Theory of Electron Tunneling in Semiconductor Junctions*

D. R. FREDKIN†
Palmer Physical Laboratory, Princeton University, Princeton, New Jersey

AND

G. H. WANNIER
Physics Department, University of Oregon, Eugene, Oregon

(Received June 13, 1962)

A formalism is developed for calculating the electric current in a semiconductor junction of the Esaki type due to direct tunneling of electrons between the valence and conduction bands. This formalism expresses the current at any temperature and bias in terms of a tunneling amplitude, whose absolute square is essentially the probability that an electron incident on the junction from either side will be seen eventually on the other side. The methods of time-independent perturbation theory are used to calculate the tunneling amplitude for a model in which the junction electric field is piecewise constant and the effective-mass approximation is applicable. The result of this calculation is used to compute explicitly the tunneling current as a function of bias and temperature. The formalism also yields a general variational principle for the tunneling amplitude.

I. INTRODUCTION

THE study of interband effects in the motion of crystal electrons in an external electric field has been of interest almost as long as the existence of bands has been understood.[1–3] In the case of a uniform electric field, it is possible to show that interband effects can be rigorously eliminated by a suitable redefinition of the bands which allows for polarization effects.[4,5] Thus, in a sense, there is no question of interband transitions in a uniform field.

In recent years, with the development of the Esaki diode, interest has turned to the behavior of crystal electrons in an inhomogeneous external electric field, and specifically to the phenomena which arise in a narrow, heavily doped p-n junction of the Esaki type.[6–14] Figure 1 illustrates such a junction schematically in the familiar deformed band picture. We see that a conduction electron on the n side of the junction may be energetically capable of existing as a valence electron on the p side, and conversely. However, in order to make such a transition, an electron must, in the deformed-band picture, traverse the forbidden gap between the valence and conduction bands.

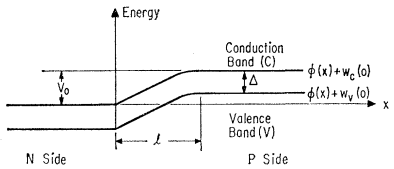

FIG. 1. Deformed band diagram of an Esaki junction.

In order to have a definite physical question to answer, we may set ourselves the task of computing the current density due to interband transitions. A part of this current originates in direct transitions, without benefit of phonon cooperation, intermediate impurity states, or the like, and it is to this part that we confine ourselves in what follows. The existence of such direct tunneling has been confirmed experimentally.[15,16]

In the past, calculations of this type have either assumed a homogeneous field or have relied on the WKB approximation. The homogeneous field calculations[7–10] are irrelevant because they are intrinsically incapable of describing a steady state of charge transport. Although, owing to certain approximations which we make, and which we discuss in Sec. V below, there are similarities between the homogeneous field calculations and the dynamical part of the present work, the intimate interlocking of dynamics and statistics in our approach prevents comparison with the results of the homogeneous field calculations. The recent work of Kane,[13] which adapts the homogeneous field type of calculation to a calculation of the current, is an exception. However, Kane's method is quasiclassical, and is therefore much less powerful than ours. Our specific example, which involves approximations whose domain of validity is essentially the same as that of the quasiclassical method, gives a result which would agree with that of Kane if he had not made an error in his calculation. This will be discussed more thoroughly in Sec. VI

* This work was begun while both authors were at the Bell Telephone Laboratories, Murray Hill, New Jersey.

† Present address: Physics Department, University of California, San Diego, La Jolla, California.

[1] C. Zener, Proc. Roy. Soc. (London) **145**, 523 (1934).
[2] W. V. Houston, Phys. Rev. **57**, 184 (1940).
[3] P. Feuer, Phys. Rev. **88**, 92 (1952).
[4] G. H. Wannier, Phys. Rev. **117**, 432 (1960).
[5] D. R. Fredkin and G. H. Wannier, Bull. Am. Phys. Soc. **6**, 108 (1961); and Phys. Rev. **125**, 1910 (1962).
[6] L. Esaki, Phys. Rev. **109**, 603 (1958).
[7] L. V. Keldysh, Soviet Phys.—JETP **6**, 763 (1958).
[8] L. V. Keldysh, Soviet Physics—JETP **7**, 665 (1958).
[9] E. O. Kane, J. Phys. Chem. Solids **12**, 181 (1959).
[10] W. Franz, Z. Naturforsch. **14a**, 415 (1959).
[11] P. J. Price and J. M. Radcliffe, IBM J. Research and Develop. **3**, 364 (1959).
[12] P. N. Argyres and B. Lax, Bull. Am. Phys. Soc. **6**, 105 (1961).
[13] E. O. Kane, J. Appl. Phys. **32**, 83 (1961).
[14] "Symposium on Electron Tunneling in Solids, January 30, 31, 1961" [J. Phys. Chem. Solids (to be published)].
[15] E. O. Kane and J. V. Morgan, Phys. Rev. Letters, **3**, 466 (1959).
[16] A. G. Chynoweth and R. A. Logan, Phys. Rev. **118**, 1470 (1960).

below. The WKB type calculations[12,14] seem to be of doubtful validity in that they rely on an *ad hoc* matching procedure of wave functions which are analogous, from the present point of view, to modulations of distinct modes of a waveguide.

In Sec. II we define a complete, linearly independent set of "tunneling states," which are eigenstates of the full Hamiltonian which have asymptotic properties analogous to those used in the usual quantum theory of scattering. Although these states may not be mutually orthogonal, a detailed analysis of the motion of wave packets clearly defines their interpretation. The crucial quantity turns out to be an amplitude $\mathcal{T}$ which is analogous to the usual scattering amplitude; except for a kinematic factor $|\mathcal{T}|^2$ is the probability of tunneling through the junction in either direction.

In Sec. III we dispose of the statistical aspects of the current. A formula is given which requires only knowledge of the dynamical quantity $\mathcal{T}$ to be a completely known expression for the current.

In Sec. IV formalism is developed for calculating $\mathcal{T}$ using a division of the complete Hamiltonian into an in-band part and an interband part. $\mathcal{T}$ is expressed in terms of an amplitude $\mathcal{A}$, for which a stationary expression is given. However, when the simplest trial functions are used in this variational principle, it reduces to a well-known form which was implicit in an earlier approach to the tunneling problem[14] and which has also been suggested from the point of view of time-dependent perturbation theory.[14,17]

In Sec. V, the formalism is applied to the simple case of a piece-wise constant external field, neglecting the influence of all but the valence and conduction bands. An explicit formula, (92), is obtained for the current density at any bias and at any temperature such that the electrons are degenerate. In the limit of zero temperature, this formula yields the curve shown in Fig. 2.

In Sec. VI, some improvements and extensions of the techniques of this paper are indicated. Some of these are purely arithmetic, involving approximations which can be avoided if one seeks only numerical answers from a computer. On the other hand, it is possible to use our methods for the treatment of more complex "indirect" tunneling processes.

II. STATIONARY STATES AND THE MOTION OF WAVE PACKETS

For dynamical purposes, the tunneling problem is one dimensional, and we shall so treat it. The trivial modifications necessary for the three-dimensional case will be discussed in Sec. III. Let us first define some notation: Bloch functions are normalized in a unit cell,

$$\int_{\Omega_0} |b_r(x,k)|^2 dx = 1. \tag{1}$$

[17] J. Bardeen, Phys. Rev. Letters **6**, 57 (1961).

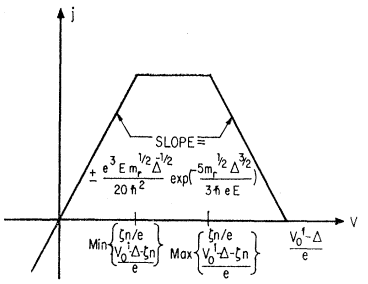

FIG. 2. Tunneling current density j as a function of bias V at zero temperature.

For each band, the Wannier functions[18] are defined by

$$a_r(x-\rho) = \int_Z e^{-ik\rho} b_r(x,k) \frac{\Omega_0 dk}{2\pi}, \tag{2}$$

where Z is the Brillouin zone, ρ is a lattice translation, and Ω_0 is the volume of the unit cell; the Wannier functions are correctly normalized

$$\int |a_r(x-\rho)|^2 dx = 1. \tag{3}$$

Also, (2) can be inverted to yield

$$b_r(x,k) = \sum_\rho e^{ik\rho} a_r(x-\rho). \tag{4}$$

We shall denote by $Z^\pm$ the positive (negative) half of the Brillouin zone.

For each energy $\mathcal{E}$ such that $w_c(0) \leq \mathcal{E} \leq w_v(0) + V_0$, define the stationary state $B(x,\mathcal{E})$ by requiring that

$$\begin{aligned} B(x,\mathcal{E}) &\sim b_c(x,k) + \mathcal{R}(\mathcal{E}) b_c(x,-k), \quad x \to -\infty, \\ B(x,\mathcal{E}) &\sim \mathcal{T}(\mathcal{E}) b_v(x,k'), \quad x \to +\infty. \end{aligned} \tag{5}$$

Here $k>0$, $k'<0$, $w_c(k) = \mathcal{E} = w_v(k') + V_0$. $w_{c,v}(k)$ are the energies of $b_{c,v}(x,k)$ in the absence of an external electric field, and V_0 is the total potential diffrerence across the junction (see Fig. 1). Here and in the following, we use units for which $\hbar = 1$. For simplicity, we assume that $w_c(k)$ $[w_v(k)]$ has only one extremum inside the Brillouin zone, which is a minimum (maximum) at $k=0$, and that both bands are nondegenerate. If we recall that in a stationary state the net flux departing at infinity must be zero, we immediately obtain an "optical theorem"

$$|\mathcal{R}(\mathcal{E})|^2 + |\mathcal{T}(\mathcal{E})|^2 [v_v(\mathcal{E})/v_c(\mathcal{E})] = 1. \tag{6}$$

Here $v_v(\mathcal{E}) = dw_v(k')/dk' > 0$, and $v_c(\mathcal{E}) = dw_c(k)/dk > 0$.

Although the $B(x,\mathcal{E})$ are not a complete set of states, they are sufficient for the study of the motion of a wave packet which, for $t \to -\infty$, represents a conduction electron incident on the junction from the left with average energy less than $w_v(0) + V_0$. It turns out that the probability of reflection, which is the probability that $x<0$ for $t \to +\infty$, is $|\mathcal{R}(\mathcal{E})|^2$, and the probability

[18] G. H. Wannier, *Elements of Solid-State Theory* (Cambridge University Press, New York, 1959), pp. 173–174.

of tunneling, which is the probability that $x>l$ for $t\to+\infty$, is $|\mathcal{T}(\mathcal{E})|^2 v_v(\mathcal{E})/v_c(\mathcal{E})$.

As has already been pointed out, our set $B(x,\mathcal{E})$ of states is incomplete; obviously they cannot describe a packet approaching the junction from the right. Consider the effect of time reversal, K, on $B(x,\mathcal{E})$. If the original Bloch states satisfy

$$Kb_r(x,k)=b_r(x,-k), \tag{19}$$

which they certainly may in the absence of degeneracy, we have

$$\begin{aligned} &KB(x,\mathcal{E})\sim b_c(x,-k)+\mathcal{R}(\mathcal{E})^* b_c(x,k), \quad x\to-\infty,\\ &KB(x,\mathcal{E})\sim \mathcal{T}(\mathcal{E})^* b_v(x,-k'), \quad x\to+\infty, \end{aligned} \tag{20}$$

with the usual relations among k, k', and $\mathcal{E}$. Define

$$B'(x,\mathcal{E})=[KB(x,\mathcal{E})-\mathcal{R}^*(\mathcal{E})B(x,\mathcal{E})]/\mathcal{I}(\mathcal{E})^*. \tag{21}$$

Then

$$\begin{aligned} &B'(x,\mathcal{E})\sim \mathcal{T}'(\mathcal{E})b_c(x,-k), \quad x\to-\infty,\\ &B'(x,\mathcal{E})\sim b_v(x,-k')+\mathcal{R}'(\mathcal{E})b_v(x,k'), \quad x\to+\infty, \end{aligned} \tag{22}$$

where

$$\begin{aligned} &\mathcal{T}'(\mathcal{E})=(1-|\mathcal{R}(\mathcal{E})|^2)/\mathcal{T}(\mathcal{E})^*=\mathcal{T}(\mathcal{E})v_v(\mathcal{E})/v_c(\mathcal{E}),\\ &\mathcal{R}'(\mathcal{E})=-\mathcal{R}(\mathcal{E})^*\mathcal{T}(\mathcal{E})/\mathcal{T}(\mathcal{E})^*. \end{aligned} \tag{23}$$

Note that

$$|\mathcal{R}'(\mathcal{E})|^2+|\mathcal{T}'(\mathcal{E})|^2[v_c(\mathcal{E})/v_v(\mathcal{E})]=1, \tag{24}$$

as a consequence of (6), as indeed must be true *a priori*.

It is not hard to see that the $B'(x,\mathcal{E})$, together with the $B(x,\mathcal{E})$, from a complete, linearly independent set of stationary states. We can now analyze the motion of a wave packet representing, for $t\to-\infty$, a valence electron incident on the junction from the right with average energy $\mathcal{E}$ greater than $w_c(0)$. We would find that the reflection probability is $|\mathcal{R}'(\mathcal{E})|^2$, and that the tunneling probability is $|\mathcal{T}'(\mathcal{E})|^2 v_c(\mathcal{E})/v_v(\mathcal{E})=|\mathcal{T}(\mathcal{E})|^2 \times v_v(\mathcal{E})/v_c(\mathcal{E})$.

It should be emphasized that the set $B(x,\mathcal{E})$, $B'(x,\mathcal{E})$ may not be orthogonal. A detailed wave packet analysis shows that this point is irrelevant for the determination of the tunneling probability. The probability of tunneling though the junction in either direction is $|\mathcal{T}(\mathcal{E})|^2 \times v_v(\mathcal{E})/v_c(\mathcal{E})$. The two bands apparently enter this expression asymmetrically only because $\mathcal{T}(\mathcal{E})$ was defined asymmetrically.

Such a wave packet analysis also leads to an expression for the transit time of an electron through the junction in terms of $\mathcal{T}$. Namely, if the transit time, t, is defined as the time for a packet to go from $\rho_0\ll 0$ to $\rho_1\gg l$, minus $|\rho_0|/v_c+(l-\rho_1)/v_v$, then

$$t=l/v_v+\mathrm{Im}(d/d\mathcal{E})\ln\mathcal{T}.$$

III. THE TUNNELING CURRENT

Suppose now that we have succeeded in computing the tunneling amplitude $\mathcal{T}(\mathcal{E})$ which was discussed in the preceding section. The next step is to compute the electric current density j as a function of the applied bias V (considered positive in the forward direction). We shall ignore all scattering effects; scattering in the junction region leads to indirect tunneling,[8,14] which does not interest us here (see, however, the remarks in Sec. VI), while scattering outside the junction region leads to ordinary resistive effects, computable by the standard methods of transport theory,[19] which are superposed on the tunneling effects which are our primary interest.

We must first modify the notation of Sec. II to take account of three-dimensional character of the problem. We replace $\mathcal{E}$ by $\mathbf{k}$ as the parameter on which $\mathcal{I}$ depends; $\mathcal{E}=w_c(\mathbf{k})=w_v(\mathbf{k}')+V_0$. Because of the translational symmetry in the plane parallel to the junction, $\mathbf{k}_\perp'=\mathbf{k}_\perp$. We may regard $\mathbf{k}_\perp$ as an innocent bystander, serving only to determine the relations among $\mathcal{E}$, k_{11}, and k_{11}'. We must take $v_c(\mathbf{k})=\partial w_c(\mathbf{k})/\partial k_{11}$, $v_v(\mathbf{k})=\partial w_v(\mathbf{k}')/\partial k_{11}'$. Everything done in Sec. II is now correct in a three-dimensional setting.

To compute the tunneling current, we use kinetic theory together with the tunneling probability defined in Sec. II and the exclusion principle; the justification for this is the same as in ordinary transport theory. Define

$$f_n(\mathbf{k})=\left[\exp\left(\frac{w_c(\mathbf{k})-\zeta_n}{kT}\right)+1\right]^{-1}, \tag{25}$$

$$f_p(\mathbf{k})=\left[\exp\left(\frac{w_c(\mathbf{k})-\zeta_p}{kT}\right)+1\right]^{-1}. \tag{26}$$

Here ζ_n and ζ_p are the two Fermi levels and $\zeta_n-\zeta_p=eV$. The probability that an electron striking the junction from the n side will tunnel to the p side is $|\mathcal{T}(\mathbf{k})|^2 v_v(\mathbf{k})/v_c(\mathbf{k}) \times[1-f_p(\mathbf{k})]$. The number of such impacts per unit area is $f_n(\mathbf{k})2v_c(\mathbf{k})d^3k/(2\pi)^3$, so the current density due to these events is

$$e\int_{z+} f_n(\mathbf{k})[1-f_p(\mathbf{k})]|\mathcal{T}(\mathbf{k})|^2 2v_v(\mathbf{k})\frac{d^3k}{(2\pi)^3}. \tag{27}$$

Similarly, the current density due to electrons tunneling from the p side to the n side is

$$\begin{aligned} -e\int_{z-} f_p(\mathbf{k})[1-f_n(\mathbf{k})]|\mathcal{T}(\mathbf{k})|^2\frac{v_v(\mathbf{k})}{v_c(\mathbf{k})}2v_v(\mathbf{k})\frac{d^3k'}{(2\pi)^3}\\ =-e\int_{z+} f_p(\mathbf{k})[1-f_n(\mathbf{k})] \\ \times|\mathcal{T}(\mathbf{k})|^2\frac{v_v(\mathbf{k})}{v_c(\mathbf{k})}2v_c(k)\frac{d^3k}{(2\pi)^3}. \end{aligned} \tag{28}$$

[19] Reference 18, Chap. 7.

Finally, the total current density is given by

$$j=e\int_{z+}[f_n(\mathbf{k})-f_p(\mathbf{k})]|\mathcal{T}(\mathbf{k})|^2 2v_v(\mathbf{k})\frac{d^3k}{(2\pi)^3}. \quad (29)$$

It may be remarked that (29) gives a smooth j vs V curve at $V=0$, despite the asymmetrical treatment of the two bands in Sec. II, and that according to (29) there is no current when $V=0$, as is physically obvious.

IV. VARIATIONAL PRINCIPLE FOR THE TUNNELING AMPLITUDE $\mathcal{T}$

We now return to a one-dimensional setting. It is convenient to make one change in the notation of Sec. II: States will be labeled by $k>0$, where $w_c(k)=\mathcal{E}=w_v(k')+V_0$. In order to derive a variational principle for $\mathcal{T}$ we shall use the methods of formal scattering theory.[20,21] Let

$$\mathcal{H}=\mathcal{H}_0+U, \quad (30)$$

where $\mathcal{H}$ is the complete Hamiltonian, $\mathcal{H}_0$ is the in-band part of the Hamiltonian, including effects of the junction field, and U is the interband part of the Hamiltonian, vanishing for zero junction field. For each interesting k [i.e., $w_v(0)+V_0\geq w_c(k)\geq w_c(0)$] define the eigenstates $\beta(k)$, $\beta'(k)$ of $\mathcal{H}_0$ by the boundary conditions

$$\begin{aligned}&\beta(x,k)\sim b_c(x,k)+r(k)b_c(x,-k), \quad x\to-\infty,\\ &\beta(x,k)\to 0, \quad x\to+\infty,\end{aligned} \quad (31)$$

and

$$\begin{aligned}&\beta'(x,k)\to 0, \quad x\to-\infty,\\ &\beta'(x,k)\sim b_v(x,-k')+r'(k)b_v(x,k'), \quad x\to+\infty.\end{aligned} \quad (32)$$

These states form a complete orthogonal set. Also define the operators (Green's functions)

$$G_\pm(k)=(\mathcal{E}-\mathcal{H}_0\pm i\epsilon)^{-1}, \quad (33)$$

where ϵ is a small positive number which shall eventually approach zero. Now define eigenstates $B(k)$, $\bar{B}'(k)$ of $\mathcal{H}$ by the integral equations

$$B(k)=\beta(k)+G_+(k)UB(k), \quad (34)$$

$$\bar{B}'(k)=\beta'(k)+G_-(k)U\bar{B}'(k). \quad (35)$$

As the notation suggests, $B(k)$ is the same as the state with that name introduced in Sec. II. Moreover, it can be shown that

$$\mathcal{R}(k)=r(k)[1-i\Omega_0 v_c(k)^{-1}\langle\beta(k)|U|B(k)\rangle], \quad (39)$$

$$\mathcal{T}(k)=-i\Omega_0 v_v(k)^{-1}r'(k)\langle\beta'(k)|U|B(k)\rangle. \quad (40)$$

The object we wish to compute is

$$|\mathcal{T}(k)|^2[v_v(k)/v_c(k)]=[\Omega_0^2/v_c(k)v_v(k)]|\mathfrak{A}(k)|^2, \quad (41)$$

where

$$\mathfrak{A}(k)=\langle\beta'(k)|U|B(k)\rangle, \quad (42)$$

and we have used the fact, obvious because flux is conserved, that $|r'(k)|^2=1$. We now turn our attention to the calculation of $\mathfrak{A}(k)$.

Let us rewrite (42), using (34) and (35), in the form

$$\mathfrak{A}(k)=\frac{\langle\beta'(k)|U|B(k)\rangle\langle\bar{B}'(k)|U|\beta(k)\rangle}{\langle\bar{B}'(k)|U|B(k)\rangle-\langle\bar{B}'(k)|UG_+U|B(k)\rangle}. \quad (43)$$

This expression is stationary against independent variations of $B(k)$ and $\bar{B}'(k)$. To see this, write

$$\begin{aligned}\mathfrak{A}(k)[\langle\bar{B}'(k)|U|B(k)\rangle-\langle\bar{B}'(k)|UG_+U|B(k)\rangle]\\ =\langle\beta'(k)|U|B(k)\rangle\langle\bar{B}'(k)|U|B(k)\rangle.\end{aligned} \quad (44)$$

Vary $\bar{B}'(k)$:

$$\begin{aligned}\delta\mathfrak{A}(k)[\langle\bar{B}'(k)|U|B(k)\rangle-\langle\bar{B}'(k)|UG_+U|B(k)\rangle]\\ +\mathfrak{A}(k)\langle\delta\bar{B}'(k)|U|B(k)\rangle\\ =\langle\beta'(k)|U|B(k)\rangle\langle\delta\bar{B}'(k)|U|\beta(k)\rangle.\end{aligned} \quad (45)$$

By (42), $\delta\mathfrak{A}(k)=0$. Now vary $B(k)$:

$$\begin{aligned}\delta\mathfrak{A}(k)[\langle\bar{B}'(k)|U|B(k)\rangle-\langle\bar{B}'(k)|UG_+U|B(k)\rangle]\\ +\mathfrak{A}(k)\langle\beta'(k)|U|\delta B(k)\rangle\\ =\langle\bar{B}'(k)|U|\beta(k)\rangle\langle\beta'(k)|U|\delta B(k)\rangle.\end{aligned} \quad (46)$$

Because

$$\begin{aligned}\mathfrak{A}(k)=\langle\beta'(k)|U|B(k)\rangle=\langle\bar{B}'(k)|U|B(k)\rangle\\ -\langle\bar{B}'(k)|UG_+U|B(k)\rangle=\langle\bar{B}'(k)|U|\beta(k)\rangle,\end{aligned} \quad (47)$$

$\delta\mathfrak{A}(k)=0$ again.

We shall use formula (43) for an approximate calculation of $\mathfrak{A}(k)$ by substituting

$$\begin{aligned}B(k)&\approx\beta(k),\\ \bar{B}'(k)&\approx\beta'(k).\end{aligned} \quad (48)$$

This gives

$$\mathfrak{A}(k)\approx\frac{\langle\beta'(k)|U|\beta(k)\rangle^2}{\langle\beta'(k)|U|\beta(k)\rangle-\langle\beta'(k)|UG_+U|\beta(k)\rangle}. \quad (49)$$

This formula differs from the usual expression,

$$\mathfrak{A}(k)\approx\langle\beta'(k)|U|\beta(k)\rangle, \quad (50)$$

used in the transition rate formalism[17] by the second term in the denominator; because (43) is a stationary expression, one would expect (49) to be an improvement over (50). However, $G_\pm$ has no interband matrix elements, so that if U has no in-band matrix elements the correction term vanishes and (49) reduces to (50). This circumstance may explain the success of several fairly crude calculations of the tunneling current.[14]

That (50) is a good approximation to (42) when U has only interband matrix elements may be understood from the fact that only the conduction band component of $B(k)$ contributes to (42), and that component is well approximated by $\beta(k)$ for reasonable electric field

[20] B. A. Lippman and J. Schwinger, Phys. Rev. **79**, 469 (1950).
[21] M. Gell-Mann and M. L. Goldberger, Phys. Rev. **91**, 398 (1953).

strengths. This argument would fail if U had in-band matrix elements, and in that case one could expect (49) to be preferable to (50).

V. EVALUATION OF THE FORMULAS

A. Schrödinger Equation in the Wannier Representation

In view of the developments so far, we need only make the separation (30) of the Hamiltonian and determine the eigenstates $\beta(k)$ and $\beta'(k)$ of $\mathcal{H}_0$. For these concrete calculations, it is easiest to work in the Wannier representation. Moreover, this is a suitable time to work explicitly in three dimensions, since our answer is going to be used in (29). Letting $\varphi(x)$ be the potential of the external electric field, including the field at zero bias due to the donors and acceptors, we take

$$\mathcal{H}_0=w(-i\nabla_\rho)-e\varphi(\rho). \tag{51}$$

We shall make use of the separation

$$x=\rho+X \tag{52}$$

of the position operator, where X is the operator introduced by Adams,[22] and we shall assume that X has no nonvanishing in-band matrix elements (as is always true in the absence of degeneracy if the crystal is invariant under space inversion). In the spirit of the approximations (48), we shall regard X as a small quantity in terms of which we work only to first order. We may then take

$$U=eE(\rho)X; \tag{53}$$

we shall be concerned exclusively with a model for which the ambiguity of the order of $E(\rho)$ and X in (53) is irrelevant.

Although the potential $\varphi(x)$ is explicitly computable in terms of the donor and acceptor concentrations, we shall adopt the model that

$$\begin{aligned}\varphi(x)&=0, \quad x\leq 0,\\ &=-Ex, \quad 0\leq x\leq l,\\ &=-El=-V_0/e, \quad x\geq l;\end{aligned} \tag{54}$$

if certain conditions, to be enumerated later, are met, one could hope that the error caused by the assumption (54) is small.

Finally, in order to make our problem explicitly soluble, we make the effective mass approximation

$$\begin{aligned}w_c(\mathbf{k})&=\mathbf{k}^2/2m_c,\\ w_v(\mathbf{k})&=-\Delta-\mathbf{k}^2/2m_v,\end{aligned} \tag{55}$$

with positive effective masses m_c, m_v.[23] We may define the classical turning points

$$\rho_c=k_{11}^2/2m_ceE, \quad \rho_v=l-k_{11}'^2/2m_veE; \tag{56}$$

recall that

$$\mathcal{E}=k^2/2m_c=V_0-\Delta-k'^2/2m_v, \tag{57}$$

and we are only interested in energies for which k and k' are real. We shall make frequent use of the conditions

$$0\ll\rho_c\ll\rho_v\ll l, \tag{58}$$

to which we referred following (54).

Let us compute $\beta(\rho)$ (the parameter $\mathbf{k}$ shall be omitted where no confusion can result): In the region $0<\rho<l$, the Schrödinger equation for β is

$$\left(\frac{\mathbf{k}_\perp^2}{2m_c}-\frac{1}{2m_c}\frac{d^2}{d\rho^2}\right)\beta+eE\rho\beta=\mathcal{E}\beta, \tag{59}$$

where the uninteresting factor $\exp(i\mathbf{k}_\perp\cdot\boldsymbol{\varrho}_\perp)$ has been omitted. Substituting

$$\xi=\alpha_c(1-\rho/\rho_c), \tag{60}$$

where

$$\alpha_c=(\rho_ck_{11})^{2/3}, \tag{61}$$

Eq. (59) becomes

$$(d^2\beta/d\xi^2)+\xi\beta=0,$$

which is just (A1) of the Appendix. Because of (31), the appropriate solution is

$$\beta(\rho)=N_cA_2(\xi)=N_cA_2[\alpha_c(1-\rho/\rho_c)], \quad 0<\rho<l, \tag{62}$$

where N_c is a normalization constant to be determined by invoking (31). Again using (58), we know that there is a region $0<\rho\ll\rho_c$ in which the asymptotic form of A_2, obtained from (A6), (A7), and (A9), is applicable. In this region

$$\begin{aligned}\beta(\rho)\sim 2iN_c\pi^{1/2}\alpha_c^{-1/4}(1-\rho/\rho_c)^{-1/4}\\ \times\sin[\tfrac{2}{3}\alpha_c^{3/2}(1-\rho/\rho_c)^{3/2}+\tfrac{1}{4}\pi].\end{aligned} \tag{63}$$

In the spirit of (54), which treats the ends of the junction as unimportant, we match (63) to the standing wave outside the junction, namely,

$$\beta(\rho)=2iN_c\pi^{1/2}\alpha_c^{-1/4}\sin(-k_{11}\rho+\tfrac{2}{3}\alpha_c^{3/2}+\tfrac{1}{4}\pi), \quad \rho<0, \tag{64}$$

by requiring continuity of amplitude, phase, and derivative of phase. Comparing (64) with (31), we obtain

$$N_c=-\pi^{-1/2}\alpha_c^{1/4}\exp[i(\tfrac{2}{3}\alpha_c^{3/2}+\tfrac{1}{4}\pi)]. \tag{65}$$

Now on to $\beta'(\rho)$: For $0<\rho<l$, the Schrödinger equation for β' is

$$\left(-\Delta-\frac{\mathbf{k}_\perp^2}{2m_v}+\frac{1}{2m_v}\frac{d^2}{d\rho^2}\right)\beta'+eE\rho\beta'=\beta'. \tag{66}$$

Substituting

$$\eta=\alpha_v[1+(\rho-l)/(l-\rho_v)], \tag{67}$$

[22] E. N. Adams, Phys. Rev. **85**, 41 (1952); and J. Chem. Phys. **21**, 2013 (1953).

[23] Unfortunately, by making this approximation we renounce all hope of obtaining the so-called Stark ladder effects which were observed by A. G. Chynoweth, G. H. Wannier, R. A. Logan, and D. E. Thomas, Phys. Rev. Letters **5**, 57 (1960).

where

$$\alpha_v=[(-k_{11}')(l-\rho_v)]^{2/3}, \tag{68}$$

(66) becomes

$$(d^2\beta'/d\eta^2)+\eta\beta'=0, \tag{69}$$

which, again, is (A1). Because of (32), the appropriate solution is

$$\beta'(\rho)=N_vA_2(\eta)=N_vA_2\left[\alpha_v\left(1+\frac{\rho-l}{l-\rho_v}\right)\right], \quad 0<\rho<l, \tag{70}$$

where N_v is to be determined by comparison with (33). Just as in the calculation of $\beta(\rho)$, we find

$$N_v=\pi^{-1/2}\alpha_v^{1/4}\exp[-i(\tfrac{2}{3}\alpha_v^{3/2}+\tfrac{1}{4}\pi)]. \tag{71}$$

B. The Operator X

In the Bloch representation X is diagonal in k and, with our normalization (19) of the Bloch states, $X_{vc}(k)=X_{cv}(k)$. In the Wannier representation, X acts as follows:

$$X\begin{bmatrix}\psi_c(\rho)\\ \psi_v(\rho)\end{bmatrix}=\begin{bmatrix}X_{cv}(-i\nabla\rho)\psi_v(\rho)\\ X_{vc}(-i\nabla\rho)\psi_c(\rho)\end{bmatrix}. \tag{72}$$

Lacking any better information, we shall use for $X_{cv}(k)$ the function obtained from the "$\mathbf{k}\cdot\mathbf{p}$ perturbation" method[24] applied to the one-dimensional case:

$$X_{cv}(\mathbf{k})=\tfrac{1}{2}(m_r\Delta+\tfrac{1}{2}\mathbf{k}_\perp^2)[k_{11}^2+m_r\Delta+\tfrac{1}{2}\mathbf{k}_\perp^2]^{-1}, \tag{73}$$

where $1/m_r=1/m_v+1/m_c$. The most important feature of (73) is the presence of poles at $\pm i(m_r\Delta+\frac{1}{2}\mathbf{k}_\perp^2)^{1/2}$ with residues $\mp i/4$; the existence of such poles at the branch points of $w(\mathbf{k})$, with residues $\mp i/4$, is a rigorous property of $X_{cv}(\mathbf{k})$.[7]

To study the character of $X_{cv}(-i\nabla\rho)$, we must find that Green's function of $(\frac{1}{2}\mathbf{k}_\perp^2+m_r\Delta-d^2/d\rho^2)$ which is regular for $|\rho|\to\infty$; this is easily seen to be $\frac{1}{2}(m_r\Delta+\frac{1}{2}\mathbf{k}_\perp^2)^{-1}\exp[-(m_r\Delta+\frac{1}{2}\mathbf{k}_\perp^2)^{1/2}|\rho|]$. We may conclude that $X_{cv}(-i\nabla_\rho)$ is an integral operator with kernel

$$\tfrac{1}{4}\exp[-(m_r\Delta+\tfrac{1}{2}\mathbf{k}_\perp^2)^{1/2}|\rho-\rho'|]; \tag{74}$$

this expression is just the Fourier transform of (73). We may say that $X_{cv}(-i\nabla_\rho)$ has range $(m_r\Delta+\mathbf{k}_\perp^2/2)^{-1/2}$.

C. Calculation of $\mathfrak{a}$

If we assemble the pieces which we obtained in **A** and **B**, we have, using (50) and (74),

$$\begin{aligned}\mathfrak{a}(\mathbf{k})&=eE\sum_{0<\rho<l}\beta'(\rho)^*\int_{-\infty}^{\infty}d\rho'\\ &\quad\times\tfrac{1}{4}\exp[-(m_r\Delta+\tfrac{1}{2}\mathbf{k}_\perp^2)^{1/2}|\rho-\rho'|]\beta(\rho')\\ &=\frac{eE}{4\Omega_0}\int_0^l d\rho\int_{-\infty}^{\infty}d\rho'\,\beta'(\rho)^*\\ &\quad\times\exp[-(m_r\Delta+\tfrac{1}{2}\mathbf{k}_\perp^2)^{1/2}|\rho-\rho'|]\beta(\rho').\end{aligned} \tag{75}$$

[24] H. Jones, *The Theory of Brillouin Zones and Electronic States in Crystals* (North-Holland Publishing Company, Amsterdam, 1960), pp. 39–43.

Because, once more, of (58), the integral with respect to ρ' can be cut off on the left at $\rho'=0$ and on the right at $\rho'=l$. For this we must, in particular, have $\rho_v\gg(m_r\Delta+\frac{1}{2}\mathbf{k}_\perp^2)^{-1/2}$. We can, therefore, substitute (62) and (70) directly into (75):

$$\begin{aligned}\mathfrak{a}(\mathbf{k})&=\frac{eE}{4\Omega_0}N_v^*N_c\int_0^l d\rho\int_0^l d\rho'\,A_2^*\left[\alpha_v\left(1+\frac{\rho-l}{l-\rho_v}\right)\right]\\ &\quad\times\exp[-(m_r\Delta+\tfrac{1}{2}\mathbf{k}_\perp^2)^{1/2}|\rho-\rho'|]\\ &\quad\times A_2[\alpha_c(1-\rho'/\rho_c)].\end{aligned} \tag{76}$$

Just as the integral with respect to ρ' in (75) could be cut off, we can extend both integrations in (76) from $-\infty$ to $+\infty$. The resulting expression simplifies if we use Fourier transforms:

$$\begin{aligned}\mathfrak{a}(\mathbf{k})&=\frac{eE}{2\Omega_0}(m_r\Delta)^{1/2}N_v^*N_c\int_{-\infty}^{\infty}\frac{d\kappa}{2\pi}\left(\frac{l-\rho_v}{\alpha_v}\frac{\rho_c}{\alpha_c}\right)\\ &\quad\times\exp[i\kappa(\rho_v-\rho_c)]\tilde{A}_2^*\left(\frac{l-\rho_v}{\alpha_v}\kappa\right)\\ &\quad\times\frac{1}{\frac{1}{2}\mathbf{k}_\perp^2+m_r\Delta+\kappa^2}\tilde{A}_2\left(-\frac{\rho_c\kappa}{\alpha_c}\right).\end{aligned} \tag{77}$$

Using (65) and (71), (77) becomes

$$\begin{aligned}\mathfrak{a}(\mathbf{k})&=\frac{-ieE}{4\pi^2\Omega_0}\frac{(m_r\Delta)^{1/2}}{(\alpha_v\alpha_c)^{3/4}}(l-\rho_v)\rho_c\int_{-\infty}^{\infty}d\kappa\exp[i\kappa(\rho_v-\rho_c)]\\ &\quad\times\tilde{A}_2^*\left(\frac{l-\rho_v}{\alpha_v}\kappa\right)\frac{1}{\frac{1}{2}\mathbf{k}_\perp^2+m_r\Delta+\kappa^2}\tilde{A}_2\left(-\frac{\rho_c}{\alpha_c}\kappa\right).\end{aligned} \tag{78}$$

We must now evaluate the integral in (78), which, thanks to (56), (61), (68), and (A10), may be written

$$\begin{aligned}I&=4\pi^2\int_{-\infty}^{\infty}d\kappa\exp\left[i\left(\frac{\Delta+\mathbf{k}_\perp^2/2m_r}{eE}\kappa+\frac{\kappa^3}{6m_reE}\right)\right]\\ &\quad\times\frac{1}{\kappa^2+m_r\Delta+\frac{1}{2}\mathbf{k}_\perp^2}.\end{aligned} \tag{79}$$

The argument of the exponential has a saddle point at

$$\kappa_s=i[2m_r\Delta+\mathbf{k}_\perp^2]^{1/2}, \tag{80}$$

with range

$$R_s=(2m_reE)^{1/2}|\kappa_s|^{-1/2}=(2m_reE)^{1/2}(2m_r\Delta+\mathbf{k}_\perp^2)^{-1/4}. \tag{81}$$

However, if we wish to deform the path of integration so that it passes through κ_s, we must take into account the pole at

$$\kappa_p=i(m_r\Delta+\tfrac{1}{2}\mathbf{k}_\perp^2)^{1/2}=2^{-1/2}\kappa_s, \tag{82}$$

whose residue is (omitting the $4\pi^2$)

$$R_p=\exp\left(-\frac{m_r^{1/2}(\Delta+\mathbf{k}_\perp^2/2m_r)^{3/2}}{6eE}\right)\times\frac{-i}{2m_r^{1/2}(\Delta+\mathbf{k}_\perp^2/2m_r)^{1/2}}. \quad (83)$$

The saddle point integration can be performed in a straightforward way if the pole κ_p is far from the saddle point κ_s in the sense that

$$|\kappa_p-\kappa_s|\sim|\kappa_s|\gg R_s. \quad (84)$$

(84) imposes an upper limit on the field E which is satisfied in practice. Careful scrutiny of (79) shows that (84) is also the condition for the saddle-point integral to be negligible compared to R_p. We therefore write

$$I=4\pi^3 m_r^{-1/2}(\Delta+\mathbf{k}_\perp^2/2m_r)^{-1/2}\times\exp\left(-\frac{5m_r^{1/2}(\Delta+\mathbf{k}_\perp^2/2m_r)^{3/2}}{6eE}\right). \quad (85)$$

Substituting (85) in (78), and using (56), (61), and (68), we finally obtain

$$\mathfrak{a}(\mathbf{k})=\frac{-i\pi}{2\Omega_0}(m_c m_v)^{-1/2}k_{\parallel}^{1/2}(-k_{\parallel}')^{1/2}\left(1+\frac{\mathbf{k}_\perp^2}{2m_r\Delta}\right)^{-1/2}\times\exp\left(-\frac{5m_r^{1/2}(\Delta+\mathbf{k}_\perp^2/2m_r)^{3/2}}{6eE}\right). \quad (86)$$

D. Calculation of the Current

The formula (29) for the current density contains the dynamical quantity $|\mathcal{T}|^2$, or, using (41), the quantity $|\mathfrak{a}|^2$, which, we found, was given by (86). Putting this information together, we have

$$j=\frac{e}{8}\int_0^{V_0{}^i-eV-\Delta}d\mathcal{E}\,[f_n(\mathcal{E})-f_p(\mathcal{E})]\times\int_{k_\perp^2/2m_c<\mathcal{E}}\exp\left[-\gamma\left(1+\frac{k_\perp^2}{2m_r\Delta}\right)^{3/2}\right]\frac{k_\perp dk_\perp}{1+k_\perp^2/2m_r\Delta}, \quad (87)$$

where $V_0{}^i$ is the value of V_0 (see Fig. 1) for zero bias, $V=0$, and

$$\gamma=5m_r^{1/2}\Delta^{3/2}/3\hbar eE; \quad (89)$$

for the convenience of the reader, we have restored $\hbar$ in (89).

Since γ is large, we may drop the upper limit in the $k_\perp$ integration; this corresponds to writing j in the form $e^{-\gamma}\times$ (asymptotic series in E) and keeping only the first term in the asymptotic series. We now have

$$j=\frac{em_r\Delta}{12\gamma\hbar^3}e^{-\gamma}\int_0^{V_0{}^i-eV-\Delta}d\mathcal{E}\,[f_n(\mathcal{E})-f_p(\mathcal{E})]. \quad (90)$$

The integral in (90) can be explicitly evaluated, giving

$$j=\frac{em_r\Delta kT}{12\gamma\hbar^3}e^{-\gamma}\ln\left[\frac{[1+\exp(\zeta_n/kT)][1+\exp((\zeta_n+\Delta-V_0{}^i)/kT)]}{[1+\exp((\zeta_n-eV)/kT)][1+\exp((\zeta_n+\Delta-V_0{}^i+eV)/kT)]}\right]. \quad (91)$$

To facilitate comparison with other formulas in the literature, we substitute (89) in (91) and have

$$j=\frac{e^2Em_r^{1/2}\Delta^{-1/2}kT}{20\hbar^2}\exp\left[-\frac{5m_r^{1/2}\Delta^{3/2}}{3\hbar eE}\right]\ln\left[\frac{[1+\exp(\zeta_n/kT)][1+\exp((\zeta_n+\Delta-V_0{}^i)/kT)]}{[1+\exp((\zeta_n-eV)/kT)][1+\exp((\zeta_n+\Delta-V_0{}^i+eV)/kT)]}\right]. \quad (92)$$

In (91) and (92), k is, of course, Boltzmann's constant. A graph of j vs V is given in Fig. 2 for the zero-temperature limit, ignoring the change in E due to the bias; experimental curves contain the ordinary p-n junction characteristic superposed on this type of curve.

VI. DISCUSSION

The main results which have emerged from our formal machinery are the variational principle (43) for $\mathfrak{a}$ and the unambiguous prescription for the use of $\mathfrak{a}$ in calculating the current density. Although these were derived above for the special case of direct tunneling, they are widely applicable to indirect tunneling processes in which scattering from impurities or through the interactions of electrons with each other and with other excitations, such as phonons, plays a role, and to tunneling processes in the presence of a magnetic field. The structure of the variational principle suggests that the calculation might be very much improved if trial wave functions containing polarization corrections, on which the correction term $\langle\bar{B}|UG_+U|B\rangle$ could operate, were used.

The computations of Sec. V show how the formal machinery works. It is gratifying to have obtained (92), and Fig. 2, which agree with simple physical pictures. Although the treatment of the three-dimensionality of the problem was crude, and could be improved, it enabled us to simplify the arithmetic by taking into account the essential one-dimensionality of the physical situation; except for trivial numerical factors, such as the "20" in (92), a better calculation would probably agree with ours. Another strictly incorrect aspect of the computation was the lack of consistency between the $X_{cv}(k)$ and the $w_{c,v}(k)$ which we used. It is well known[24] that these quantities are related analytically. However, the $X_{cv}(k)$ which is consistent with the

assumption of parabolic bands is a constant; using this omits the Keldysh singularity, which is fatal. If both $w_{c,v}(k)$ and $X_{cv}(k)$ are taken from $k\cdot p$ perturbation theory, the calculation is consistent but it is difficult to solve the one-band wave equations for $\beta(k)$ and $\beta'(k)$, which must be known throughout the crystal. We have therefore effected a compromise in Sec. V which enables us to avoid physically uninstructive mathematical complications.

At this point it is easy to understand the sources of disagreement between (92) and the recent result of Kane to which we referred in Sec. I.[13] First, our value of γ, (89), is to be compared with Kane's value $\pi m_r^{1/2}\Delta^{3/2}/8\hbar eE$. These differ by a factor of 1.06, which is not very significant in view of the crudeness of both calculations; the origin of this discrepancy is our use of the effective-mass approximation as compared with Kane's use of the band energies $w(k)$ derived from the $k\cdot p$ perturbation method. Second, our equations would have led to Kane's result if, in (87), we had put $\mathbf{k}_\perp^2/2m_r\Delta\approx 0$. This, however, would surely be incorrect for small E, and therefore large γ, because it removes the cutoff factor in the integral over $\mathbf{k}_\perp$ in (87), and this cutoff is most significant precisely in the case of small E; of course, for large E the quasiclassical approach breaks down completely.

Finally, a word is in order about the unrealistic potential (54) which was used in Sec. V. The lack of realism was serious only in that the "corners" at $\rho=0$ and $\rho=l$ were not rounded off. But because we made repeated use of (58), the exact character of the corners was unimportant. Both the special potential (54) and the approximations based on (58) can be avoided if one is willing to do the calculations numerically.

ACKNOWLEDGMENTS

We wish to thank Dr. E. Blount, Dr. M. L. Goldberger, and Dr. M. Lax for helpful discussions.

APPENDIX

Equation $u''+xu=0$

In Sec. V we have made extensive use of the properties of the solutions of the equation

$$(d^2u/dx^2)+xu=0. \tag{A1}$$

Although these solutions are well known, and are expressible in terms of Bessel functions, it is easier to rederive them ourselves. It is evident that

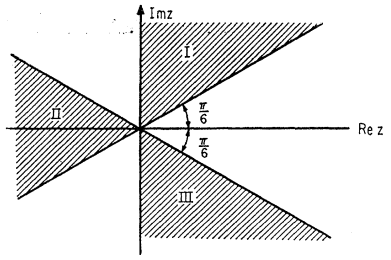

FIG. 3. Regions of the complex plane for (A3), (A4), and (A5).

$$A(x)=\int_C \exp(\tfrac{1}{3}z^3+xz)dz \tag{A2}$$

is a solution of (A1) if $\mathrm{Re}z^3\to-\infty$ at the end points of the curve C. This will be the case if the end points of C are at infinity in the shaded portion of the z plane in Fig. 3. We define

$$A_1(x)=\int_{\mathrm{II}}^{\mathrm{III}} \exp(\tfrac{1}{3}z^3+xz)dz, \tag{A3}$$

$$A_2(x)=\int_{\mathrm{III}}^{\mathrm{I}} \exp(\tfrac{1}{3}z^3+xz)dz, \tag{A4}$$

$$A_3(x)=\int_{\mathrm{II}}^{\mathrm{I}} \exp(\tfrac{1}{3}z^3+xz)dz. \tag{A5}$$

Clearly,

$$A_2(x)=A_3(x)-A_1(x). \tag{A6}$$

Using our integral representations for the A's, we can obtain asymptotic forms for $|x|\to\infty$ by an application of the saddle-point method. The saddle points of the integrand occur when $z^2=-x$. We obtain

$$A_1(x)\sim e^{-\pi i/4}\pi^{1/2}x^{-1/4}\exp(-\tfrac{2}{3}ix^{3/2}), \quad x\to+\infty, \tag{A7}$$

$$A_2(x)\sim i\pi^{1/2}(-x)^{-1/4}\exp[-\tfrac{2}{3}(-x)^{3/2}], \quad x\to-\infty, \tag{A8}$$

$$A_3(x)\sim e^{\pi i/4}\pi^{1/2}x^{-1/4}\exp(\tfrac{2}{3}ix^{3/2}), \quad x\to+\infty. \tag{A9}$$

It is also useful to know the Fourier transform of $A_2(x)$:

$$\tilde{A}_2(k)=\int_{-\infty}^{\infty} dx\, e^{-ikx}A_2(x)=2\pi i\exp(-\tfrac{1}{3}ik^3). \tag{A10}$$