}{y}, \quad (2.21)$$

where $y^2\equiv 2ma^2E/\hbar^2$, and $z\equiv ma^2V_0/\hbar^2$.

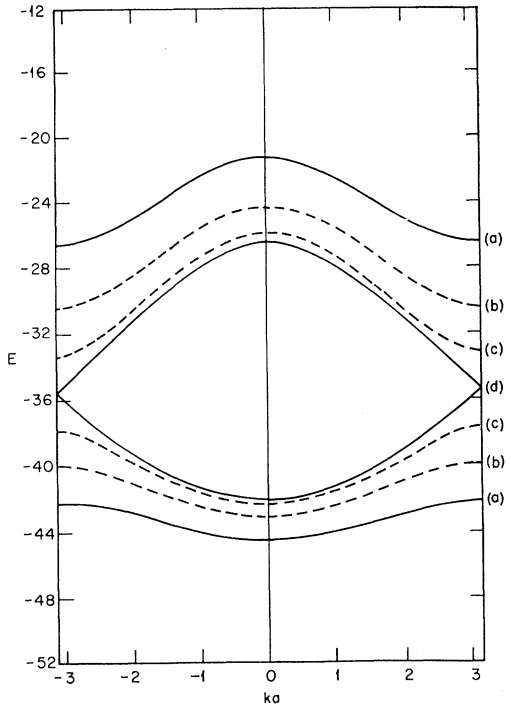

FIG. 1. Energy as a function of k: $z=-6$. (a) $\epsilon=0.15$, (b) $\epsilon=0.10$, (c) $\epsilon=0.05$, (d) $\epsilon=0$. (arbitrary units for E)

Equation (2.21) gives the energy band structure of the crystal. We still have two parameters at our discretion, the strength of the interaction z, and the amount of distortion ϵ. These parameters determine the widths of the valence and conduction bands and the size of the energy gap. As z increases in magnitude, the width of the band in the undistorted case decreases; as ϵ increases, the gap increases relative to the widths of the valence and conduction bands. As an example, the energy-band structure is given in Fig. 1 for $z=-6$ and ϵ varying from 0.15 to 0.

As can be seen, the energy gap introduced by the distortion is approximately proportional to ϵ, a result which remains valid even when ϵ becomes rather large. When z is large and thus the bands are narrow, the distortion depresses the entire lower band and elevates the entire upper band by relatively constant amounts. The distortion can be looked at as a band generalization of the Jahn-Teller effect. When the bands are extremely narrow, the gap opens up along the whole Fermi surface at once. This gap, and therefore the gain in electronic energy is proportional to ϵ. Since the loss in elastic energy is proportional to ϵ^2, the total energy will be lowered by some distortion. In the limit of infinitely narrow bands, the analysis must, of course, be equivalent to the ordinary Jahn-Teller theory.

On the other hand, when the bands are wide, a distortion only introduces a small gap at the edges of the reduced first Brillouin zone. Only a fraction, which we shall show in Sec. III is of the order of the ratio of the

energy gap to the band width, of each band is displaced. Hence the gain in electronic energy tends to be proportional to ϵ^2, and therefore cannot dominate the loss in strain energy for any finite distortion. This gives us some insight into the reason why such distortions are found, for example, in the oxides of vanadium, but not in sodium. We shall demonstrate these ideas quantitatively in Sec. III.

Returning to the model under consideration, the positions of the top of the valence and bottom of the conduction bands at $T=0$ can be expressed:

$$E_v=E_0-rE_g,$$
$$E_c=E_0+(1-r)E_g, \tag{2.22}$$

where E_0 is the common energy of the band edges in the absence of distortion, and r is a factor giving the ratio of the depression of the valence band maximum to the band gap, E_g. For sufficiently small ϵ, r always approaches the symmetric value $\frac{1}{2}$. However, as can be seen from Fig. 1, for larger ϵ, r can become significantly less than $\frac{1}{2}$. Let us write

$$r=\tfrac{1}{2}(1-\delta), \tag{2.23}$$

where δ gives the deviation of the splitting from the symmetric case. Solutions of Schrödinger's equation,[22] for both delta-function and Mathieu interactions show that the function $\delta(\epsilon)$ rises rather sharply from 0 to a constant value which depends on the strength of the interaction. For very narrow bands, the asymptotic values of δ were generally between $\frac{1}{4}$ and $\frac{1}{2}$, centering around $\frac{1}{3}$.

It is also found that the dependence of E_g on ϵ varies somewhat from strict proportionality, in the narrow band limit, as ϵ is increased. In the region of constant δ, E_g can be expressed as the linear function

$$E_g=A\epsilon-D, \tag{2.24}$$

where D is small and positive. This represents a gap which shows an upward curvature as ϵ increases. Equation (2.24) is a good approximation in any region where $E_g(\epsilon)$ has constant slope. Thus, it can also be applied in the vicinity of very small ϵ by taking $D=0$.

Using (2.22) and (2.23), the zero-temperature energy relative to that of the undistorted system is

$$E=-(NE_g/2)(1-\delta)+B\epsilon^2, \tag{2.25}$$

where the second term on the right represents the increase in elastic energy. In the region where δ is essentially constant and E_g is represented by (2.24), (2.25) becomes

$$E=(ND/2)(1-\delta)-(NA/2)(1-\delta)\epsilon+B\epsilon^2. \tag{2.26}$$

From (2.26), we can determine the equilibrium distortion as

$$\epsilon_0=(NA/4B)(1-\delta). \tag{2.27}$$

The above analysis is valid only at zero temperature and only for relatively narrow bands. As the temperature increases, electrons are excited across the gap. As in the case of antiferromagnetism, we wish to calculate the dependence of the energy gap on the concentration of carriers.

If the bands are very narrow, then for each electron excited from the valence to the conduction band, the energy is increased by E_g. Thus, at finite temperature, (2.26) must be replaced by

$$E=\frac{ND}{2}(1-\delta)-\frac{NA}{2}(1-\delta)\epsilon+nA\epsilon-nD+B\epsilon^2, \tag{2.28}$$

which gives for the equilibrium distortion

$$\epsilon=\epsilon_0\left[1-\left(\frac{2}{1-\delta}\right)\frac{n}{N}\right]. \tag{2.29}$$

As expected, the amount of distortion decreases with increasing temperature. From (2.24) and (2.29), the energy gap as a function of carrier concentration is

$$E_g=E_{g0}\left[1-\left(\frac{2}{1-\delta}\right)\left(1+\frac{D}{E_{g0}}\right)\frac{n}{N}\right], \tag{2.30}$$

where we have used the approximation that D is small compared to E_{g0}. Equation (2.30) is the relation analogous to (2.18) when the gap is due to a crystalline distortion, provided we are in a region when δ is constant and E_g is a linear function of ϵ. The value of $\tilde{\beta}$ in (2.1) is

$$\tilde{\beta}=\frac{2}{1-\delta}\frac{E_{g0}}{N}\left(1+\frac{D}{E_{g0}}\right). \tag{2.31}$$

This expression can also be used in the range where ϵ is very small, by setting $\delta=D=0$. In this region, (2.31) becomes identical to the antiferromagnetic result (2.19). This entire argument will be repeated more rigorously in Sec. III.

III. FREE ENERGY AND ELECTRICAL CONDUCTIVITY

For this section, we shall assume that it is a crystalline distortion which has brought about the energy gap. We shall write down the free energy of the system and then determine the amount of distortion ϵ, which minimizes this energy. From this, we can then find the concentration of free carriers n as a function of temperature. Since electrical conductivity for an intrinsic semiconductor can be expressed as

$$\sigma=ne(\mu_e+\mu_h), \tag{3.1}$$

where μ_e and μ_h are the electron and hole mobilities, then $n(T)$ gives the conductivity as a function of temperature for constant mobility.

Treating electrons in the valence and conduction bands as two independent phases, and ignoring inner

orbitals, the electronic free energy of the system is

$$F_{el}=n\mu_c+(N-p)\mu_v - kT\int_{-\infty}^{\infty}\ln\left[1+\exp\left(\frac{\mu_c-E}{kT}\right)\right]\rho_c(E)dE - kT\int_{-\infty}^{\infty}\ln\left[1+\exp\left(\frac{\mu_v-E}{kT}\right)\right]\rho_v(E)dE, \quad (3.2)$$

where ρ_c and ρ_v are the densities of states, μ_c and μ_v are the quasi-Fermi levels, and n and p are the concentrations of free carriers in the conduction and valence bands, respectively. To (3.2) must be added the increase in strain free energy due to the distortion. For small distortions, this may be expressed as a power series in ϵ^2:

$$F_{st}=B\epsilon^2+B'\epsilon^4+\cdots. \quad (3.3)$$

A. Narrow-Band Approximation

The oxides which exhibit semiconductor-to-metal transitions are characterized by extremely narrow $3d$ bands. For such materials, the effective masses of electrons and holes are so large that the usual parabolic band approximations are not valid except at extremely low temperatures. The physical situation is probably closer to the extreme limit of delta-function bands. Therefore, we shall analyze the free energy and conductivity first in this simple limit. We assume one $3d$ electron per cation, although any number can be treated analogously.

We take as the densities of states in the conduction and valence bands

$$\rho_c(E)=N\delta(E-E_c), \quad \rho_v(E)=N\delta(E-E_v). \quad (3.4)$$

Substituting (3.4) in (3.2), we obtain

$$F_{el}=n\mu_c+(N-p)\mu_v-NkT\ln\left[1+\exp\left(\frac{\mu_c-E_c}{kT}\right)\right] -NkT\ln\left[1+\exp\left(\frac{\mu_v-E_v}{kT}\right)\right]. \quad (3.5)$$

Minimizing (3.5) with respect the quasi-Fermi levels gives

$$\mu_c=E_c+kT\ln\frac{n}{N-n}, \quad \mu_v=E_v-kT\ln\frac{p}{N-p}. \quad (3.6)$$

Substituting (3.6) in (3.5):

$$F_{el}=(N-p)E_v+nE_c+nkT\ln\frac{n}{N-n}+pkT\ln\frac{p}{N-p} -NkT\ln\frac{N}{N-n}-NkT\ln\frac{N}{N-p}. \quad (3.7)$$

Adopting the charge-neutrality condition, $n=p$, simplifies (3.7) to

$$F_{el}=(N-n)E_v+nE_c+2nkT\ln\frac{n}{N-n} -2NkT\ln\frac{N}{N-n}. \quad (3.8)$$

The first two terms on the right of (3.8) give the energy of the system at zero temperature; the remaining two terms represent the entropy contribution to the free energy.

Consider a small distortion ϵ; the increase in strain free energy, to second order in ϵ can be written as $B\epsilon^2$. Using (2.22), (2.23), and (2.24) in (3.8), we obtain for the free energy of the system as a function of the distortion, the free-carrier concentration, and the temperature:

$$F(\epsilon,n,t)=NE_0+\frac{ND}{2}(1-\delta)\left(1-\frac{2}{1-\delta}\frac{n}{N}\right) -\frac{N}{2}(1-\delta)A\epsilon\left(1-\frac{2}{1-\delta}\frac{n}{N}\right)+B\epsilon^2 +2nkT\ln\frac{n}{N-n}-2NkT\ln\frac{N}{N-n}. \quad (3.9)$$

Minimization of (3.9) with respect to ϵ yields

$$\epsilon=\epsilon_0\left[1-\frac{2}{1-\delta}\frac{n}{N}\right], \quad (3.10)$$

where $\epsilon_0\equiv NA^2(1-\delta)/4B$ is the zero-temperature distortion. From (2.24),

$$E_g=E_{g0}\left[1-\left(\frac{2}{1-\delta}\right)\left(1+\frac{D}{E_{g0}}\right)\frac{n}{N}\right] =E_{g0}\left[1-\left(\frac{2}{1-\bar{\delta}}\right)\frac{n}{N}\right],$$

where $\bar{\delta}\equiv\delta+D(1-\delta)/E_{g0}$. Substitution of (3.10) into (3.9) gives

$$F(n,T)=NE_0-\frac{N(1-\bar{\delta})}{4}E_{g0}\left(1-\frac{2}{1-\bar{\delta}}\frac{n}{N}\right)^2 +2nkT\ln\frac{n}{N-n}-2NkT\ln\frac{N}{N-n}. \quad (3.11)$$

The first term on the right of (3.11) is a constant representing the electronic energy of the undistorted system, and can be dropped by setting NE_0 as the zero

of energy. We may then write (3.11) as

$$\frac{F(x,y)}{2NkT}=-\frac{y(1-\bar{\delta})}{4}\left(1-\frac{2}{1-\bar{\delta}}x\right)^2 +x\ln x+(1-x)\ln(1-x), \quad (3.12)$$

where $x\equiv n/N$ and $y\equiv E_{g0}/2kT$. Minimizing (3.12) with respect to x shows

$$y=\ln\frac{1-x}{x}\Big/1-\left(\frac{2}{1-\bar{\delta}}\right)x. \quad (3.13)$$

Equation (3.13) is just what we would expect from application of Fermi-Dirac statistics to a system described by (3.4) and (2.30).[22] Inversion of (3.13) gives the carrier concentration as a function of temperature. Since, for a constant mobility, electrical conductivity is proportional to carrier concentration, it is useful to perform this inversion. For the case of Boltzmann statistics, (3.13) can be inverted analytically. The Boltzmann approximation consists in replacing the numerator of the right side of (3.13) by $\ln(1/x)$, and is valid as long as x does not get too large. In this approximation, with substitutions

$$\eta\equiv xe^{y},$$
$$\tau\equiv\exp\left[\left(\frac{2}{1-\bar{\delta}}\right)ye^{-y}\right], \quad (3.14)$$

(3.13) becomes

$$\ln\tau=(\ln\eta)/\eta.$$

The function $\eta(\tau)$ is plotted in Fig. 2. For a given τ, corresponding to a particular temperature, there are two values of η, corresponding to two carrier concentrations. However, the upper intersection corresponds to a maximum of the free energy and thus has no physical significance. The lower intersection is easily shown to correspond with a relative minimum in the free energy.

The function $\eta(\tau)$ can be expressed explicitly by

$$\eta(\tau)\equiv\lim_{n\to\infty}\eta_n, \quad (3.15)$$

where $\eta_0=\tau$ and $\eta_{n+1}=\tau^{\eta_n}$. In Fig. 2 the point $T=0$ corresponds to $\tau=1$. As the temperature is increased, τ increases monotonically. At a given temperature T_0, corresponding to $\tau_0=e^{1/e}$, a singularity exists, which would correspond to a second-order transition to a metallic phase. Near this transition it can be shown that the gap disappears like $(T_0-T)^{1/2}$. In fact, this point is only reached when $\bar{\delta}$ is exactly 0. For all finite $\bar{\delta}$, a smaller temperature T_0 exists at which point the free energy of the metallic phase becomes lower than that of the semiconducting phase, and a first-order transition occurs. To analyze this, we must look at (3.12). At a given temperature, the free energy has a local minimum

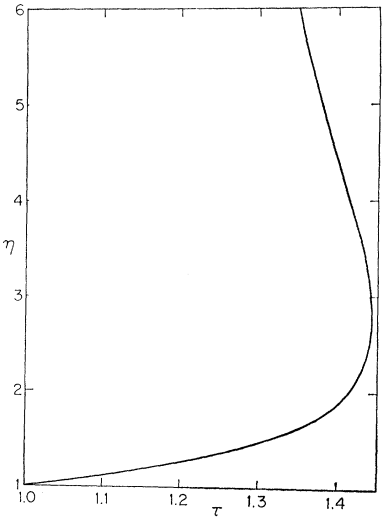

FIG. 2. The function $\eta(\tau)$.

given by the lower branch of the inversion of (3.13); there is also a maximum at a higher x corresponding to the upper branch of this inversion. For still greater x, the free energy is a monotonically decreasing function. The largest physical value of x is the "metallic" state, $x=\frac{1}{2}$, which represents a narrow half-filled band. Clearly, when x becomes large, the approximations of a constant δ and a linear E_g given in (2.24) break down, since otherwise (2.24) would imply a negative energy gap. However, we are not interested in this intermediate range. As we have noted, near the above-described metallic state, ϵ is very small, and the free-energy expression (3.12) will apply provided we take $\bar{\delta}-D=0$, or $\bar{\delta}=0$. Thus, the free energy of the metallic state is simply

$$F_{\mathrm{Met}}/2NkT=\tfrac{1}{2}\ln\tfrac{1}{2}+\tfrac{1}{2}\ln\tfrac{1}{2}=-0.693. \quad (3.16)$$

The free energy of the semiconducting phase is obtained by substituting the lower branch of the inversion of (3.13) into (3.12). The temperature at which (3.16) drops below this value is the temperature of a first-order semiconductor-to-metal transition. The transition temperature is plotted as a function of $\bar{\delta}$ in Fig. 3. Away from $\bar{\delta}=0$, the maximum value of x in the semiconducting region is quite small, and thus ϵ and E_g do not vary much below the transition. This means that the ϵ dependence of δ and E_g will not change significantly in the semiconducting region, and thus that the analysis that went into obtaining (3.12) is consistent.

Figure 3 gives E_{g0}/kT_0 as a function of $\bar{\delta}$. The second-order transition temperatures E_{g0}/kT_c are also shown as a dotted line. For a given material, δ, D, and thus $\bar{\delta}$ can be expected to be relatively constant as pressure or stress is applied. If so, E_{g0}/kT_0 will be constant. This can be expressed as

$$\frac{d\ln E_{g0}}{dX}=\frac{d\ln T_0}{dX}, \quad (3.17)$$

where X is any external parameter which does not change the value of $\bar{\delta}$ for the material. Equation (3.17) can be tested experimentally, and is an important prediction of this theory.

As we remarked in Sec. III, for crystalline distortions in narrow bands, values of δ centered around $\frac{1}{3}$. Using this as a typical $\bar{\delta}$, we find

$$E_{g0}/kT_0=8.10. \tag{3.18}$$

It is also important to calculate the value of x just before the transition, since this will give the jump in free-carrier concentration at T_0. These are plotted as a function of $\bar{\delta}$ in Fig. 4. For $\bar{\delta}=\frac{1}{3}$,

$$n_0/N=0.023. \tag{3.19}$$

Thus, the carrier concentration jumps by about a factor of 50 at T_0. Since it is likely that the mobility in extremely narrow bands changes considerably from the semiconducting to the metallic state, we cannot predict the metallic conductivity from this analysis. However, the low value of n/N found in (3.19) is essential to confirm the validity of the assumption that $\bar{\delta}$ does not change in the semiconducting region. We can also use the information provided by (3.19) to estimate the jump in mobility from the observed conductivity discontinuity.

The narrow-band limit is an extreme idealization which is never actually obtained. In reality, the bands must have some width. The case of a Gaussian broadening about delta-function bands can be handled without much difficulty, and can be solved analytically for Boltzmann statistics,[22] which is a good approximation if

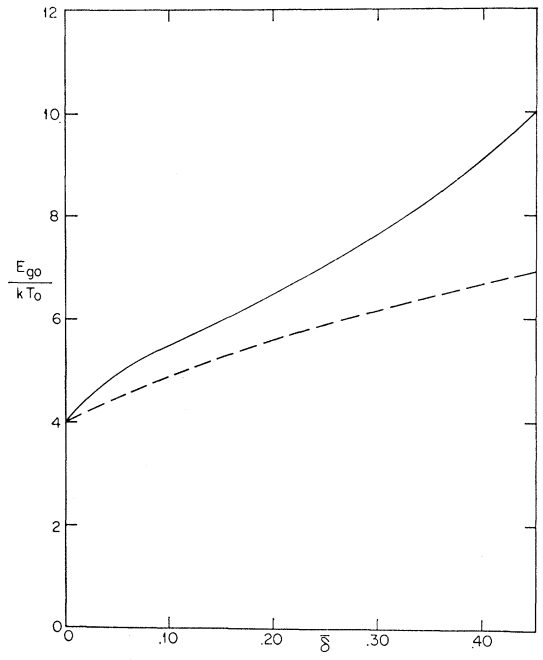

FIG. 3. E_{g0}/kT_0 as a function of $\bar{\delta}$; narrow-band limit.

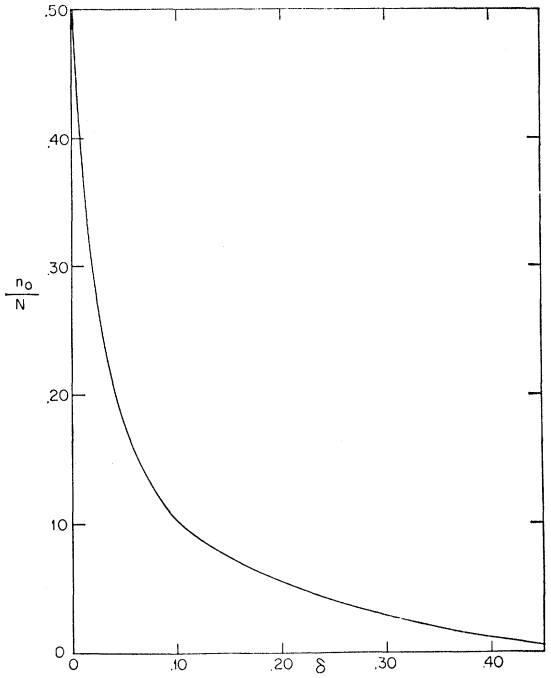

FIG. 4. n_0/N as a function of $\bar{\delta}$; narrow-band limit.

the spread is small. A Gaussian spread of a narrow band furthermore is probably a reasonable representation for a band in which conduction occurs by uncorrelated hopping. For the densities of states given by

$$\rho_c(E)=\frac{N}{\pi\lambda}\exp-\left[\frac{(E-2E_g/3)^2}{\lambda^2}\right],$$

$$\rho_v(E)=\frac{N}{\pi\lambda}\exp-\left[\frac{(E+E_g/3)^2}{\lambda^2}\right],$$

it is found that the solution can be mapped onto the one for $\lambda=0$ by introducing a renormalized $\bar{\delta}$:

$$\bar{\delta}_{\text{eff}}=\bar{\delta}+(1-\bar{\delta})\frac{\lambda^2}{4k^2T_0^2}.$$

Thus, introducing a small spread to the bands tends to raise E_{g0}/kT_0 somewhat. This means that the transition occurs at a slightly lower temperature then in the zero-bandwidth case, as one might intuitively expect. For $\lambda/E_{g0}=0.01$ and $\bar{\delta}=\frac{1}{3}$, $E_{g0}/kT_0=8.10$, the same value as when $\lambda=0$; for $\lambda/E_{g0}=0.05$, $E_{g0}/kT_0=8.24$. However, we note that the effective band gap should more precisely be taken as $(E_{g0})_{\text{eff}}=E_{g0}-2\lambda$, since this represents the energy separation between appreciable densities of states in the two bands. Thus, for the first example above, $(E_{g0})_{\text{eff}}/kT_0=7.94$, and for the second example, $(E_{g0})_{\text{eff}}/kT_0=7.42$. Thus, for a given energy gap the transition temperature is raised as the bands are widened, as would be expected.

B. Effective-Mass Approximation

We also can consider the situation where the valence- and conduction-band structure in the semiconducting state is well represented by ellipsoidal constant energy surfaces. We can then define density-of-state effective masses, m_e for electrons, m_h for holes, and write

$$\rho_c(E)=\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}[E-E_0-(1-r)E_g]^{1/2},$$
$$\rho_v(E)=\frac{1}{2\pi^2}\left(\frac{2m_h}{\hbar^2}\right)^{3/2}[E_0-rE_g-E]^{1/2}. \tag{3.20}$$

Sufficiently near the band extrema, (3.20) is always a good approximation. But for large m_e or m_h, the equations do not apply for a very large energy range. The effective-mass approximation assumes that (3.20) would remain appropriate for all values of E which have a finite probability of electron or hole occupation at the temperature under consideration. In this range of applicability, the densities of states given by (3.20) arise, for example, from the spherical bands

$$E_c(k)=E_0+[(1-r)^2E_g^2+a_e^2k^2]^{1/2},$$
$$E_v(k)=E_0-[(rE_g)^2+a_h^2k^2]^{\frac{1}{2}}, \tag{3.21}$$

where

$$a_e^2\equiv\frac{\hbar^2(1-r)E_g}{m_e};\quad a_h^2\equiv\frac{\hbar^2rE_g}{m_h}.$$

The bands given by (3.21) are a somewhat more plausible representation of the band structure away from the band edges than the ordinary effective mass approximation represented in Eq. (3.20). When the gap is large, (3.21) reduces to the usual effective-mass bands. Substituting (3.20) into (3.2), and taking into account the fact that the total concentration of states in each band is N, we obtain the following expression for free energy:

$$F_{el}=N\bar{E}_v+n\mu_c-p\mu_v$$
$$-\frac{2}{3}\frac{2}{\sqrt{\pi}}A_eT^{3/2}(kT)F_{3/2}\left[\frac{\mu_c-E_0-(1-r)E_g}{kT}\right]$$
$$-\frac{2}{3}\frac{2}{\sqrt{\pi}}A_hT^{3/2}(kT)F_{3/2}\left[\frac{E_0-rE_g-\mu_v}{kT}\right], \tag{3.22}$$

where

$$A_e\equiv\left(\frac{1}{4}\right)\left(\frac{2m_ek}{\pi\hbar^2}\right)^{3/2};\quad A_h\equiv\left(\frac{1}{4}\right)\left(\frac{2m_hk}{\pi\hbar^2}\right)^{3/2}, \tag{3.23}$$

and

$$F_n(\zeta)=\int_0^\infty\frac{x^n dx}{\exp(x-\zeta)+1}. \tag{3.24}$$

Minimization of (3.24) with respect to the quasi-Fermi levels gives

$$n=\frac{2}{\sqrt{\pi}}A_eT^{3/2}F_{1/2}\left[\frac{\mu_c-E_0-(1-r)E_g}{kT}\right],$$
$$p=\frac{2}{\sqrt{\pi}}A_hT^{3/2}F_{1/2}\left[\frac{E_0-rE_g-\mu_v}{kT}\right]. \tag{3.25}$$

In the Boltzmann limit, Eq. (3.25) gives for the carrier concentrations,

$$n=A_eT^{3/2}\exp\left[\frac{\mu_c-E_0-(1-r)E_g}{kT}\right],$$
$$p=A_hT^{3/2}\exp\left[\frac{E_0-rE_g-\mu_v}{kT}\right], \tag{3.26}$$

and the electronic free energy becomes simply

$$F_{el}=N\bar{E}_v+n\mu_c-p\mu_v-kT(n+p). \tag{3.27}$$

From (3.26), the quasi-Fermi energies are

$$\mu_c=E_0+(1-r)E_g+kT\ln\frac{n}{A_eT^{3/2}},$$
$$\mu_v=E_0-rE_g-kT\ln\frac{p}{A_hT^{3/2}}, \tag{3.28}$$

which substituted into (3.27) yields

$$F_{el}=N\bar{E}_v+nE_0+n(1-r)E_g-pE_0-prE_g$$
$$+nkT\ln\frac{n}{A_eT^{3/2}}+pkT\ln\frac{p}{A_hT^{3/2}}-kT(n+p). \tag{3.29}$$

If we assume the condition of charge neutrality, $n=p$, Eq. (3.29) now simplifies to

$$F_{el}=N\bar{E}_v+nE_g+2nkT\ln\frac{n}{eA^*T^{3/2}}, \tag{3.30}$$

where $A^*\equiv(A_eA_h)^{1/2}$. Equation (3.30) is the wide-band analogy to (3.8), and if the valence and conduction bands were rigidly displaced by a small distortion, then all results would follow just as in the narrow band approximation. However, as can be seen from Fig. 1, this is not the case, the distortion being localized near the band edges, so that $\bar{E}_v$ decreases less rapidly than E_g. Although the thermodynamic functions could be worked out with the approximation (3.20), in order to obtain the displacement of $\bar{E}_v$ in terms of E_g, it is necessary to use the more realistic model (3.21), since the averaging involved in $\bar{E}_v$ includes states in the

valence band far away from the gap. We thus obtain

$$\bar{E}_v = E_0 - \frac{3}{k_0^3}\int_0^{k_0} [(rE_g)^2 + a_h^2 k^2]^{1/2} k^2 dk$$

$$= E_0 - \left(\frac{3}{8y_0^3}\right)(rE_g)\{y_0(1+2y_0^2)(1+y_0^2)^{1/2} - \ln[y_0 + (1+y_0^2)^{1/2}]\}, \quad (3.31)$$

where $y_0 \equiv a_h k_0 / rE_g$ is determined from the requirement that the concentration of states in the band is N. In the wide-band limit where the effective-mass approximation is most appropriate, y_0 is large, and the complicated expression (3.31) simplifies to

$$\bar{E}_v = (E_0 - (3/4)a_h k_0) - \frac{15}{8}\frac{r^2}{a_h k_0} E_g^2. \quad (3.32)$$

Thus, in this limit, the reduction in electronic energy is proportional to the square of the distortion, as we indicated in Sec. IID. Since the cost in strain energy is still proportional to ϵ^2, such distortions do not occur when the bands are wide.

On the other hand, if the bands are narrow, y_0 is small, and (3.31) results in

$$\bar{E}_v = E_0 - (rE_g)(1 + (3/10)y_0^2). \quad (3.33)$$

Substitution of the leading term of (3.33) into (3.30), adding the increase in elastic free energy, and using (2.24) gives

$$F = NE_0 - rN\left(1 - \frac{1}{r}x\right)(A\epsilon - D) + B\epsilon^2 + 2NkTx \ln\frac{Nx}{eA^* T^{3/2}}. \quad (3.34)$$

From (3.34), we immediately obtain the equilibrium distortion as

$$\epsilon = (NA/2B)r(1 - (1/r)x) = \epsilon_0(1 - (1/r)x), \quad (3.35)$$

which recaptures the result (3.10). Elimination of ϵ from (3.34) gives

$$\frac{F(x,y)}{2NkT} = \frac{E_0}{2kT} - \frac{\bar{r}}{2}\left(1 - \frac{1}{\bar{r}}x\right)^2 y + x\ln(Cxy^{3/2}), \quad (3.36)$$

where $\bar{r} \equiv r(1 - D/E_{g0})$, y is defined as in Sec. A, and $C \equiv N(2k)^{3/2}/eE_{g0}^{3/2}A^*$.

Minimizing (3.36) with respect to x,

$$eCxy^{3/2} = e^{-y(1-(1/\bar{r})x)}, \quad (3.37)$$

which is the analogue of (3.13), and is just what one would expect from naive application of Boltzmann statistics. It may be solved explicitly with the substitution

$$Q \equiv \exp\left[\frac{y^{-1/2}e^{-y}}{reC}\right],$$

$$Z \equiv eCy^{3/2}e^y x.$$

Then (3.37) becomes

$$Z = Q^z = \eta(Q), \quad (3.38)$$

the same function as defined in (3.15). As in Sec. A, the solution of (3.38) which minimizes the free energy has a singularity which would correspond to a second-order transition to a metallic state. The point y_c corresponds to $Q_c = e^{1/e}$, or

$$y_c^{-1/2} = \bar{r}Ce^{y_c}. \quad (3.39)$$

Let

$$S \equiv (\bar{r}C)^2 y_c,$$

$$\ln(1/R) \equiv 2(\bar{r}C)^{-2}. \quad (3.40)$$

Then (3.39) can be written

$$S = R^s = \eta(R),$$

and the critical temperature is given by

$$2y_c = S\ln(1/R) = \eta(R)\ln(1/R). \quad (3.41)$$

From (3.40), the parameter $\ln(1/R)$, which determines y_c, is

$$\ln(1/R) = \left(\frac{e}{N\bar{r}}\right)^2 \left(\frac{m^* E_{g0}}{2\pi\hbar^2}\right)^3. \quad (3.42)$$

Thus, $\ln(1/R)$ is essentially a measure of the ratio of band gap to band width. The quantity $2y_c \equiv E_{g0}/kT_c$ is plotted as a function of $\ln(1/R)$ in Fig. 5. For narrow bands, as we have assumed, $\ln(1/R)$ is large, and E_{g0}/kT_c is a slowly varying function of R. Since in this region $E_{g0}/kT_c > 5$, Boltzmann statistics are appropriate. There appears to be some overlap between the effective-mass approximation and the example of a

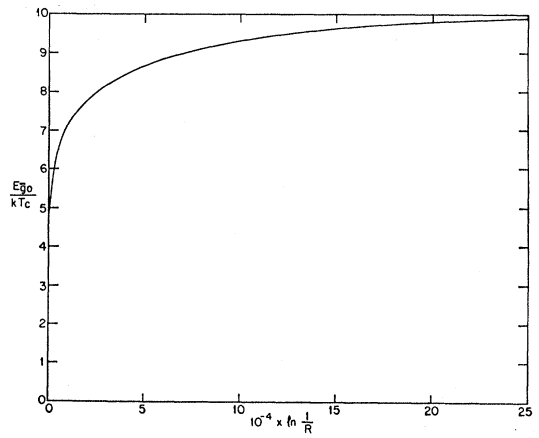

FIG. 5. E_{g0}/kT_0 as a function of $\ln(1/R)$; effective-mass approximation, Boltzmann statistics.

Gaussian spread around the narrow-band limit given in the last section. From (3.42) and the definitions of y_0 and a_h, we can write

$$y_0^2=(9\pi e^2/8\bar{r}^5)^{1/3}[\ln(1/R)]^{1/3}.$$

For $\bar{r}=\frac{1}{3}$ and $\ln(1/R)=1.0\times10^4$, this results in $3y_0^2/10=0.2$, close to the upper limit of validity of the expansion in (3.33). Thus, $\ln(1/R)$ must be of this order of magnitude or larger for Fig. 5 to be applicable. Expressing $\ln(1/R)$ in terms of the effective-mass approximation band width, $E_b\equiv\hbar^2k_0^2/2m$, we find

$$\ln(1/R)=(81\pi e/64)(E_{g0}/E_b)^3.$$

For a band-gap to band-width ratio of 10, $\ln(1/R)=1.1\times10^4$, a value sufficiently large so that (3.33) is valid. In this case, (3.41) yields $E_{g0}/kT_c=7.3$. For the same band-gap to band-width ratio in the Gaussian bands of the last section, we found $(E_{g0})_{\text{eff}}/kT_0=7.4$.

Just below the transition, $\eta=e$. Thus, from (3.37), at this point

$$x(T_c)=(1/C)y_c^{-3/2}e^{-y_c}. \tag{3.43}$$

But (3.39) and (3.43) imply

$$x(T_c)=\bar{r}/y_c. \tag{3.44}$$

When Boltzmann statistics are applicable, y_c is significantly larger than 1; also $\bar{r}$ is never greater than $\frac{1}{2}$ and can be much less. Thus, the fraction of carriers excited before the occurrence of a second-order transition can be quite small. If a first-order transition occurs at a lower temperature than T_c, the critical carrier concentration is still lower than that given by (3.44). Using (3.36) and (3.44), the free energy of the semiconducting state at T_c can be evaluated as

$$\frac{F(y_c)}{2NkT}=\frac{E_0}{2kT}-\frac{\bar{r}}{2}(y_c+y_c-1). \tag{3.45}$$

Equation (3.45) is valid provided that the gap just below T_c is still sufficiently large so that the approximation (3.33) in appropriate.

In order to determine the order of the transition, we must calculate the free energy of the metallic state. We shall assume that the effective-mass approximation is a good description also of the metallic band structure, although the effective mass in the metallic state $\bar{m}$ need not bear any relation to the semiconducting electron and hole effective masses. With $\rho(E)=CE^{1/2}$, the metallic free energy is

$$F_{Met}=NE_F-CkT\times\int_0^\infty dE\,E^{1/2}\ln\left[1+\exp\left(\frac{E_F-E}{kT}\right)\right]. \tag{3.46}$$

Proceeding as before, but applying the condition of degeneracy, we obtain the usual result

$$F_{\text{Met}}=\tfrac{3}{5}NE_0-(\pi^2Nk^2T^2/4E_0), \tag{3.47}$$

where we have identified the Fermi energy of the metallic state with E_0, the energy of the top of the semiconducting valence band when the gap has shrunk to zero. In terms of $\bar{m}$,

$$E_0=(\hbar^2/2\bar{m})(3\pi^2N)^{2/3}, \tag{3.48}$$

where we have taken the bottom of the band as the zero of energy. From (3.47) and (3.48),

$$\frac{F_{Met}}{2NkT}=\frac{3}{5}\frac{E_0}{2kT}-\left(\frac{\pi^2}{3N}\right)^{2/3}\frac{\bar{m}E_{g0}}{8\hbar^2}y^{-1}. \tag{3.49}$$

The metallic free energy given by (3.49) must be compared to the minimum semiconducting free energy given by (3.36) and (3.37). If (3.49) drops below the minimum of (3.36) at some temperature T_0, then this is the temperature of a first-order semiconductor-to-metal transition. If this does not occur below T_c, then T_c represents the point of a second-order transition. If (3.45) is valid, then the order of the transition can be determined by evaluating (3.49) at $y=y_c$ and comparing the result to (3.45). Using (3.39) with $\bar{r}=\frac{1}{2}$, the proper metallic value, and using the definition of C, we find

$$\frac{F_{Met}(y_c)}{2NkT}=\frac{3}{5}\frac{E_0}{2kT}-\frac{\pi}{2}\left(\frac{\pi}{12e}\right)^{2/3}\frac{\bar{m}}{m^*}y_c^{-2/3}e^{2y_c/3}. \tag{3.50}$$

If (3.50) is lower than (3.45), a first-order transition takes place at $T_0<T_c$. If (3.50) lies above (3.45), the transition is a second-order one at $T_0=T_c$. We can expect E_0 to be small compared to E_{g0}, since the bands are relatively narrow. It can then be seen from (3.45) and (3.50) that the transition will be first order unless y_c is small. But the approximations which led to (3.45) are consistent only with a large y_c. If the initial gap is small, Fermi statistics must be used. This analysis has been carried out[22] and it is found that Fermi statistics raise T_c, and thus lower y_c. This effect is enhanced if the electron and hole effective masses differ significantly. Thus, as a practical matter, in the range in which the effective-mass approximation can be used, the transition will always be first order.

We have tacitly been assuming that the energy gap in the semiconducting state is a direct one, the valence band maximum and conduction band minimum occurring at the same point in **k** space. If this is not the case, two new possibilities emerge, at least in the wide-band approximation. Firstly, the effective $\bar{r}$ can be significantly increased if the real value of the indirect gap E_g is much smaller than the theory of Sec. II indicated. This can be taken into account by using an effective $\bar{r}$ given by

$$\bar{r}_{\text{eff}}\equiv\bar{r}\frac{(E_g)_{\text{real}}}{(E_g)_{\text{apparent}}},$$

where $(E_g)_{\text{apparent}}$ is the direct gap in **k** space. Since a

small $\bar{r}$ favors a first-order transition and, according to (3.41), (3.42), and (3.44), a low-carrier concentration on the semiconducting side of the transition, it is seen that the existence of an indirect gap enhances the sharpness of the transition. Secondly, the possibility of a transition to a semimetallic state rather than a fully metallic state exists. The analysis of such a transition is completely straightforward and can be carried out by retaining the expressions for F_{e1} in the form (3.24). The semimetallic state is that which the Fermi functions can be replaced by their value under degenerate statistics:

$$F_{3/2}(\zeta)=\tfrac{2}{5}\zeta^{5/2}+(\pi^2/4)\zeta^{1/2},$$
$$F_{1/2}(\zeta)=\tfrac{2}{3}\zeta^{3/2}+(\pi^2/12)\zeta^{-1/2}.$$

Once again, it is found that the semiconductor-to-semimetal transition can be either first or second order, with the first-order situation much more probable. The possibility of a second order in this case exists because the system can become degenerate before the direct gap has disappeared. This possibility is rendered more unlikely by the requirement that the semimetallic state also have lower free energy than the fully metallic state in order for it to be stable.

C. General Case

In Sec. B we saw from Eq. (3.32) that in the wide-band limit, distortion of the lattice would not take place even at absolute zero. In this section we shall examine the zero-temperature energy of the crystal as a function of fractional distortion ϵ for a general band model which permits adjustment to represent any ratio of band width to band gap, and shall determine under what conditions the crystal can spontaneously distort. We use a model in which the bands are spherical around the conduction and valence band edges, the energies being given by

$$E_v(k)=E_0-[(rE_g)^2+4\beta^2\sin^2(\tfrac{1}{2}ka)]^{1/2},$$
$$E_c(k)=E_0+[(1-r)^2E_g^2+4\beta^2\sin^2(\tfrac{1}{2}ka)]^{1/2}. \quad (3.51)$$

This form of function is that obtained by nearest-neighbor tight-binding calculations in a one-dimensional crystal, and the model is simply assumed to have spherical symmetry when extended to three dimensions. Although no real band structure has exactly this form, the model is probably sufficiently realistic for our purposes.

We can obtain the effective-mass approximation from (3.51) by taking β large compared to E_g, or the narrow-band limit by taking β small compared to E_g. $E_g=0$ gives metallic bands whereas $\beta=0$ leads to the zero band-width case.

In order to determine whether the crystal will distort we want to evaluate the total energy of electrons in a filled valence band as a function of the energy gap. Thus, we are interested in

$$\bar{E}_v=\sum_k E_v(k)$$
$$=NE_0-3N(rE_g)\int_0^1 dz\, z^2\left(1+\lambda^2\sin^2\frac{\pi z}{z}\right)^{1/2}, \quad (3.52)$$

where $\lambda\equiv 4\beta/E_g$. The integral in (3.52) can be related to elliptic integrals of the second kind, but for the present we can restrict ourselves to the two limits of wide bands ($\lambda^2\gg 1$) and narrow bands ($\lambda^2\ll 1$). In the wide-band limit, we find

$$\bar{E}_v=N\left[E_0-\frac{48\beta}{\pi^2}\left(1-\frac{2}{\pi}\right)\right]-\frac{K}{\beta}(rE_g)^2-\frac{\pi^2(rE_g)^3}{384\beta^2}, \quad (3.53)$$

where

$$K=-\frac{3}{\pi}\sum_{n=1}^{\infty}(-1)^h n^2\ln\left(1-\frac{1}{4n^2}\right)$$
$$\approx 0.150.$$

This limit is the one where the gap E_g is small compared to β, which is always true for sufficiently small distortions. In this case $E_g=A\epsilon$, and the total energy of the system at $T=0$ to second order in ϵ is

$$E=N\left[E_0-\frac{48\beta}{\pi^2}\left(1-\frac{2}{\pi}\right)\right]+\left(B-\frac{KA^2r^2}{\beta}\right)\epsilon^2, \quad (3.54)$$

where $B\epsilon^2$ is the increase in elastic energy brought about by the distortion. Since we expect $B\gg A$, (3.54) shows that there is always a local minimum of $E(\epsilon)$ at $\epsilon=0$. If β is large enough that (3.53) remains a good approximation as ϵ gets large, it is clear that the crystal will not distort. However, when the bands are quite narrow, β will be small, and for sufficiently large ϵ, the opposite limit, ($\lambda^2\ll 1$), becomes appropriate. Then, from (3.52),

$$\bar{E}_v=NE_0-N(rE_g)\left[1+\frac{3\lambda^2}{2\pi^2}\left(1+\frac{\pi^2}{6}\right)\right].$$

In this region, (2.24) should apply, and thus the total zero-temperature energy of the system can be written

$$E=N(E_0-rD)-NAr\epsilon-(\Delta/\epsilon)+B\epsilon^2, \quad (3.55)$$

where $\Delta\equiv(1+6/\pi^2)N\beta^2/A$. Equation (3.55) exhibits a local minimum at a finite ϵ:

$$\epsilon_0=\frac{NAr}{2B}\left[1-\frac{4B^2\beta^2}{N^2A^4r^3}\left(1+\frac{6}{\pi^2}\right)\right], \quad (3.56)$$

so that the crystal will distort if the energy at this ϵ_0 is lower than the energy at $\epsilon=0$. At ϵ_0, the energy can be expressed

$$E=N\left\{\left(E_0-\frac{rD}{2}\right)-\left[\frac{r}{2}Er_0+\frac{2.4\beta^2}{E_{g0}}\right]\right\}, \quad (3.57)$$

where E_{g0} is the energy gap introduced by ϵ_0.

The energy of the local minimum which must occur at $\epsilon=0$ is obtained directly from (3.54) as

$$E=N[E_0-1.77\beta]. \tag{3.58}$$

Comparison of (3.57) and (3.58) shows that the crystal will distort if and only if

$$\frac{rE_{g0}}{2\beta}+\frac{2.4\beta}{E_{g0}}+\frac{rD}{2\beta}>1.77. \tag{3.59}$$

Ignoring D compared to E_{g0}, and taking $r=\frac{1}{3}$ as a typical value, (3.59) becomes

$$\beta<0.12E_{g0}. \tag{3.60}$$

Since the band width in the model under consideration is 4β, the criterion for distortion is that the band width be less than half the band gap. In general, when spherical bands are not assumed, the width of the band will be a larger multiple of the overlap integral β, so that the precise criterion is not quite as strict as this. However, since the gaps introduced by distortions of this type are of the order of tenths of electron volts, the distortions will take place only when the band width is of the same order of magnitude or smaller.

It might appear at first that the foregoing discussion is inconsistent with the conclusion following Eq. (3.42) that raising the band width relative to the band gap raises the transition temperature. However, the latter conclusion applies only for *fixed* band gap. In the present case, the expression (3.56) for ϵ_0 shows that the band gap itself decreases with increasing band width, and this tends to lower the transition temperature more than the slow change of y_c with $\ln(1/R)$ tends to raise it. Hence for fixed values of B, and A, and r the transition temperature will be lowered by increasing β as it should be.

Finally, note that the condition (3.60) corresponds to $0.3y_0^2<0.43$ and thus the expansion (3.33) is valid for the range of band-gap to band-width ratios for which the crystal will distort.

IV. DISCUSSION

We have shown that a narrow-band crystal, whose structure is such that it would otherwise give a half-filled conduction band, can lower its ground-state energy by distorting, and thus produce an insulating state at zero temperature. Wider band crystals of this type will not distort, but can lower their energy by ordering antiferromagnetically and also achieve low-temperature semiconduction. In both cases the gap introduced must decrease with thermal excitation of carriers, and a transition to a metallic state occurs at a given temperature which can be calculated from the band parameters of the system.

We wish to apply the theory to the narrow $3d$ bands of the transition metal oxides. The limit of zero width bands is, of course, much too extreme to be accurate, but the introduction of a small spread to the bands has very little effect on the final results. In particular, the expressions (3.17) and (3.18), which are the ones most easily experimentally tested, are virtually unchanged by Gaussian spreads of width up to one-quarter of the zero-temperature gap. Furthermore, in Sec. IIIC, we show how a general band model gives the narrow band results in one limit and the effective-mass approximation in the other. The conclusions here indicate that the narrow-band limit is the consistent one to use in discussing crystalline distortions.

The theory of Sec. III is worked out only for the case of a crystalline distortion. The analogous calculations for an antiferromagnet differ somewhat and will be detailed elsewhere, although the results are very similar. In the narrow-band limit, whereas a crystalline distortion always leads to a first-order transition, a saturated antiferromagnet exhibits a second-order transition. In general, either order transition is possible, although an antiferromagnet is more likely to undergo a second-order transition than is a distorted crystal, all other things being the same.

A first-order transition to a metallic state occurs when the free energy of such a state falls below the local minimum of $F(n,T)$ for the semiconducting state. However, the local minimum continues to exist right up to the second-order transition temperature, T_c. This local minimum might lead to a metastable state when the material in the semiconducting state is heated, and this would show up experimentally as a hysteresis. This hysteresis will not occur when the transition is second order, and this can provide a quick determination of the order of the transition directly from the electrical conductivity data. Thus, the results of Morin[18] indicate that V_2O_3, VO, and VO_2 undergo first-order transitions, whereas the transition in Ti_2O_3 is second order.

ACKNOWLEDGMENTS

We should like to thank Julius Feinleib, William Paul, Henry Ehrenreich, and Arthur Bienenstock for many helpful discussions at all stages of this work.