Electrons in Perturbed Periodic Lattices*

J. C. SLATER
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received June 17, 1949)

It is shown that the motion of an electron in a periodic potential, such as is found in a solid, plus a slowly varying perturbative potential, can be derived from the energy in the periodic lattice alone, as a function of momentum or wave number. A Schrödinger equation is set up, in which the Hamiltonian is the sum of this energy in the periodic lattice—the momentum being replaced by a differential operator—and of the perturbative potential energy. The resulting wave function modulates atomic functions to provide a solution of the perturbed problem. This method is applied to give proofs of simple theorems in conduction theory, to discuss the energy levels of impurity atoms in a semiconductor, and to consider excitons; all are problems which have been considered before, but which are treated more straightforwardly by the present method. Applying the method statistically, the combined Poisson's equation and Fermi-Dirac statistics is set up for impurities in metals and semiconductors, and for the theory of rectifying barriers.

I. INTRODUCTION

MANY of the most important problems in the theory of solids concern the motion of electrons in perturbed periodic lattices. Examples of such problems are: the effect of impurities, of the P- or N-type, in semiconductors; the behavior of rectifying barriers, either between semiconductors and metals, or between P- and N-type semiconductors; and the behavior of optically excited energy levels in crystals. All these problems have received much discussion in the literature, by methods involving various approximations to the solution of the wave-mechanical problem of the motion of electrons in a perturbed periodic lattice. It is the purpose of this paper to point out that there is a quite general theorem in wave mechanics, regarding the motion of an electron in such a perturbed lattice, which serves to unify the treatment of all these problems. This theorem was essentially discovered by Wannier,[1] and used by him in discussing excited energy levels of crystals. That problem, unfortunately, is one of the most complicated to which the method can be applied, and the theorem has tended to lie buried in Wannier's paper, attracting little attention, without general realization of the important simpler problems to which it is applicable. James[2] has independently arrived at many of the qualitative results of the present paper, but is apparently unaware of the importance of Wannier's theorem, and bases his conclusions on quite different and less powerful methods of discussing the problem, applicable only in the one-dimensional case, though he has carried his approximations one step further than we have. A number of other writers have used similar less powerful methods for special problems.[3]

In the present paper, we shall first state the general theorem, and prove it by a method similar to that used by Wannier for his special problem. We shall then apply it to discussion of the important problems, such as impurity semiconductors and excited energy levels, which result in discrete energy levels. For problems of rectifying barriers and surface states, a statistical approach is more appropriate, and we point out the relation of our theorem to self-consistent field methods, and thence to statistical treatment by the Fermi-Dirac statistics. This leads us to a discussion of rectifying barriers and surface states, not essentially different from that appearing in the literature, but somewhat more general and unified.

II. THE MOTION OF ELECTRONS IN PERTURBED PERIODIC LATTICES

In Appendix I we give a general proof of our theorem; in this section we shall merely state it and discuss its applications. The theorem is one which starts by assuming that the problem of the motion of electrons in a given periodic lattice has been solved, and uses that solution as the starting point for discussion of the problem in which the potential is the sum of the original periodic potential and an additional potential varying only slightly from atom to atom of the periodic potential. We first remind the reader of the nature of the solution of the periodic potential problem.[4] We describe the solution as if we were considering a metal, although the extension to non-metals with more than one atom per cell presents no difficulties. We surround each atom by an appropriate polyhedral cell, the vectors from the origin (located at the nucleus of one of the atoms) to other atoms being denoted by Q_k, so that the potential is unchanged when we make the translation Q_k. Then each solution of the wave equation can be characterized by a vector quantity p, of the dimensions of a momentum, such that the wave function is multiplied by a

* This work has been supported in part by the Signal Corps, the Air Materiel Command, and the ONR.

[1] G. H. Wannier, "Structure of electronic excitation levels in insulating crystals," Phys. Rev. **52**, 191 (1937).

[2] H. M. James, "Electronic states in perturbed periodic systems," unpublished report of Contract No. W-36-039-SC-32020, Department of Physics, Purdue University. The writer is indebted to Professor K. Lark-Horovitz, director of the project, for the privilege of inspecting this report. See also paper in Phys. Rev., this issue.

[3] For instance, S. C. Tibbs, Trans. Faraday Soc. **35**, 1471 (1939), who discussed excited energy levels in NaCl, and S. Peckar, J. Phys. U.S.S.R. **10**, 431 (1946).

[4] We use substantially the notation in J. C. Slater, Rev. Mod. Phys. **6**, 209 (1934).

factor $\exp[(i/\hbar)p\cdot Q_k]$ when we make the translation Q_k. Bloch[5] made a well-known approximation to the form of the wave function by starting with functions $u(q-Q_k)$ (where q is the vector position of the point where we are finding the wave function) representing the wave function of an electron at a vector displacement $q-Q_k$ from the nucleus of the kth atom, in the case where that atom alone was present; such a function is generally called an atomic function. He then set up the approximate wave function

$$b(p, q)=\sum_k \{\exp[(i/\hbar)p\cdot Q_k]\}u(q-Q_k), \qquad (1)$$

which clearly has the required periodicity property and which, at the same time, behaves in the neighborhood of each atom like an atomic wave function.

This Bloch function $b(p, q)$ suffers from two defects. First, it is not an exact solution of the problem; secondly, the functions $u(q-Q_k)$, for different k's, are not orthogonal to each other, so that in calculating any sort of integrals over the Bloch functions we meet integrals coming from lack of orthogonality. Wannier[1] showed that both difficulties can be overcome by setting up a new set of atomic functions $a(q-Q_k)$ (see Appendix I for their definition), similar to the u's near each atom, but oscillating and falling off in amplitude like the function $(\sin x)/x$ at a distance from the atom. These functions have the properties that they are normalized, are exactly orthogonal to each other, and when they are substituted in an expression like (1) they form an exact solution of the problem of the periodic lattice. Thus the real solution $\psi_0(p, q)$ can be written in the form

$$\psi_0(p, q)=\sum_k 1/N^{\frac{1}{2}}\{\exp[(i/\hbar)p\cdot Q_k]\}a(q-Q_k). \qquad (2)$$

The factor $(1/N^{\frac{1}{2}})$, where N is the total number of atoms in the crystal, is introduced so that ψ_0 will be normalized when integrated over the whole volume of the crystal. We notice from (2) that the function $\psi_0(p, q)$ is periodic in p-space, or momentum space: if p is increased by one of the vectors P_j of the reciprocal lattice, defined by the relation $P_j\cdot Q_k=\text{integer}\times h$, the expression on the right side of (2) is unchanged. Then the energy $E_0(p)$, the energy of the level associated with a given p, will likewise be a periodic function of p, being unchanged when p increases by any one of the P_j's. All solutions can then be obtained by allowing p to range over the interior of the central zone in momentum space; it is easily shown that, when p is quantized by the boundary conditions appropriate to a finite crystal with N atoms, there will be N allowed stationary states. For a given p, there will, of course, be an infinite number of energy levels, just as for an isolated atom. The N levels continuously joined together, corresponding to the various allowed p's, form an energy band; we see that there are an infinite number of such bands. Their properties and relations to the theory of metals (where they overlap) and semiconductors or insulators (where gaps in energy remain between them) are well known.

Now we are ready to consider our problem of the perturbed periodic lattice. Let the Hamiltonian function of the unperturbed problem (the kinetic energy plus periodic potential energy) be H_0, so that the ψ_0's satisfy the equation

$$H_0\psi_0(p, q)=E_0(p)\psi_0(p, q). \qquad (3)$$

Then we wish to find functions $\psi_n(q)$, n being a quantum number, satisfying

$$H\psi_n(q)=E_n\psi_n(q), \qquad (4)$$

where $H=H_0+H_1$, H_1 being the slowly varying function of q. We try to express the ψ_n's in the form:

$$\psi_n(q)=\sum_k \Psi_n(Q_k)a(q-Q_k). \qquad (5)$$

That is, we try to find a function $\Psi_n(q)$ which we can use to modulate the atomic functions $a(q-Q_k)$ to get the correct solution of the problem, replacing the exponential function $(1/N^{\frac{1}{2}})\exp[(i/\hbar)p\cdot q]$ which is used in the problem of the unperturbed periodic potential, in the solution (2). Our theorem now states that $\Psi_n(q)$ satisfies the following differential equation, provided H_1 varies slowly with position, so that it does not change its value greatly from one atom to the next:

$$[E_0[-i\hbar(\partial/\partial q)]+H_1(q)]\Psi_n(q)=E_n\Psi_n(q). \qquad (6)$$

Here the first term $E_0(-i\hbar\partial/\partial q)$ stands as an abbreviation for the differential operator in which E_0, regarded as a function of the three rectangular components of the vector p, is transformed by replacing p_x by $-i\hbar\partial/\partial x$, p_y by $-i\hbar\partial/\partial y$, etc., as in the ordinary kinetic energy operator in Schrödinger's equation. In (6) we then have a Schrödinger equation for $\Psi_n(q)$, in which the perturbative potential H_1 appears as the potential energy, while the kinetic energy operator is derived from the energy $E_0(p)$ of the unperturbed problem by replacing p by a differential operator. It is this theorem which is proved in Appendix I, and which was applied to the problem of excited energy levels by Wannier.[1] By means of it, we effectively reduce the problem of electrons in periodic lattices and additional perturbing potentials to a problem much like that of free electrons in the perturbing potential (as we shall show in the next section) and hence make the problem of electrons in periodic lattices not much more complicated than free-electron theory.

III. APPLICATIONS OF THE GENERAL THEOREM

The first application which we shall make of our general theorem (6) is to the motion of wave packets of electrons in the perturbed periodic lattice. We can set up such wave packets just as well from the functions Ψ_n, which modulate the atomic functions, as from the functions ψ_n, which take into account the oscillations in the neighborhood of each atom. It is a well-known

[5] F. Bloch, Zeits. f. Physik **52**, 555 (1928).

theorem of quantum mechanics that the center of gravity of a wave packet moves according to the classical Hamiltonian equations. Since (6) is derived from the Hamiltonian $E_0(p)+H_1(q)$, we see that the equations of motion of the packet (writing them in terms of their rectangular components) are:

$$\frac{dx}{dt}=\frac{\partial E_0}{\partial p_x},\quad \frac{dy}{dt}=\frac{\partial E_0}{\partial p_y},\quad \frac{dz}{dt}=\frac{\partial E_0}{\partial p_z}, \tag{7}$$

and

$$\frac{dp_x}{dt}=-\frac{\partial H_1}{\partial x},\quad \frac{dp_y}{dt}=-\frac{\partial H_1}{\partial y}\quad \frac{dp_z}{dt}=-\frac{\partial H_1}{\partial z}. \tag{8}$$

Both these theorems are familiar, but they are ordinarily derived by much more involved methods than are used here.[6] In (7) we see the formula for the velocity of a wave packet in terms of the gradient of the function E_0 in momentum space, and in (8) the statement that the momentum p of a packet is governed by the classical equation of motion in terms of the additional force resulting from the perturbation H_1. These two results are the basis of most of the band theory of electrical conduction in solids, but it has hardly been realized that they form merely the classical Hamiltonian equations of the Hamiltonian of Eq. (6).

Our next example will be the behavior of wave packets near the bottom or top of an energy band, and hence the concept of effective mass. At the bottom of a band, provided the axes are properly oriented, the energy E_0 may be written in the form

$$E_0(p)=E_1+\frac{p_x^2}{2m_x}+\frac{p_y^2}{2m_y}+\frac{p_z^2}{2m_z}. \tag{9}$$

Here m_x, m_y, m_z are three coefficients of the dimensions of masses and E_1 the energy of the bottom of the band. In this case, (7) and (8) become

$$\frac{dx}{dt}=\frac{p_x}{m_x},\quad \frac{dp_x}{dt}=-\frac{\partial H_1}{\partial x},\qquad m_x\frac{d^2x}{dt^2}=-\frac{\partial H_1}{\partial x}, \tag{10}$$

with similar equations for the y- and z-components, showing that the packet obeys an ordinary equation of motion corresponding to a particle of mass m_x for the x-coordinate, m_y for the y-coordinate, and m_z for the z-coordinate. Similarly, near the top of a band, we have:

$$E_0(p)=E_1-\frac{p_x^2}{2m_x}-\frac{p_y^2}{2m_y}-\frac{p_z^2}{2m_z}, \tag{11}$$

where these m_x's are different from those in (9), but still positive. Hence the equations corresponding to (10) are

$$m_x(d^2x/dt^2)=\partial H_1/\partial x, \text{ etc.}, \tag{12}$$

showing that a packet at the top of a band will be accelerated by an external field in the opposite direction to a particle of positive mass exposed to the perturbative force. It is well known, and we need not repeat the discussion, that this leads a hole in an almost filled band, near the top of that band, to be accelerated as a positively charged particle of electronic charge and mass m_x, m_y, m_z would be.

If we are near the top or bottom of a band, so that one of the approximations (9) or (11) is correct, it is clear that Eq. (6) reduces to a Schrödinger equation of the conventional type, with a quadratic differential operator for the kinetic energy (though, in general, with different coefficients for the x-, y-, and z-derivatives). We can thus solve it by conventional methods, resulting in discrete energy levels as in atomic problems. We shall shortly give examples of this situation. In case the energy is such that these approximations are not appropriate, the problem becomes more complicated, higher derivatives entering the differential equation. In such circumstances, at least in one-dimensional problems, the most appropriate method of solution would presumably be the WKB approximation. This depends on finding the momentum p, which is equal to h/λ, where λ is the wave-length, in terms of position. From the equation $E_0(p)+H_1(q)=E$ we can find p, and hence set up the wave function and quantum condition.

There is a useful graphical method of discussing the solution, in the one-dimensional case (James[2] makes considerable use of this method). This is shown in Fig. 1. In Fig. 1(a), we show a schematic periodic potential, with its energy bands. In Fig. 1(b), we draw the energy bands, pushed upward for each value of x by the amount $H_1(x)$, the potential energy. We also draw a horizontal line of constant energy, E. We now see that the kinetic energy and momentum are determined by the position of E with respect to the energy bands, just as they would be in the absence of H_1. For instance, at point A in Fig. 1(b), the energy E is in the same position with respect to the energy band which the energy E' occupies in Fig. 1(a). Thus, at this point A, in the presence of the potential H_1, the de Broglie wave-length of the function Ψ will be the same as the de Broglie wavelength of the sinusoidal function in Fig. 1(a), corresponding to energy E'. Our graphical representation of Fig. 1(b) thus has many of the characteristics of an energy diagram in classical mechanics, in which potential energy and total energy are plotted as functions

[6] R. Peierls, Zeits. f. Physik **53**, 255 (1929); F. Bloch, Zeits. f. Physik **52**, 555 (1928); A. Sommerfeld and H. Bethe, *Handbuch der Physik* (Verlag. Julius Springer, Berlin, 1933), second edition, vol. XXIV, pp. 374–375, 506–509; H. Jones and C. Zener, Proc. Roy. Soc. A144, 101 (1934); C. Zener, Proc. Roy. Soc. A145, 523 (1934); L. Brillouin, *Les Electrons dans les Metaux du Point de Vue Ondulatoire* (Hermann and Cie, Paris, 1934); J. C. Slater, Rev. Mod. Phys. **6**, 209, 259–262, Appendix VI (1934); Mott and Jones, *Properties of Metals and Alloys* (Oxford University Press, New York, 1936), p. 92–96; W. V. Houston, Phys. Rev. **57**, 184 (1940); and many other references.

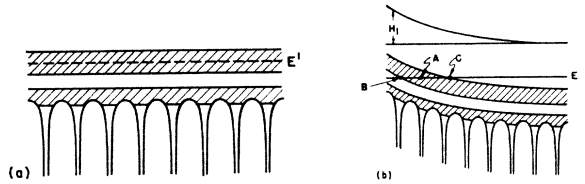

FIG. 1. (a) Periodic potential, with energy bands (shaded). (b) Energy bands and potential pushed upward by amount H_1 (shown in upper curve). Line of constant energy E cuts band, at point A, at same relative height as energy E' in Fig. 1(a). B and C, reversing points of oscillation of particle.

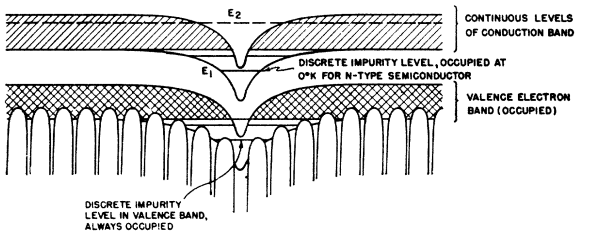

FIG. 2. Continuous and discrete levels surrounding impurity atom in N-type semiconductor. E_1, energy of a discrete level; E_2, energy in continuum.

of x, the difference giving the kinetic energy. When E, in Fig. 1(b), lies outside any of the energy bands, the kinetic energy is negative, the wave-length imaginary, and the wave function is damped off exponentially. When E is inside one of the bands, kinetic energy is positive, and the wave-length real. A classical particle of energy E, moving according to the classical Hamiltonian $E_0(p)+H_1(q)$, would then oscillate between points like B and C, Fig. 1(b), reversing at each point as its kinetic energy and momentum become zero, and a quantized particle will obey a quantum condition. Such a picture, as James has emphasized, allows us to deduce the nature of the stationary states and wave functions in such a problem. If, for instance, the external field represented by H_1 is constant, so that the energy bands are tilted at a constant angle, the electron will oscillate back and forth in coordinate space with a very large amplitude (for a small external field), at the same time having its position in the energy band go from bottom to top and back again, in a way familiar in the theory of electrical conduction.

Let us now consider the application of our theorem to cases actually met in the theory of solids. First we take P- and N-type impurity atoms. The N-type is a little easier to understand, and we deal with it first. It is usually an atom substituted for one of the atoms normally present in the lattice, and containing more valence electrons than the atom which it replaces; for definiteness, we may be considering an atom of P or As in a lattice of Si or Ge. If the atom loses an electron, it has enough remaining electrons to fit properly into the lattice, but it then carries a positive charge, which introduces a Coulomb potential (modified as to its absolute value by the dielectric constant of the material) into the lattice. Thus the energy bands, as modified by this Coulomb potential, will be as shown in Fig. 2. In these bands, we clearly have the possibility of discrete energy levels, of a hydrogen-like sort, at energies such as E_1. Just compensating these levels, which if they were occupied would introduce extra charge near the impurity, we see that with higher energies in the band, such as E_2, the electrons will effectively be repelled like positive charges as they approach too close to the impurity atom, their kinetic energy going to zero and the electrons being turned back at the point where the line at height E_2 emerges from the top of the energy band. Thus these other states will provide less than the normal charge near the impurity, so that, if all the states of a band are occupied, we shall still have just enough charge to account for one electron per atom of each spin in the band. In the N-type semiconductor, these modified levels are all occupied in the lower, or valence-electron band. We still have one electron per impurity atom left over, however, and this at low temperature will go into the lowest discrete level below the upper (or conduction) band, but at slightly higher temperatures will go into one of the conduction levels.

The P-type semiconductor is handled similarly. A P-type impurity atom normally contains one less valence electron than the atom of the lattice, e.g., an atom of Al or B in a lattice of Si or Ge. If we provide such an atom with an extra electron, to make it a negative ion, it has the right number of electrons to fit into the lattice. Then the modified energy bands will be as in Fig. 3, clearly giving discrete levels lying above the bands, with compensating diminished electron density in those wave functions lying at the bottom of the band. In the neutral crystal, there must be one less electron per impurity atom than the number necessary to fill the complete valence-electron band, so that at low temperatures this electron will be missing from the discrete level lying above the band; but at higher temperatures it will often be missing from one of the continuous levels lying in the band, leading therefore to hole conduction.

The case of excited energy levels in crystals, which has often been discussed (see, for instance, Wannier[1]), is more involved, but similar to these cases of impurities. It is, perhaps, easiest to understand in the case where a tightly bound, or x-ray, electron is excited to the conduction band.[7] If an atom of the crystal has lost one of its inner electrons, it acts approximately, so far as its valence electrons are concerned, like an atom with a nuclear charge greater by one unit; the missing electron can act like an additional valence electron. Thus the atom temporarily plays the role of an N-type impurity atom, and will set up discrete energy levels as in Fig. 2. When the corresponding emission spectrum is observed, resulting from an electron in the valence-

[7] The application of the theory to this case, the soft x-ray problem of Skinner and O'Bryan and other writers, is discussed by F. Seitz, *Modern Theory of Solids* (McGraw-Hill Book Company, Inc., New York, 1940) p. 436–438.

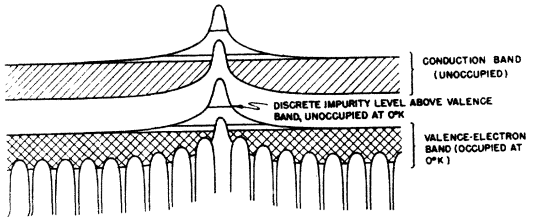

FIG. 3. Continuous and discrete levels surrounding impurity atom in P-type semiconductor.

electron band falling down into the empty x-ray level, there is the possibility that the electron may come from one of the levels in the continuum of the valence-electron band (the possibility considered by Skinner and O'Bryan, discussed in many references, some given in reference 7), but also the possibility that it may come from one of the discrete levels lying below the valence-electron band, resulting therefore in a longer wave radiation than we should otherwise find. Such tails are observed in the soft x-ray emission spectra, and Seitz suggests this interpretation of them.

The really optical case of excitation, where an electron is removed from the valence-electron band to a conduction band, is more complicated, in that both electron and hole are readily mobile. We may then best describe it essentially as Wannier did in reference 1, and as Frenkel had done earlier.[8] Considering it classically, the electron in the conduction band attracts the hole in the valence-electron band, and, since each has a comparable effective mass, they each execute hydrogen-like orbits about their center of mass, resulting in certain discrete states. These discrete states lie below the continuous states; that is, the electron effectively is at the bottom of the conduction band, like the discrete states in an N-type semiconductor as in Fig. 2, while the positive hole is at the top of the valence electron band, like the discrete states in a P-type semiconductor as in Fig. 3. The electron and hole together form a stable structure, however, which, because of the mobility of both electron and hole, is free to wander through the crystal, forming what has been called an exciton (see reference 8). Being a neutral structure, it carries no current. Less energy is required to set up such an exciton than to raise an electron from the valence-electron band to the conduction band, leaving both electron and hole dissociated from each other and free to move, so that the wave-length for absorption to this exciton level, which does not result in photo-conductivity, is longer than for the limit of photo-conductivity. It is well known that such exciton levels exist, for instance, in the alkali halides.[9]

[8] J. Frenkel, Phys. Rev. **37**, 17 (1931); **37**, 1276 (1931); Physik Zeits. Sowjetunion **9**, 158 (1936).

[9] See, for instance, J. C. Slater and W. Shockley, Phys. Rev. **50**, 705 (1936), in which the theory of the exciton is considered without benefit of Wannier's theorem.

IV. STATISTICAL TREATMENT OF PERTURBED PERIODIC LATTICES

A study of the stationary states of the electrons in the perturbed lattice, such as we have been making in the preceding section, is really only half the problem; we wish, as well, to ask which levels will be occupied, which ones empty. In thermal equilibrium, which is the only case we shall consider, we must then supplement our theory by use of the Fermi-Dirac statistics: the average number of electrons in a state of total energy E, with a given spin, is $1/\{\exp[(E-E_F)/kT]+1\}$, where E_F is the electrochemical potential or Fermi level. From this fact, supplemented by the knowledge of the wave functions Ψ_n, we can find the average charge density at each point of the lattice. It is by no means necessarily true that this charge density will automatically come out zero; hence we have space charge, and from this space charge we can compute an electrostatic potential by Poisson's equation. We can then apply a self-consistent condition, in essentially the sense of Hartree: we can demand that the potential energy of an electron in this potential be the same quantity $H_1(q)$ which is responsible for perturbing the energy bands.

To set up our self-consistent condition, we must first find the net charge density as a function of position arising from our assumed energy bands with the assumed Fermi level. First we shall find this in the case of the unperturbed periodic potential. In this potential, let the number of energy levels per unit volume, in the energy range dE, be $n(E)dE$; this can be found, as is well known, from the volume of momentum space lying between surfaces $E_0(p)=E$ and $E_0(p)=E+dE$, since states are distributed in momentum space with uniform density in the periodic case. In forbidden bands of energy, $n(E)$ is, of course, equal to zero. Let N_0 be the number of electrons per unit volume necessary to render the crystal electrically neutral. Then the excess number of electrons per unit volume, with an arbitrary value of E_F, is

$$N(E_F)=\int_{-\infty}^{\infty} n(E) \times\{\exp[(E-E_F/kT)+1]\}^{-1}dE-N_0. \quad (13)$$

Ordinarily, we determine E_F by the condition that N must be zero, or the lattice uncharged, so that we set $N(E_F)$ in (13) equal to zero. Our present problem, however, is different: we wish to investigate the results of volume charge in the lattice, resulting from N being not zero but, instead, a slowly varying function of position. If we have such a volume charge, then the potential in the lattice will differ from its periodic value by a slowly varying function determined from the volume charge by Poisson's equation. We let H_1 be the potential energy of an electron in this slowly varying potential, as before. Since the charge density is $-Ne$, where e is the magni-

tude of the electronic charge, $\nabla^2 H_1 = -Ne^2/\epsilon$, where ϵ is the permittivity of the material, or the dielectric constant times ϵ_0.

In the presence of the slowly varying electrostatic potential, $n(E)$ will, of course, no longer be given as it was previously. We know from the preceding sections, however, that the effect of H_1 will be to push up the energy bands with respect to their original position by an amount equal to the local value of H_1. It seems reasonable that the number of excess electrons per unit volume will then be given by $N(E_F - H_1) = N(\zeta)$, where the function N is as defined in (13), and where we introduce the abbreviation $\zeta = E_F - H_1$. This is an assumption much like that made in the well-known Thomas-Fermi method of discussing atomic structure, where we assume at every point of space that the statistical distribution of electron energies is what would be found for free electrons moving in a constant potential equal to the local value of the actual potential. It takes no specific account of the discrete energy levels, but merely handles them in a statistical or averaged-out way. Our method differs from the Thomas-Fermi method in three respects: we are handling our kinetic energy by the energy-band method, so that it is given by $E_0(p)$ instead of the usual expression; we are dealing with a modulating function $\Psi_n(q)$, instead of with the actual wave function $\psi(q)$; and we are handling our statistics in a form appropriate to an arbitrary temperature, rather than for the absolute zero of temperature as is done in the ordinary Thomas-Fermi method.

When we make the assumption above, we can write Poisson's equation in the following form:

$$\nabla^2 \zeta = N(\zeta) e^2/\epsilon, \tag{14}$$

where we have used the fact that E_F must be constant over-all space, to satisfy the condition for thermal equilibrium in the Fermi statistics, so that its Laplacian is zero. In Eq. (14), supplemented by Eq. (13) for the function N, we have the general formulation of the problem of setting up the electrostatic potential within a solid in thermal equilibrium. This equation has, of course, been used and solved in special cases by many writers. It has essentially been used by Schottky[10] in an extensive series of papers, and is similarly used by Mott and Gurney, and by Bethe.[11] Fan[12] has carried out careful studies of the contact between metals and between a metal and a semiconductor, which are complete enough so that many of our results will be merely a restatement of some of Fan's conclusions. Quite recently, Markham and Miller[13] have used essentially similar methods in closely related problems. Many other writers are aware of these methods of handling the problem. To give a complete picture, we shall state some of the methods of solving this equation, and some of the applications to well-known cases, as well as some new aspects of the problem.

The nature of the solution depends on the form of the function $N(\zeta)$. In Fig. 4 we show this function for two familiar cases: the metal and the intrinsic semiconductor. In the first case, N increases very rapidly as ζ departs from the value appropriate for no charge, and over a considerable range it can be treated as proportional to $\zeta - \zeta_0$, where ζ_0 is the value associated with no charge. In the semiconductor, however, N increases very slowly with $\zeta - \zeta_0$, behaving approximately as a hyperbolic sine, although when ζ becomes so large, negatively or positively, that the Fermi level penetrates either the lower valence-electron band or the upper conduction band, N begins to get very large, negatively or positively as the case may be. If $N(\zeta)$ can be approximated as $a(\zeta - \zeta_0)$ (where a is a constant) as it can over a considerable range in these two cases, then Eq. (14) takes on the mathematical form of the wave equation, and solutions can be set up by familiar methods. Thus, if we are dealing with a one-dimensional problem,we have solutions $\zeta - \zeta_0 = \exp(\pm x/X)$, where $X = (\epsilon/ae^2)^{\frac{1}{2}}$, and where x is the coordinate in the direction in which the potential is changing. As shown by Fan,[12] this quantity X for a metal is very small—of the order of magnitude of an angstrom unit; on the other hand, for an intrinsic semiconductor it becomes large and in the limit of zero temperature for this case it becomes infinite. For a three-dimensional problem of spherical symmetry, we similarly have $\zeta - \zeta_0 = \exp(\pm r/X)/r$, where r is the distance from the center, and X is as given above.

We can now examine several applications of these simple results. First we consider the metal, and the one-

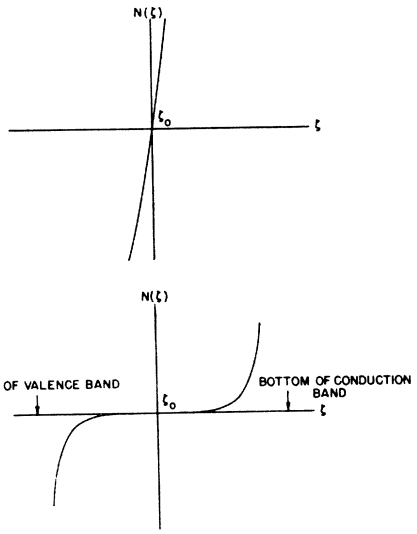

FIG. 4. $N(\zeta)$ as function of ζ. Above, metal; below, intrinsic semiconductor.

[10] W. Schottky, Zeits. f. Physik **118**, 539 (1941); other references quoted in this paper.

[11] N. F. Mott and R. W, Gurney, *Electronic Processes in Ionic Crystals* (Oxford University Press, New York, 1940), Chapter V; H. A. Bethe, M.I.T. Radiation Laboratory Report 43-12 (November 23, 1942).

[12] H. Y. Fan, Phys. Rev. **62**, 388 (1942).

[13] J. J. Markham and P. H. Miller, Jr., Phys. Rev. **75**, 959 (1949).

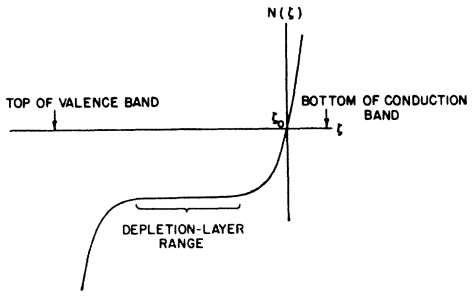

FIG. 5. $N(\zeta)$ *vs.* ζ for N-type semiconductor.

dimensional problem; this is met in investigating the surface charge at the free surface of a metal in an external electric field. Clearly, we can adjust boundary conditions so that the electric field at the surface resulting from the gradient of H_1 has any desired value, and yet the resulting potential will penetrate into the metal only to a distance of the order of a single atomic layer, with the related charge density confined to this same small depth. This allows us, in other words, to assume an arbitrary surface charge on the surface layer of atoms of a metal, of suitable amount to match any external boundary conditions. Similarly, if two metals are close to each other and connected electrically, so that they must bear surface charges enough to produce the difference of potential equal to their difference of work function, these surface charges will be formed according to this same method; and as the metals are brought into contact, the double layer between them is formed from surface charges of the same variety at the surfaces of each. This description of the double layer has been worked out in detail by Fan.[12] Another example relating to the metal comes from the spherically symmetrical problem. If we had an impurity atom in a metal, of the type which we have in a semiconductor, and which we discussed in Section II, it would produce a local singularity in the potential. The solution which we should have to use would then be of the form $\zeta=\zeta_0$ $=$constant $\exp(-r/X)/r$, showing a suitable singularity at $r=0$, but decreasing exponentially to zero in a distance of atomic dimensions. In other words, the conduction electrons would shield the impurity atom so completely that it would not produce appreciable perturbation of potential beyond its nearest neighbor atoms. This, of course, is a well-known result.

In an intrinsic semiconductor, we may consider these same two problems, remembering that here X is very large. This means that in such a material, which is practically an insulator, we can accumulate a practically negligible volume charge in the interior, so that if the whole material (including the surface layers) behaves in the same way, we cannot have a thin surface layer of charge as we can in a metal. Instead, if we have such an insulator in an external electric field normal to the surface, the field penetrates the surface, the normal component of D being continuous as in the usual theory of dielectrics. To account for surface charges which unquestionably can build up on the surface of a dielectric, as, for instance, by bombarding with electrons, which have no chance to leak off, we must introduce surface states, capable of holding extra charge; we postpone discussion of such surface states to the next section. In the interior of an intrinsic semiconductor,we may use the spherical solution of our equation to discuss an impurity atom; and we find, with our large X, that the field is essentially an inverse square field, the effect of the dielectric being seen only in the dielectric constant. Thus we have correctly drawn our perturbed energy bands around impurity atoms, in Figs. 2 and 3, as though the potential varied inversely as the distance from the impurity center, without the type of shielding found in the metal.

V. IMPURITY SEMICONDUCTORS AND RECTIFYING BARRIERS

The method of treatment we have used in the preceding section handles the action of impurity atoms on a microscopic scale, asking how the potential behaves around each impurity atom; Section II handled similarly the energy levels of such impurity atoms on the same microscopic scale. In treating impurity semiconductors, however, it is usually more convenient to treat average behavior over a volume which is small compared to the thickness of a rectifying barrier, but large compared to the distance between impurity atoms; or, alternatively, it is better somehow to average the impurity levels, so that we do not have to consider the fine-grained inhomogeneities arising from the discrete impurity atoms. When we do this, we have a different distribution of energy levels, $n(E)$, for now we include an appropriate number of levels per unit volume arising from the discrete levels in N- or P-type impurity atoms, and hence located just below or just above the continuous bands. Also in computing the net amount of charge per unit volume, we must take account of

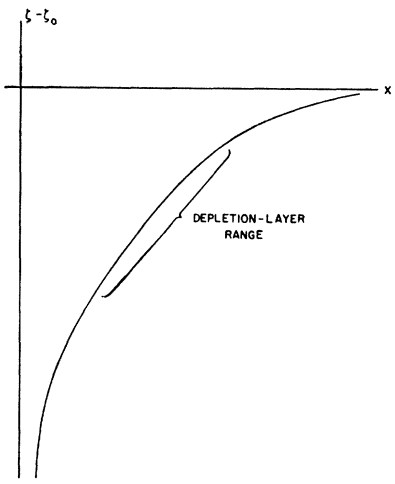

FIG. 6. $\zeta-\zeta_0$ *vs.* x, for N-type semiconductor.

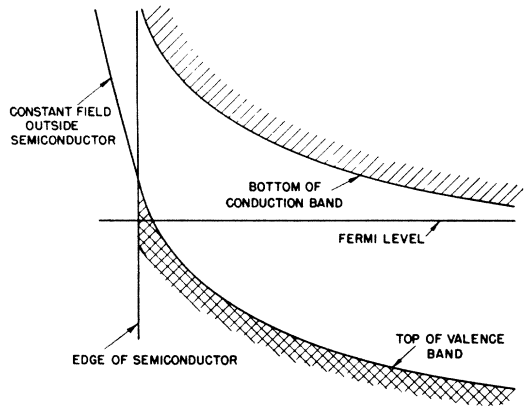

FIG. 7. Energy bands at boundary of N-type semiconductor, large external electric field, but no double layer.

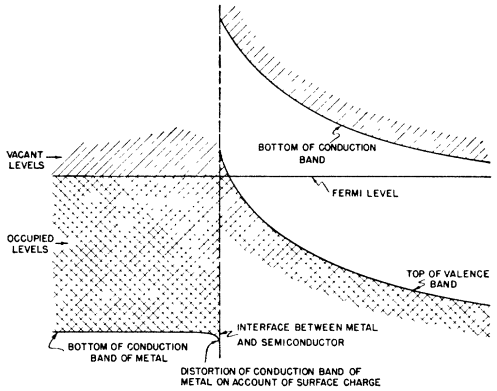

FIG. 8. Energy bands at interface between metal and semiconductor, no surface states.

the charge furnished by these impurity ions. When we do this, we may get for $N(\zeta)$ for, say, a material containing N-type impurity atoms uniformly distributed, a curve of the nature shown in Fig. 5.

In Fig. 5, there are shown several distinct regions, with different behavior of $N(\zeta)$ in each. First, for ζ less than ζ_0, there is a long region in which N is negative and approximately constant. This is the range in which the impurity atoms have lost their extra electrons, so that they yield a positive space charge. This is the region in which Schottky's depletion-layer theory is appropriate. As ζ decreases still further, there is a very rapid and large decrease in N; this arises when ζ is so low that it begins to empty the levels in the valence-electron band. In this region, the material would act like a P-type semiconductor, and with still further decrease of ζ it would show metallic properties, whereas in the depletion-layer region there are practically no holes in the lower band, or electrons in the upper band, so that the material acts like an intrinsic semiconductor, or practically as an insulator. Proceeding in the other direction, of increasing ζ, we find that N is zero when $\zeta=\zeta_0$, and beyond that point N starts to increase very rapidly. This is the range where there are enough electrons to start filling the conduction band. Of course, even with $N=0$, there are some electrons in the conduction band, raised by thermal agitation from the donator impurity atoms; but this number very rapidly increases with increase of ζ, so that the material becomes a much better conductor, and soon acquires metallic properties.

With as complicated a function $N(\zeta)$ as is given in Fig. 5, it is clear that the simple approximation used previously (of setting it proportional to $\zeta-\zeta_0$) is inadequate, and we must use the whole form of the function. Even in this case, in the one-dimensional problem, we can integrate Poisson's equation, (14), as is done, for instance, by Mott and Gurney[11] and by Fan.[12] The equation becomes $d^2\zeta/dx^2=f(\zeta)$, where $f(\zeta)$ is a function of ζ. Mathematically, this is similar to a one-dimensional equation of motion in mechanics, $md^2x/dt^2=f(x)$, where $f(x)$ is the force; and, as in the mechanical case, we can integrate by a method entirely equivalent to the energy integral in mechanics. When we do this, carrying out the integrations numerically if $N(\zeta)$ is as complicated as in Fig. 5, we can find the relation between ζ and x. We find that there are solutions approaching ζ_0 asymptotically for large values of x, one for ζ greater than ζ_0, the other for ζ less than ζ_0; all such solutions differ from each other only by uniform translation along the x-axis. In Fig. 6 we show such a solution for the case coming into the problem of a rectifying barrier at the surface of an N-type semiconductor.

There are sections of the curve of Fig. 6 which are associated with the various parts of Fig. 5. As ζ departs only slightly from ζ_0, we are in a linear part of Fig. 5. Here we have the Laplacian of $\zeta-\zeta_0$ equal to a constant times $\zeta-\zeta_0$, so that we can set up an exponential solution for $\zeta-\zeta_0$ as a function of x, just as we did with the metals. We find, however, a much slower exponential variation, extending over much greater distances. Next we have the region where ζ is considerably less than ζ_0, so that we are in the region of the depletion layer. In this region, as we have mentioned earlier, $N(\zeta)$ is practically constant, and our solution agrees exactly with Schottky's,[10] leading to ζ as a quadratic function of x. This parabolic function, of course, fits smoothly to the exponential function which holds as ζ approaches its asymptotic value ζ_0. Finally, when ζ gets so small that electrons of the lower band begin to be removed, ζ begins to decrease very rapidly with change of x. This is because we are now meeting high positive volume charges, arising from the emptying of this lower band, and the electric field can change with position about as rapidly as in a metal (where we have already seen that we can accommodate a large enough charge in a layer of atomic thickness to be equivalent to a surface charge). This happens here as ζ penetrates into the lower band, and the reason for its happening is essentially the large reservoir of charge available, similar to the case of a metal. In the nature of things, however, this surface

layer of charge can only be positive, since it arises from holes in the valence-electron band. If there is a negative surface charge, the curve corresponding to Fig. 6 must rise rather than fall at the surface of the material.

The result shown in Fig. 6 can now be used to discuss the boundary between a metal and an N-type semiconductor, or between such a semiconductor and a vacuum. If we used this result straightforwardly, we should draw the following conclusions. We should conclude that at the boundary between a semiconductor and a vacuum, in the absence of an external electric field, the potential in the semiconductor would be constant. If an external field were impressed, then the surface charge to terminate the lines of force would be actually distributed through the whole depth of a depletion layer, instead of being located on the surface, as in a metal. The only exception would come if the field were so strong that the Fermi level dipped down into the valence-electron band at the surface; then any remaining charge required to terminate the lines of force would be located almost exactly at the surface. In other words, the total amount of charge which can be distributed through the interior of the volume in the depletion layer is limited. In such an extremely large field, the energy bands would look as in Fig. 7, and the external field would be indicated by the slope of the curve to the left of the solid. If now the semiconductor were placed in contact with a metal between which the difference of work function was so great as to require on the semiconductor a positive charge, and on the metal a negative charge, great enough to produce a field of the magnitude shown in Fig. 7 in the double layer between the two materials, then we should find the situation shown in Fig. 8. Here the Fermi level would come slightly below the top of the valence-electron band of the semiconductor just at the surface, and the resulting situation would be almost independent of the work function of the metal, provided only that it was different enough from that of the semiconductor to require a large enough double layer.

It is well known that the situation we have just described does not fit the observations, at least in germanium and silicon, two semiconductors which are very well understood as a result of the large amount of work done on them during the war and since at Purdue University, the Bell Telephone Laboratories, the University of Pennsylvania, and elsewhere. Meyerhof[14] in a set of measurements on contact difference of potential between silicon and metals found definite evidence that our simple picture is wrong, and his effect was explained by Bardeen[15] with his theory of surface states. Since then, the group at the Bell Telephone Laboratories has arrived at substantially the following conclusions regarding the surface of germanium, explainable in terms of surface states.[16] At a free surface between germanium and air, there is good evidence that in the absence of an external field there is, nevertheless, a well-formed positively charged depletion layer below the surface, compensated by an equal negative surface charge. In an external electric field, the extra surface charge required to terminate the lines of force appears just on the surface, rather than in the depletion layer in the interior. And when contact is made between the semiconductor and a metal, the required double layer adjusting the Fermi levels of the two to coincidence is made up of a surface charge of the usual sort at the surface of the metal, and a surface charge of opposite sign on the surface of the germanium.

It thus appears that the surface layer of atoms on a germanium crystal must behave differently from the interior and, to explain this, Bardeen introduces the idea of surface states. The action can be described, roughly, as if there were part of a monomolecular layer of metal on the surface of the germanium, whose work function differed from that of a hypothetical germanium which lacked the surface states by something like the amount considered in Fig. 8. This layer of metal would have to acquire enough negative charge to raise its Fermi level—normally far below that of the germanium —up to equality with that of the germanium. Having the large reservoir of electrons characteristic of a metal, it could acquire any other amount of surface charge necessary to compensate for an applied external field, or for a double layer arising when another metal made contact with it. Thus the rectifying barrier would remain something like that of Fig. 8, which is not unlike what is observed (see, for instance, reference 16), independent of the material of the metal making contact.

In this description of the surface, which is substantially that suggested by the group at the Bell Telephone Laboratories, it is not clear whether the surface states arise from real impurities on the surface (either metallic or, at any rate, setting up a distribution of states of the sort characteristic of a metal), or whether they are inherent in the germanium itself. The evidence indicates that the surface states depend on surface conditions, suggesting impurities, and certainly it is very difficult to get the surface really clean. Furthermore, the absence of surface conductivity suggests that the surface states are localized at widely separated spots on the surface, as if they arose from impurities. On the other hand, the photoelectric experiments of Apker, Taft, and Dickey,[17] performed on the cleanest surfaces obtainable, gave evidence of the same sort of situation observed in ordi-

[14] W. E. Meyerhof, Phys. Rev. **71**, 727 (1947).
[15] J. Bardeen, Phys. Rev. **71**, 717 (1947).

[16] W. H. Brattain and W. Shockley, Phys. Rev. **72**, 345 (1947); W. H. Brattain, Phys. Rev. **72**, 345 (1947); J. Bardeen and W. H. Brattain, Phys. Rev. **74**, 230 (1948); W. H. Brattain and J. Bardeen, Phys. Rev. **74**, 231 (1948); W. Shockley and G. L. Pearson, Phys. Rev. **74**, 232 (1948); J. Bardeen and W. H. Brattain, Phys. Rev. **75**, 1208 (1949). The author is indebted to Dr. Bardeen for an opportunity to see this latter paper before its appearance in print.
[17] Apker, Taft, and Dickey, Phys. Rev. **73**, 46 (1948).

nary germanium, suggesting that the situation may be inherent in a clean germanium surface. This is not impossible: the spacing between the surface layer of germanium atoms and the next layer below it may well be different from that between layers in the interior because of the unbalanced forces near the surface; this would bring about a distortion of the energy bands near the surface, quite aside from anything we have considered, which might have the effect of making the surface layer of atoms behave quite differently from the interior. At any rate, it seems to be empirically clear that we must treat the surface layer of such a crystal as a different material from the interior; hence, to apply the arguments of the present theory to the boundary layer, we must consider the interior and the surface separately, and consider the boundary conditions at the interface between them, as well as at the interface between the surface and air or another conductor.

VI. ACKNOWLEDGMENTS

Most of the material described in this paper was prepared for presentation at the M.I.T. Conference on Physical Electronics in April 1949, and at a preliminary session before it, and the writer is indebted to Professor W. B. Nottingham for the stimulus to prepare it, and for valuable discussion. He is also indebted to Professor K. Lark-Horovitz, Professor H. M. James, and Dr. H. Y. Fan, of Purdue University, for valuable comments and much information about the semiconductor program at Purdue; and to Dr. J. Bardeen, Dr. W. H. Brattain, and Dr. W. Shockley, of the Bell Telephone Laboratories, for valuable comments and suggestions.

APPENDIX I

We start with the functions $\psi_0(p, q)$, satisfying Eq. (3). From them we form the functions

$$a(q-Q_k)=N^{-\frac{1}{2}}\Sigma_p\{\exp[(-i/\hbar)p\cdot Q_k]\}\psi_0(p, q), \quad (1A)$$

where the sum is over all the quantum states p. These functions are the normalized atomic functions of Wannier, written in our notation; we refer to reference 1 for their properties. Equation (2) follows at once from (1A) by multiplying by $\exp[(i/\hbar)p'\cdot Q_k]$, summing over k, and using the theorem

$$\Sigma_k \exp[(i/\hbar)(p'-p)\cdot Q_k]=0 \quad \text{if} \quad p'\neq p,$$
$$=N \quad \text{if} \quad p'=p.$$

Now we set up the solution (5) for Eq. (4). In (4) we write $\psi_n(q)$ in the form (5), multiply both sides of the equation by $a^*(q-Q_m)$, and integrate over q, obtaining the equation

$$0=\Sigma_k\int a^*(q-Q_m)(H_0+H_1-E_n)\Psi_n(Q_k)a(q-Q_k)dq. \quad (2A)$$

We remember that H_0 is an operator operating on a function of q, H_1 is a slowly varying function of q, and E_n is a constant. On account of the orthogonality of the a's, proved by Wannier, $\int a^*(q-Q_m)a(q-Q_k)dq=0$ if $m\neq k$, 1 if $m=k$. Thus the term in E_n reduces to $-E_n\Psi_n(Q_m)$. So far as the term in H_1 is concerned, let us assume that H_1 is so slowly varying with q that it can be regarded as approximately constant over the atomic wave function of an atom. Then, again using the orthogonality of the a's, this term reduces to $H_1(Q_m)\Psi_n(Q_m)$. If H_1 varies more rapidly, we can use the deviations from this result as a starting point for a higher order of approximation.

The other term of (2A), the one in H_0, must be handled differently. We rewrite $a^*(q-Q_m)$ and $a(q-Q_k)$, as they appear in (2A), by using (1A), and make use of the fact that the ψ_0's satisfy Eq. (3). Then we have

$$\Sigma_k\int a^*(q-Q_m)H_0\Psi_n(Q_k)a(q-Q_k)dq$$
$$=\sum_{k,p,p'}(1/N)\Psi_n(Q_k)\{\exp[(i/\hbar)(p'\cdot Q_m-p\cdot Q_k)]\}$$
$$\times\int\psi_0^*(p', q)H_0\psi_0(p, q)dq.$$

We use (3), and the orthogonality of the functions ψ_0, for different p's, to show that $\int\psi_0^*(p', q)H_0\psi_0(p, q)dq=E_0(p)\delta(p', p)$. When we substitute this above, we have

$$\Sigma_k\int a^*(q-Q_m)H_0\Psi_n(Q_k)a(q-Q_k)dq$$
$$=\sum_{k,p}(1/N)\Psi_n(Q_k)E_0(p)\exp[(i/\hbar)p\cdot(Q_m-Q_k)]$$
$$=\sum_{s,p}(1/N)\Psi_n(Q_m-Q_s)E_0(p)\exp[(i/\hbar)p\cdot Q_s], \quad (3A)$$

where we have substituted $Q_m-Q_k=Q_s$.

We can now rewrite this expression (3A). We recall that $E_0(p)$ is a periodic function in the p-space, having the periodicity of the reciprocal lattice. Thus it can be written in the form

$$E_0(p)=\sum_{Q_k}A(Q_k)\exp[(-i/\hbar)p\cdot Q_k], \quad (4A)$$

where the Fourier coefficients $A(Q_k)$ are given by

$$A(Q_k)=\Sigma_p(1/N)E_0(p)\exp[(i/\hbar)p\cdot Q_k].$$

Thus the right-hand side of (3A) can be expressed in the form

$$\Sigma_s A(Q_s)\Psi_n(Q_m-Q_s). \quad (5A)$$

Now we expand $\Psi_n(Q_m-Q_s)$ in Taylor's series about the point Q_m. We have

$$\Psi_n(Q_m-Q_s)=\Psi_n(Q_m)-\frac{d}{dq}\Psi_n(Q_m)(Q_s)$$
$$+\frac{1}{2!}\frac{d^2}{dq^2}\Psi_n(Q_m)(Q_s^2)+\cdots$$
$$=\left[\exp\left(-Q_s\frac{d}{dq}\right)\right]\Psi_n(Q_m). \quad (6A)$$

In this expression we have written only the case where q is a scalar quantity, but an exactly analogous form holds if it is a vector, $Q_s(d/dq)$ being replaced by the scalar product $Q_s\cdot\nabla$, where the ∇ operator denotes vector differentiation with respect to the components of the vector q. We now use the result (6A) to modify (5A), and it takes on the form

$$\Sigma_s A(Q_s)[\exp(-Q_s\cdot\nabla)]\Psi_n(Q_m).$$

Comparison with Eq. (4A) shows that this is what we should get if we took $E_0(p)$, replaced p in it by the operator $(\hbar/i)\nabla$, and allowed this to operate on $\Psi_n(Q_m)$. When we combine this with the expressions for the terms in H_1 and E_n in Eq. (2A) which we have already discussed, we see that we are led to Eq. (6), which we wished to prove.