Exact effective-mass theory for heterostructures

Bradley A. Foreman*
School of Electrical Engineering, Phillips Hall, Cornell University, Ithaca, New York 14853-5401
(Received 25 January 1995; revised manuscript received 29 June 1995)

An exact effective-mass differential equation is derived for electrons in heterostructures. This equation is exactly equivalent to the Schrödinger equation, and is obtained by applying a k-space transformation of variables to the Burt envelope-function theory in which the Brillouin zone is mapped onto the infinite real axis. The mapping eliminates all nonlocal effects and long-range Gibbs oscillations in the Burt theory, producing an infinite-order differential equation in which interface effects are strongly localized to the immediate vicinity of the interface. A general procedure is given for obtaining finite-order boundary conditions from the infinite-order equation; the second-order theory reduces to the BenDaniel-Duke model with a δ-function potential at the interface. The derivation is presented for a simple one-dimensional crystal but can easily be generalized for more complex situations.

I. INTRODUCTION

Nearly 60 years have elapsed since Wannier's proposal of a full-Brillouin-zone effective-mass theory for crystals with slowly varying inhomogeneities.[1] In that time, a considerable effort[2–32] has gone into extending this theory to include the types of rapid inhomogeneities that occur in common semiconductor heterostructures—namely, far too abrupt for conventional slow-variation techniques[1,2,24,27,29] to be used, but sufficiently small in magnitude that a single-band description remains valid throughout the structure. Despite this effort, no satisfactory generalization of the Wannier-Slater theory has yet been found, leading many to conclude that none will ever be found.[22,33] Such judgments may be a bit premature, however, since most previous analyses of abrupt heterojunctions[8,9,13,22,26] have been content to stop with the derivation of second-order connection rules for the envelope functions across the interface, without seriously considering whether it might be possible to extend the use of differential equations into the interface region itself. Only in recent years has this issue been confronted directly, most notably in the exact envelope-function theory proposed by Burt.[15–20] Burt's theory has provided a great deal of insight into why the effective-mass method works so well at an abrupt junction; it remains, nonetheless, a fundamentally nonlocal theory, hence it cannot produce an exact equation of motion having the local differential form that is usually desired.

This paper describes a theoretical approach that yields, to the author's knowledge, the first exact effective-mass differential equation for electrons in heterostructures. The method is based on a k-space transformation of Burt's envelope-function theory that maps the (finite) Brillouin zone onto the infinite real axis. This mapping eliminates all undesirable effects arising from the finite bandwidth in k space, including both long-range Gibbs oscillations and nonlocal effects. The resulting infinite-order differential equation has excellent analytical properties, since the perturbations arising at a heterojunction decay as Gaussian functions (as opposed to the sinc functions found in Burt's formulation). The derivation of the correct macroscopic boundary conditions for slowly varying bulk envelopes is then straightforward (although not trivial), allowing one to see clearly why the neglect of the higher-order differential operators is valid even at an abrupt junction.

The great power of the Wannier-Slater theory[1] lies in its broad scope. By deliberately ignoring the question of how the bulk band structure is to be calculated, this theory allows us to make some very general statements about how an electron behaves under the influence of a slowly varying perturbation. The present work aims to achieve this sort of generality for heterostructures. It is therefore concerned not with any particular method of solution for the heterostructure problem, but rather the form of the equations that the envelope function must obey, given that such a solution exists. Indeed, one of the central messages of this paper is that by custom tailoring our definition of what an envelope function is, we can adjust the equation of motion to have the properties that we desire.

For clarity and simplicity the derivation will be presented for a one-dimensional, lattice-matched, spinless system in which a single-band description is valid over the entire energy range of interest. The extension of this theory to more complex situations presents no difficulty in principle once the basic concepts are understood, but for now it is best to consider only the simplest case possible. The discussion begins in Sec. II with a general examination of the mathematical properties of the functions we will be considering, focusing on some of the problems associated with restricting the envelopes to the first Brillouin zone. The actual derivation begins in Sec. III, where Burt's coupled integrodifferential equations are recast as pure integral equations; the coupling between bands is then eliminated, resulting in a single-band integral equation. In Sec. IV an attempt is made to rewrite

this integral equation as a local differential equation, but it is found that nonlocal terms must always be included if the bandwidth of the envelopes is to remain finite. The solution to this problem is the k-space mapping technique described above and presented in Sec. V, which produces an exact infinite-order differential equation. The justification for discarding the higher-order differential operators is developed in Sec. VI, along with a simple method for determining the interface connection rules in any finite-order theory. A review of the results is presented in Sec. VII.

II. QUASICONTINUUM

Before beginning the derivation that is the subject of this paper, it is worthwhile to spend some time examining the mathematical properties of the functions that arise in the Burt envelope-function theory. The single most important property is that all envelope functions are *strictly* limited to wave vectors within the first Brillouin zone. This constraint is imposed in order to establish a unique relationship between the envelope-function theory and the microscopic theory (the Schrödinger equation) from which it was derived. To be specific, if we define the Fourier transform according to

$$\begin{aligned} F(x)&=\frac{1}{\sqrt{2\pi}}\int F(k)e^{ikx}dk \\ F(k)&=\frac{1}{\sqrt{2\pi}}\int F(x)e^{-ikx}dx \end{aligned}, \tag{2.1}$$

$(-\pi/a \le k \le \pi/a)$ for functions of a single variable and

$$\begin{aligned} H(x,x')&=\frac{1}{2\pi}\int\int e^{ikx}H(k,k')e^{-ik'x'}dk\,dk' \\ H(k,k')&=\frac{1}{2\pi}\int\int e^{-ikx}H(x,x')e^{ik'x'}dx\,dx' \end{aligned}, \tag{2.2}$$

$[-\pi/a \le (k,k') \le \pi/a]$ for functions of two variables, then both $F(k)$ and $H(k,k')$ are nonzero only in the range $-\pi/a \le (k,k') \le \pi/a$, where a is the lattice constant. Functions satisfying this constraint are often referred to as "quasicontinuum" functions[34] because they can be completely specified (via the sampling theorem[35]) in terms of their values at the lattice points, hence they are in a sense both continuous and discrete. This means that there is a fundamental mathematical equivalence between the exact envelope-function theory for electrons and the corresponding theory for phonons.[34,36]

This wave-vector restriction is nothing new—it is, after all, an explicit part of the Luttinger-Kohn effective-mass theory,[2] and is implicit in the Wannier-Slater formulation[1] as well—but it is only since the advent of the Burt theory[15] that it has assumed an important role in the analysis of heterostructures. This is because most of the early heterostructure models were developed by grafting together two bulk envelopes (with suitable boundary conditions at the interface), and in the bulk one can choose to work in the limit of arbitrarily long wavelengths. However, if one treats the heterostructure as a whole, one is forced to deal with functions that vary rapidly on the scale of the unit cell, and the zone-boundary region begins to exert a noticeable influence on the properties of the envelope functions.

To see this, we can start by looking at one of the most rapidly varying quasicontinuum functions, namely the sinc function δ_B, which is defined by

$$\begin{aligned} \delta_B(x)&=\frac{\sin(\pi x/a)}{\pi x} \\ \delta_B(k)&=\frac{1}{\sqrt{2\pi}}B(k) \end{aligned}, \tag{2.3}$$

where

$$B(k)=\begin{cases} 1 & \text{if } |k|<\pi/a \\ 0 & \text{if } |k|>\pi/a \end{cases}. \tag{2.4}$$

This function, shown in Fig. 1(a), is just the first-zone part of the Dirac δ function. When convolved with any given function f, δ_B acts as an ideal low-pass filter:

$$\int \delta_B(x-x')f(x')dx'=f_B(x) , \tag{2.5a}$$

where

$$f_B(k)=B(k)f(k) . \tag{2.5b}$$

In other words, δ_B truncates the given function to Fourier components within the first Brillouin zone. For quasicontinuum functions, $f=f_B$, so δ_B merely acts as an ordinary δ function.

Another frequently encountered function is the unit step, which is useful in representing the change in material properties between two bulk media in a heterostructure. The best-known step function is the discontinuous Heaviside function θ:

$$\begin{aligned} \theta(x)&=\begin{cases} 1 & \text{if } x>0 \\ 0 & \text{if } x<0 \end{cases} \\ \theta(k)&=\frac{1}{\sqrt{2\pi}}\left[\pi\delta(k)+\frac{1}{ik}\right] \quad (-\infty \le k \le \infty) . \end{aligned} \tag{2.6}$$

This is, of course, not a quasicontinuum function, since an infinite bandwidth is needed to represent the discontinuity. The quasicontinuum version of the unit step may be obtained by inserting (2.6) into (2.5), which yields

$$\begin{aligned} \theta_B(x)&=\int_{-\infty}^{x}\delta_B(x')dx'=\frac{1}{2}+\frac{1}{\pi}\mathrm{Si}(\pi x/a) , \\ \theta_B(k)&=B(k)\theta(k) , \end{aligned} \tag{2.7}$$

where $\mathrm{Si}(x)$ is the sine integral.[37] The step functions θ and θ_B are depicted in Fig. 1(b).

From the behavior of δ_B and θ_B in Fig. 1, one of the most undesirable features of the quasicontinuum formalism should now be apparent, namely that any rapid

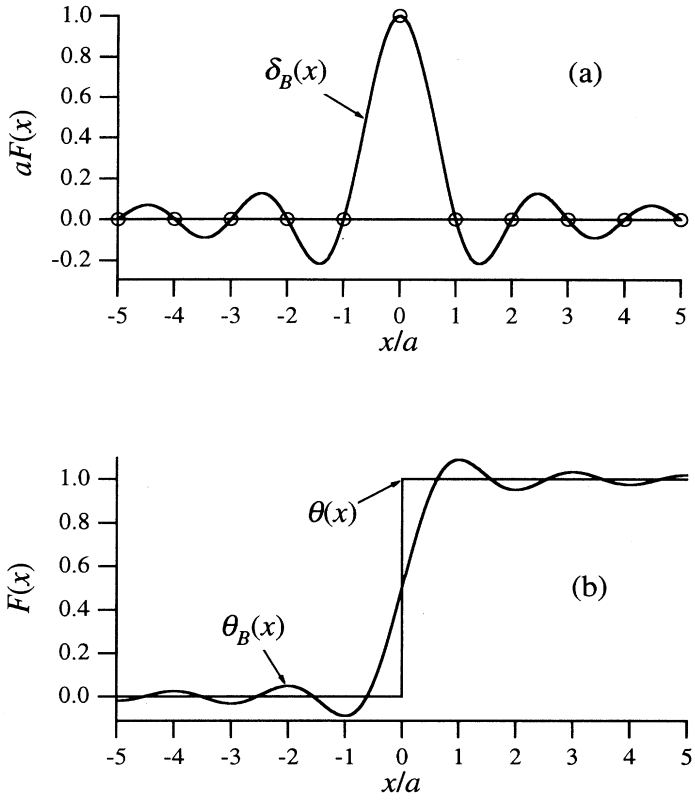

FIG. 1. Quasicontinuum functions: (a) δ function, (b) step function.

change is always accompanied by the appearance of long-range oscillations. These oscillations, commonly known as Gibbs oscillations, are of substantial magnitude even many lattice constants away from the discontinuity. They arise from the abrupt truncation of the quasicontinuum functions at the zone boundary. Although they are in a sense mathematically "real"—since the quasicontinuum is an exact representation of the underlying microscopic theory—one should be very cautious in interpreting the physical significance of these oscillations.

For example, consider the linear chain of atoms represented by the circles in Fig. 1(a), where only a single atom is displaced from equilibrium. This is clearly a strongly localized displacement pattern. The quasicontinuum envelope representing this displacement is proportional to $\delta_B(x)$, which vanishes at all lattice points other than the one displaced. Overall, however, $\delta_B(x)$ is very poorly localized, dropping off in magnitude only as $1/x$. Therefore, if one does not take into account the fact that the envelope has physical meaning *only* at the lattice points, it is easy to be seduced into attributing more significance to these long-range oscillations than actually exists.

In addition to this pitfall of a physical nature, the Gibbs oscillations lead to another problem of a strictly mathematical character. As an example, suppose that we wish (for reasons to be seen below) to calculate the mth-order moment of inertia I_m $(m \geq 0)$ associated with the displacement pattern in Fig. 1(a). If we perform this calculation as a discrete sum, there is no problem; taking the displaced atom to be at $x_n = na$, we have simply

$$I_m = \sum_{n'} x_{n'}^m \delta_{nn'} = x_n^m , \tag{2.8a}$$

where $\delta_{nn'}$ is the Kronecker delta. This sum involves only a single nonzero term, so there is no question of its convergence. On the other hand, if we perform the calculation in the quasicontinuum, we must be more careful:

$$I_m = \int x^m \delta_B(x - x_n) dx \ . \tag{2.8b}$$

Since $\delta_B(x)$ decays only as $1/x$, this integral fails to converge for $m \geq 1$, and for $m \geq 2$ it exhibits an unbounded oscillatory behavior. The *center* of these oscillations is the value $I_m = x_n^m$ found in (2.8a), and indeed this is the answer one obtains upon transforming (2.8b) to k space,[34] but this is by no means an adequate substitute for actual convergence. Thus, despite the formal mathematical equivalence between the discrete expression (2.8a) and its quasicontinuum counterpart (2.8b), the latter has such poor convergence properties as to make it totally useless as a numerical tool for calculations of this type.

These examples clearly show the disadvantages of the quasicontinuum formalism in terms of both physical interpretation and mathematical utility. However, given the unique relationship that exists between the quasicontinuum and the original microscopic theory, it is not easy to see how these difficulties could be circumvented without abandoning the exactness of the theory. For the moment, therefore, these problems will be set aside (to be revisited in Sec. V), turning our attention instead to the

derivation of envelope-function equations from the Schrödinger equation.

III. INTEGRAL-EQUATION THEORY

The starting point for the present analysis is the one-dimensional Schrödinger equation

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2}+V(x)\psi(x)=E\psi(x)\ , \tag{3.1}$$

where m is the free-electron mass and $V(x)$ is a nonperiodic microscopic potential. In the Burt envelope-function theory,[16,19] the wave function ψ is expanded over a complete orthonormal set of basis functions that have the periodicity of the underlying Bravais lattice:

$$\psi(x)=\sum_n F_n(x)U_n(x)\ , \tag{3.2}$$

where

$$U_n(x)=U_n(x+a)\ . \tag{3.3}$$

The basis functions U_n have the same form throughout the heterostructure, independent of the local material composition (else they would not be periodic); they are not necessarily the zone-center Bloch functions for *any* of the bulk media making up the structure. Because they are periodic, the U_n have Fourier coefficients only at the reciprocal-lattice vectors. The envelope functions F_n must therefore be restricted to wave vectors within the first Brillouin zone; otherwise, the product in (3.2) leads to overlap in k space, and the envelopes cannot be uniquely determined from the wave function ψ.

In representation (3.2), the Schrödinger equation (3.1) takes the form of an infinite set of coupled integrodifferential equations for the envelopes F_n:[16,19]

$$-\frac{\hbar^2}{2m}\frac{d^2F_n}{dx^2}-\frac{i\hbar}{m}\sum_m p_{nm}\frac{dF_m}{dx}+\sum_m\int W_{nm}(x,x')F_m(x')dx'=EF_n(x)\ , \tag{3.4}$$

where

$$\begin{aligned}
p_{nm}&=\int_{\substack{\text{unit}\\ \text{cell}}}U_n^*\left[-i\hbar\frac{dU_m}{dx}\right]\frac{dx}{a}\ ,\\
W_{nm}(x,x')&=T_{nm}\delta_B(x-x')+V_{nm}(x,x')\ ,\\
T_{nm}&=\int_{\substack{\text{unit}\\ \text{cell}}}U_n^*\left[-\frac{\hbar^2}{2m}\frac{d^2U_m}{dx^2}\right]\frac{dx}{a}\ ,\\
V_{nm}(x,x')&=\int\delta_B(x-x'')U_n^*(x'')V(x'')U_m(x'')\times\delta_B(x''-x')dx''\ .
\end{aligned} \tag{3.5}$$

In these equations p_{nm} is the momentum matrix element (a constant), T_{nm} is the kinetic-energy matrix element (also a constant), and $V_{nm}(x,x')$ is the nonlocal potential-energy operator. Note that V_{nm} is nonlocal (except in bulk media) even though the microscopic potential V is local; also note that every function appearing in Eq. (3.4) is a quasicontinuum function.

The fundamental problem we are now faced with is how to reduce the infinite set of coupled equations (3.4) to a single uncoupled equation without introducing any approximations. The first step is to rewrite Eq. (3.4) in the form of a pure integral equation:

$$\sum_m\int H_{nm}(x,x')F_m(x')dx'=EF_n(x)\ , \tag{3.6}$$

in which the nonlocal Hamiltonian operator $H_{nm}(x,x')$ is given by

$$H_{nm}(x,x')=-\frac{\hbar^2}{2m}\delta_{nm}\delta_B''(x-x')-\frac{i\hbar}{m}p_{nm}\delta_B'(x-x')+W_{nm}(x,x')\ . \tag{3.7}$$

Attention will now be focused on a single band s, which is assumed to overlap no other bands over the energy range of interest (e.g., the Γ_6 conduction band in most direct-gap III-V compounds). The equation of motion (3.6) for this band may be written as

$$\int H_{ss}(x,x')F_s(x')dx'+\sum_r\int H_{sr}(x,x')F_r(x')dx'=EF_s(x)\ , \tag{3.8}$$

in which the second term describes the coupling to all other remote bands r. The equations of motion for these remote bands are

$$\sum_{r'}\int[H_{rr'}(x,x')-E\delta_{rr'}\delta_B(x-x')]F_{r'}(x')dx'+\int H_{rs}(x,x')F_s(x')dx'=0\ . \tag{3.9}$$

To proceed further we need to eliminate the interband coupling by solving Eq. (3.9) for F_r as a function of F_s. This equation can be solved exactly if we define the inverse of a general operator A by the relation

$$\sum_k\int A_{ik}(x,x'')A_{kj}^{-1}(x'',x')dx''=\delta_{ij}\delta_B(x-x')\ , \tag{3.10}$$

which may be written in abbreviated notation as $AA^{-1}=1$. The operator needed to solve Eq. (3.9) is the Green function

$$G=(E1-\tilde{H})^{-1}\ , \tag{3.11}$$

in which $\tilde{H}$ is the matrix obtained by deleting the sth row and column of H. The ability to calculate the inverse (3.11) is not essential and will not be considered in this paper. The primary concern here is the *existence* of such an inverse, since its existence allows us to determine the correct *form* of the solution to (3.9) without performing any detailed calculations. This inverse will fail to exist only if the energy E crosses over into the range of energies represented by the remote bands r. This is not a problem unless the band s is degenerate with one or more of these bands, in which case the band r is not really "remote." Such a degeneracy may occur either within a given medium, as in the Γ_8 valence bands of most zinc-blende compounds, or across the heterojunction, as in the overlap of the Γ_6 and Γ_8 bands that occurs in InAs/GaSb

heterostructures. Within the present formalism these cases must be treated using a multiband approach, which is outside the scope of this paper. It will be assumed in what follows that no such degeneracy exists.

The remote-band envelope function F_r is therefore given by the following solution to Eq. (3.9):

$$F_r(x)=\sum_{r'}\int\int G_{rr'}(x,x')H_{r's}(x',x'')F_s(x'')dx'dx'' . \tag{3.12}$$

This may be substituted into Eq. (3.8) to obtain the desired single-band equation of motion:

$$\int H(x,x')F(x')dx'=EF(x) , \tag{3.13}$$

in which $F\equiv F_s$ is the single-band envelope function and

$$H(x,x')=H_{ss}(x,x')+\sum_{r,r'}\int\int H_{sr}(x,x'')G_{rr'}(x'',x''') \times H_{r's}(x''',x')dx''dx''' \tag{3.14}$$

is the nonlocal single-band Hamiltonian. In Eq. (3.13) all explicit reference to the band index s has been dropped for simplicity.

Within the specified energy range, the integral equation (3.13) is still exactly equivalent to the Schrödinger equation. However, since our ultimate goal is to eliminate the appearance of any explicit nonlocal effects, it is obvious that we will need to find some way to transform this equation into a local differential equation. The derivation of such an equation is considered in Sec. IV below, but before beginning the derivation it is convenient to introduce a few (not very restrictive) assumptions in order to simplify the results. In what follows it will be assumed that: (i) the functions U_n do not merely have the symmetry of the Bravais lattice, but are basis functions for the space group of the underlying crystal structure; (ii) there exists *some* Hamiltonian with the symmetry of this space group that the functions U_n diagonalize; and (iii) the U_n may be chosen to be strictly real.

As an example, consider the case of an $InP/In_{0.53}Ga_{0.47}As$ superlattice, for which the space group of the underlying crystal structure is the zinc-blende group T_d^2. Now imagine the spectrum of Hamiltonians obtained by continuously deforming the microscopic Hamiltonian of bulk InP into that of bulk (virtual-crystal) $In_{0.53}Ga_{0.47}As$. Assumptions (i) and (ii) will be satisfied by taking the U_n to be the zone-center eigenfunctions for any of these Hamiltonians. Assumption (iii) then follows directly from time-reversal symmetry,[38,39] since these are Γ-point basis functions in a zinc-blende crystal.

Using condition (iii) it is easy to show that T_{nm} and V_{nm} are real and p_{nm} is purely imaginary. Since the matrix p_{nm} is both Hermitian and imaginary, its diagonal elements must be zero. In addition, the multiband Hamiltonian H_{nm} is both Hermitian $[H_{nm}(x,x')=H_{mn}^*(x',x)]$ and real, thus it is symmetric $[H_{nm}(x,x')=H_{mn}(x',x)]$. Therefore the single-band Hamiltonian H is also Hermitian, real, and symmetric:

$$H(x,x')=H(x',x) . \tag{3.15}$$

This symmetry property will significantly reduce the complexity of the differential equations considered below.

IV. LOCAL TRANSFORMATION

The appearance of nonlocal interactions in an envelope-function theory is, again, not a new concept (dating back at least to Luttinger and Kohn[2]), but, like the quasicontinuum restrictions on the envelopes, its significance in a heterostructure has been mostly ignored until recently. The first detailed exploration of nonlocal effects at a heterojunction was performed by Burt,[16,19] who showed that the exact equation of motion can be replaced, to a very good approximation, with a local differential equation. The goal of this section is to determine whether it is possible to achieve this result without any approximations. It turns out to be impossible to do this within the quasicontinuum formalism, but in making the attempt we will gain valuable insight into how such a transformation *can* be achieved.

In an inhomogeneous medium, one can use the chain rule to rearrange the terms in a differential equation in a variety of different but equivalent ways. To narrow the scope somewhat, the present analysis will focus solely on attempting to rewrite the integral equation (3.13) in the form of the following infinite series:

$$H_0(x)F(x)+\frac{d}{dx}\left[H_2(x)\frac{dF}{dx}\right] +\frac{d^2}{dx^2}\left[H_4(x)\frac{d^2F}{dx^2}\right]+\cdots=EF(x) . \tag{4.1}$$

The specific form shown here (i.e., that of a generalized Sturm-Liouville equation) is desirable because of its high symmetry; note that there are no odd-order terms, and that the differential operators are symmetric with respect to the various coefficients. The lack of odd-order terms is a consequence of the symmetry (3.15) of the nonlocal Hamiltonian.

We may begin the analysis by transforming the equation of motion (3.13) to k space:

$$\int H(k,k')F(k')dk'=EF(k) . \tag{4.2}$$

Now since the desired form (4.1) contains a series of terms involving products of functions of x, and since a product in x space corresponds to a convolution in k space, it is reasonable to start by trying to isolate the $(k-k')$ dependence of the operator $H(k,k')$. The easiest way to accomplish this is to transform variables by rotating the axes by $\pi/4$ in both x space and k space:[34]

$$X=(x-x')/\sqrt{2},\quad X'=(x+x')/\sqrt{2} ,$$
$$K=(k'+k)/\sqrt{2},\quad K'=(k'-k)/\sqrt{2} . \tag{4.3}$$

To go along with this change of variables, it is convenient to define an operator $\hat{H}$ by the relations

$$\hat{H}(X,X')=H[x(X,X'),x'(X,X')]\ ,$$
$$\hat{H}(K,K')=H[k(K,K'),k'(K,K')]\ . \tag{4.4}$$

The rotated variables are useful because they allow us to distinguish the effects caused by nonlocality (X) from those due to inhomogeneity (X'). (This may be seen from the fact that $\hat{H}$ is independent of X' in a homogeneous medium.) These variables also make it easier to treat the symmetry properties of the Hamiltonian, since the symmetry (3.15) simply means that $\hat{H}$ must be even function of X:

$$\hat{H}(X,X')=\hat{H}(-X,X')\ . \tag{4.5}$$

Therefore, if we rewrite the Fourier transform (2.2) in terms of $\hat{H}$,

$$\hat{H}(K,K')=\frac{1}{2\pi}\int\int e^{-iKX}\hat{H}(X,X')e^{iK'X'}dX\,dX'\ , \tag{4.6}$$

only the even part of the integrand makes a nonzero contribution:

$$\hat{H}(K,K')=\frac{1}{2\pi}\int\int \cos(KX)\hat{H}(X,X')e^{iK'X'}dX\,dX'\ . \tag{4.7}$$

To proceed further we need to write out the Taylor-series expansion for the cosine function in (4.7). Since all terms in this series are functions of K^2X^2, we may then substitute

$$K^2=K'^2+2kk'\ . \tag{4.8}$$

Collecting all terms of the same order in kk', this yields the following series expansion for H:

$$H(k,k')=\frac{1}{\sqrt{2\pi}}B(k)B(k')\sum_{n=0}^{\infty}(-kk')^nH_{2n}(k-k')\ , \tag{4.9}$$

where the coefficients H_{2n} are defined by

$$H_{2n}(k-k')=\frac{2^n}{\sqrt{2\pi}}\int\int e^{iK'X'}\hat{H}(X,X')X^{2n}\times\sum_{j=n}^{\infty}\frac{j!}{n!(j-n)!}\left[\frac{(-iK'X)^{2(j-n)}}{(2j)!}\right]\times dX\,dX'\ . \tag{4.10}$$

Equation (4.9) may now be transformed back to x space, which yields

$$H(x,x')=\sum_{n=0}^{\infty}\int\delta_B^{(n)}(x-x'')H_{2n}(x'')\delta_B^{(n)}(x''-x')dx''\ , \tag{4.11}$$

where $\delta_B^{(n)}(x)=d^n\delta_B/dx^n$, and $H_{2n}(x)$ is the Fourier transform of (4.10). Finally, if we substitute this series expansion into the nonlocal equation of motion (3.13), the following result is obtained:

$$\int\delta_B(x-x')\left\{\sum_{n=0}^{\infty}\frac{d^n}{dx'^n}\left[H_{2n}(x')\frac{d^nF}{dx'^n}\right]\right\}dx'=EF(x)\ . \tag{4.12}$$

This equation, although now very close to the desired form (4.1), is still nonlocal. The "local transformation" advertised in the title of this section has therefore failed to live up to its name. The reason why the last remnant of nonlocality in (4.12) cannot be eliminated is that the product of H_{2n} and $F^{(n)}$ will occupy, in general, a region in k space *twice* the size of the first Brillouin zone. The convolution with respect to δ_B is therefore needed in order to truncate these out-of-zone terms and restore the left-hand side to the quasicontinuum. (The only case in which this convolution is unnecessary is that of a homogeneous medium, where the coefficients H_{2n} are independent of x.) The problem cannot be alleviated, for example, by limiting the envelopes to the inner half of the Brillouin zone,[32] since the left- and right-hand sides of the equation would then still occupy different regions in k space. As a consequence, in an inhomogeneous medium, the exact equation of motion for *any* envelope function of finite bandwidth must necessarily be nonlocal.

This is rather disheartening, since it means that, in addition to the long-range Gibbs oscillations that appear in the coefficients $H_{2n}(x)$, we are now forced to contend with a further smearing effect arising from the convolution in (4.12). If we remain with the quasicontinuum formalism, the only way to eliminate these problems is to discard them outright. The justification for these approximations has already been presented by Burt,[16,19] but it would clearly be to our advantage to reformulate the theory in such a way as to avoid completely the need for any such compromise. This may be achieved through a properly chosen transformation of variables, as demonstrated below.

V. METACONTINUUM

A. Basic properties

The root of all the mathematical problems described in the preceding sections lies in the finite bandwidth of the quasicontinuum functions. This, in turn, can be traced back to the requirement that there be an unambiguous, one-to-one correspondence between each function in the envelope-function space (in this case, the quasicontinuum) and its analog in the microscopic space (i.e., the space of all sufficiently regular square-integrable wave functions ψ). However, the only reason for transforming the theory from the microscopic space to the quasicontinuum in the first place was the resulting simplification of the equations of motion. Now that a number of problems have cropped up in the quasicontinuum, there is certainly nothing that forbids us from transforming the theory yet again in order to eliminate these problems, provided that the transformation is one to one.

The desired transformation would be a mapping of the quasicontinuum, which is defined on the finite interval $-\pi/a\le k\le\pi/a$, onto a space defined on the infinite interval $-\infty\le q\le\infty$, where q plays the same role in this space as k in the quasicontinuum. Ideally, this transformation would also convert oscillatory functions such as $\delta_B(x)$ into more strongly localized functions such as the

Gaussian (e^{-x^2}), but without altering any significant physical content of the theory. A mapping which possesses all of these properties is

$$k=\frac{\pi}{a}\mathrm{erf}\left[\frac{qa}{2\sqrt{\pi}}\right] \quad (-\pi/a\le k\le\pi/a,\ -\infty\le q\le\infty)\ , \tag{5.1}$$

where

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{5.2}$$

is the Gaussian error function. The power series expansion for this mapping is[37]

$$k=q\sum_{n=0}^{\infty}\frac{(-1)^n}{n!(2n+1)}\left[\frac{qa}{2\sqrt{\pi}}\right]^{2n}\ , \tag{5.3}$$

which reduces to $k=q$ in the limit of small q. One can see from the graph in Fig. 2 that q is indeed very close to k in the vicinity of the zone center, the difference becoming appreciable only near the zone boundary.

In addition to the mapping (5.1), we need to establish transformations for the various quasicontinuum functions and equations, the object being that the equation of motion have the same form in both q space and k space. To transform integrals, one needs the relation

$$dk=G^2(q)dq\ , \tag{5.4}$$

where

$$G(q)=e^{-q^2a^2/8\pi} \tag{5.5}$$

is a suitably normalized Gaussian function. The appropriate definitions for q-space functions and operators are

$$\begin{aligned} f(q)&\equiv G(q)F[k(q)]\ ,\\ h(q,q')&\equiv G(q)H[k(q),k'(q')]G(q')\ , \end{aligned} \tag{5.6}$$

where the transformed functions are distinguished notationally through the use of lower-case letters. One may then construct x-space functions through the Fourier transforms

$$\begin{aligned} f(x)&=\frac{1}{\sqrt{2\pi}}\int f(q)e^{iqx}dq\\ f(q)&=\frac{1}{\sqrt{2\pi}}\int f(x)e^{-iqx}dx \end{aligned}\ , \tag{5.7}$$

$(-\infty\le q\le\infty)$ and

$$\begin{aligned} h(x,x')&=\frac{1}{2\pi}\int\int e^{iqx}h(q,q')e^{-iq'x'}dq\,dq'\ ,\\ h(q,q')&=\frac{1}{2\pi}\int\int e^{-iqx}h(x,x')e^{iq'x'}dx\,dx' \end{aligned} \tag{5.8}$$

$[-\infty\le(q,q')\le\infty]$, which are identical in form to the earlier relations (2.1) and (2.2) for the quasicontinuum; note, however, that $f(x)\neq F(x)$ [see Eq. (5.10) below].

An apt name for the resulting envelope-function space would be the "metacontinuum," reflecting both its origin as a transformation of the quasicontinuum and its ability (shown below) to transcend all of the flaws in the quasicontinuum. Note that the defining relation (5.1) for the metacontinuum is not in itself unique, since there exists an infinity of other mappings that would achieve approximately the same effect. [Consider, for example, the mappings $k=(\pi/a)\tanh(qa/\pi)$ and $k=(2/a)\arctan(qa/2)$; the latter is just the bilinear transformation used in digital filter design.[40]] However, once a particular definition has been chosen, it establishes a one-to-one correspondence between the metacontinuum and the quasicontinuum; hence, an envelope-function theory based on the metacontinuum formalism is still exactly equivalent to the original microscopic theory. The specific mapping (5.1) was selected here because of its close relation to the Gaussian function, which has ideal mathematical properties in a number of ways.

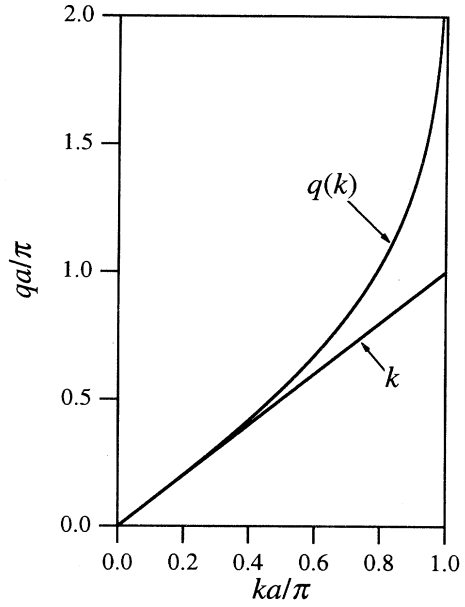

FIG. 2. The metacontinuum mapping function (5.1).

One of the most important of these properties is that definition (5.6) always generates *entire* functions—that is, functions which are analytic (infinitely differentiable) for all $|x|<\infty$. For a function to be entire, it is sufficient[41] that its Fourier transform go to zero faster than $|q|^{-n}$ for any finite n as $|q|\to\infty$, which the factor $G(q)$ in (5.6) obviously does. Not all possible mappings $k(q)$ lead to this property—for example, although $k=(\pi/a)\tanh(qa/\pi)$ does generate entire functions, the choice $k=(2/a)\arctan(qa/2)$ yields functions such as $e^{-|x|}$ with slope discontinuities.

The relationship between the metacontinuum and the quasicontinuum was presented above in Fourier space, but it may also be given directly in x space. If we define the transformation operator

$$T_{GB}(x,x')=\frac{1}{2\pi}\int G(q)e^{iqx}e^{-ikx'}dq\ , \tag{5.9}$$

then the metacontinuum function $f(x)$ is obtained from

the quasicontinuum function $F(x)$ via the integral

$$f(x)=\int T_{GB}(x,x')F(x')dx' , \tag{5.10}$$

which will usually be written as $f(x)=M[F(x)]$ for short. The inverse relation $F(x)=M^{-1}[f(x)]$ is given by

$$F(x)=\int T_{BG}(x,x')f(x')dx' , \tag{5.11}$$

in which

$$T_{BG}(x,x')=T_{GB}(x',x) . \tag{5.12}$$

Although $f(x)$ and $F(x)$ are not equal, they do have the same area, since

$$\begin{aligned}\int F(x)dx &= \sqrt{2\pi}F(k)|_{k=0} \\ &= \sqrt{2\pi}f(q)|_{q=0} \\ &= \int f(x)dx .\end{aligned} \tag{5.13}$$

Also, the area under the product of any two quasicontinuum functions is equal to the area under the corresponding metacontinuum product:

$$\begin{aligned}\int F_1(x)F_2(x)dx &= \int F_1(-k)F_2(k)dk \\ &= \int f_1(-q)f_2(q)dq \\ &= \int f_1(x)f_2(x)dx .\end{aligned} \tag{5.14}$$

In particular, this means that

$$\int |F(x)|^2dx = \int |f(x)|^2dx , \tag{5.15}$$

so the concept of probability density can still be used in the metacontinuum in theory.

B. Transformation of functions

As an example, if we apply transformation (5.10) to the quasicontinuum delta function δ_B, the result is

$$M[\delta_B(x)]=\delta_G(x) , \tag{5.16a}$$

where

$$\begin{aligned}\delta_G(x) &= \frac{\sqrt{2}}{a}e^{-2\pi x^2/a^2} , \\ \delta_G(q) &= \frac{1}{\sqrt{2\pi}}G(q) .\end{aligned} \tag{5.16b}$$

This function, shown in Fig. 3(a), has the desired Gaussian shape in both x space and q space. It is easy to verify that $\delta_G(x)$ and $\delta_B(x)$ satisfy Eqs. (5.13) and (5.15); in particular, both functions are normalized to unit area.

There are many different possible definitions for a metacontinuum step function; three examples are

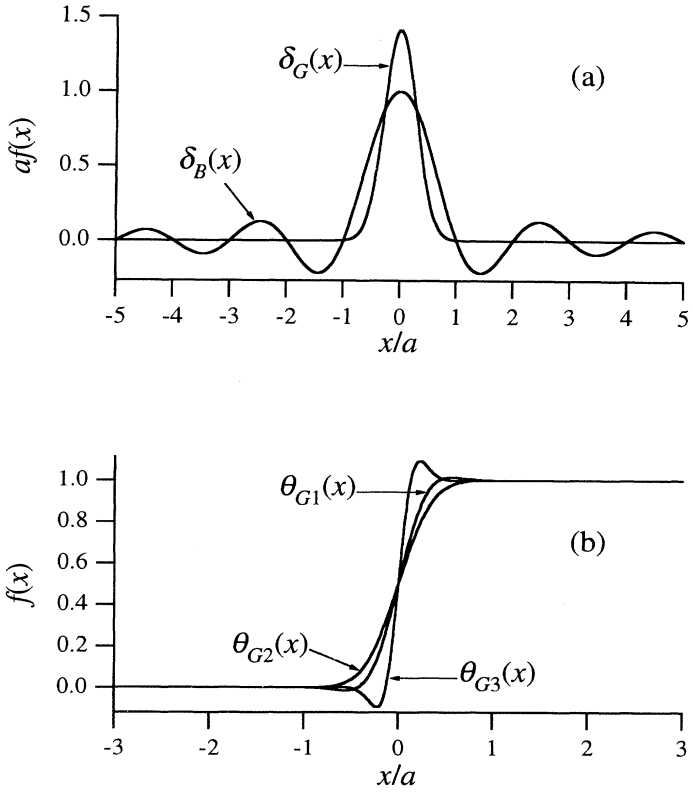

FIG. 3. Metacontinuum functions: (a) δ function, (b) step functions.

$$\theta_{G1}(x)=M[\theta_B(x)]\ ,$$
$$\theta_{G2}(x)=\int_{-\infty}^{x}\delta_G(x')dx'=\tfrac{1}{2}[1+\mathrm{erf}(\sqrt{2\pi}x/a)]\ , \qquad (5.17)$$
$$\theta_{G3}(x)=M\left[a\sum_{n=0}^{\infty}\delta_B(x-na-a/2)\right]\ .$$

These functions are shown in Fig. 3(b). They differ slightly in the range $|x|<a$, but outside this range they are all effectively constant.

The functions (5.16) and (5.17) are useful mainly in describing the behavior of the material properties at an interface. For the electron envelope function itself, consider the following example:

$$\mathcal{f}(x)=e^{-\alpha|x|}\ ,$$
$$\mathcal{f}(k)=\frac{1}{\sqrt{2\pi}}\frac{2\alpha}{\alpha^2+k^2}\quad(-\infty\le k\le\infty)\ , \qquad (5.18)$$

where script letters are used to indicate that this is neither a quasicontinuum nor a metacontinuum function. This type of envelope arises in the second-order effective-mass theory from either an attractive δ-function potential[9] or a change in sign of the effective mass at an interface;[21] it provides an extreme example of the slope discontinuity normally associated with a heterojunction. The envelope (5.18) is shown in Fig. 4 (for the case $\alpha=1/a$) along with its quasicontinuum truncation $[F(k)=B(k)\mathcal{f}(k)]$ and its metacontinuum transform $[f(q)=G(q)\mathcal{f}(k)]$. The quasicontinuum envelope $F(x)$ shows the usual Gibbs oscillations, but the only difference between the metacontinuum envelope $f(x)$ and the original envelope $\mathcal{f}(x)$ is the smoothing of the slope discontinuity at the interface. Therefore, although the definition of the envelope functions may be a bit more complicated in the metacontinuum than in the quasicontinuum, the end result is an envelope which is much closer to that found in the standard effective-mass theory.

This process may be continued indefinitely, but it should be clear from the examples given here that the mapping (5.1) has succeeded in its first goal, which was the elimination of the Gibbs oscillations that plagued the quasicontinuum theory. The metacontinuum functions are strongly localized and well behaved, with the dominant feature being the smoothing of any transitions that occur on a scale more rapid than the lattice constant. Although they have an infinite bandwidth in q space, they drop off very rapidly with increasing q due to the influence of the function $G(q)$.

C. Transformation of operators

The next step is to show that the same conclusions hold true for nonlocal operators. The single-band Hamiltonian $H(x,x')$ is a rather complicated function, so the following examples will be worked out on the simpler multiband Hamiltonian $H_{nm}(x,x')$. The kinetic-energy (T_{nm}) term is easiest, since it is proportional to δ_B, which transforms as

$$M[\delta_B(x-x')]=\delta(x-x')\ ; \qquad (5.19)$$

note the difference between this operator relationship and the earlier relation (5.16) given for functions. The p_{nm}

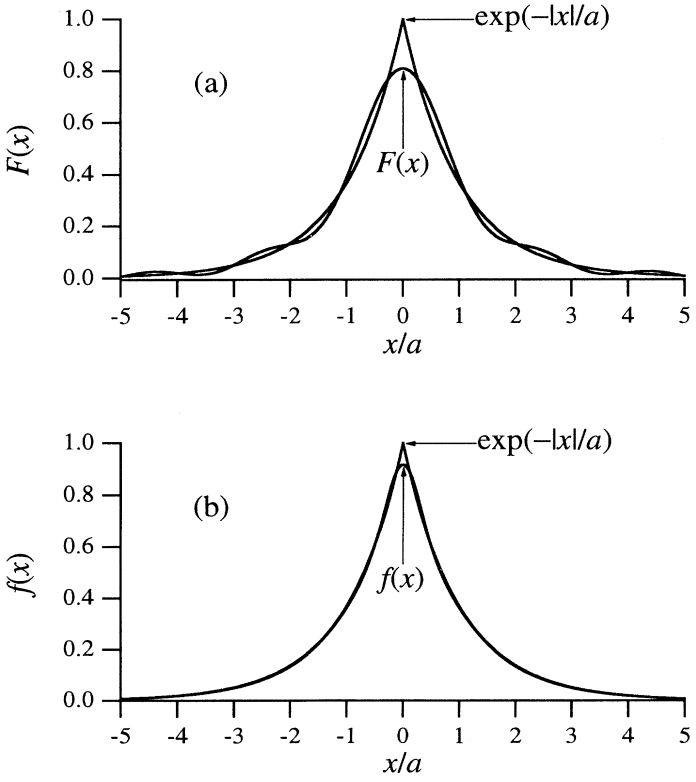

FIG. 4. Envelope function $\mathcal{f}(x)=\exp(-|x|/a)$. (a) Quasicontinuum truncation. (b) Metacontinuum transform.

term is slightly more complicated, because the act of differentiation is not the same in the metacontinuum and the quasicontinuum (since $k \neq q$):

$$M[\delta_B'(x-x')] = \delta'(x-x') + \frac{a^2}{12\pi}\delta'''(x-x') + \cdots = \sum_{n=0}^{\infty} \frac{1}{n!(2n+1)} \left[\frac{a^2}{4\pi} \right]^n \times \delta^{(2n+1)}(x-x') . \quad (5.20)$$

The right-hand side is, however, still strictly local. A similar result may be obtained for the transform of the free-electron term in the Hamiltonian (3.7).

To treat the nonlocal potential $V_{nm}(x,x')$, it is first convenient to define the following microscopic function:

$$V_{nm}^{(\mathrm{mic})}(x) = U_n^*(x) V(x) U_m(x) . \quad (5.21)$$

The definition (3.5) of $V_{nm}(x,x')$ is therefore equivalent to

$$V_{nm}(k,k') = \frac{1}{\sqrt{2\pi}} B(k) B(k') V_{nm}^{(\mathrm{mic})}(k-k') . \quad (5.22)$$

Now in a homogeneous medium, the microscopic potential $V(x)$ is periodic, so this reduces to

$$V_{nm}(k,k') = B(k) B(k') V_{nm}^{[0]} \delta(k-k') , \quad (5.23)$$

where $V_{nm}^{[0]}$ is the average value of $V_{nm}^{(\mathrm{mic})}(x)$. The nonlocal potential is therefore of the form

$$V_{nm}(x,x') = V_{nm}^{[0]} \delta_B(x-x') , \quad (5.24a)$$

or, in the metacontinuum,

$$v_{nm}(x,x') = V_{nm}^{[0]} \delta(x-x') . \quad (5.24b)$$

The situation is not quite so simple if the medium is inhomogeneous. In general we have

$$v_{nm}(x,x') = \frac{1}{(2\pi)^{3/2}} \int\int G(q) e^{iqx} V_{nm}^{(\mathrm{mic})}(k-k') \times G(q') e^{-iq'x'} dq\, dq' . \quad (5.25)$$

To isolate the nonlocal effects we may introduce the axis rotation (4.3), which yields

$$\hat{v}_{nm}(X,X') = \frac{1}{(2\pi)^{3/2}} \int\int G(Q) e^{iQX} V_{nm}^{(\mathrm{mic})}(-\sqrt{2}K') \times G(Q') e^{-iQ'X'} dQ\, dQ' , \quad (5.26)$$

where $Q = (q'+q)/\sqrt{2}$ and $Q' = (q'-q)/\sqrt{2}$. Now if K' were independent of Q, we could integrate with respect to Q immediately and find

$$\hat{v}_{nm}(X,X') \stackrel{?}{=} \frac{1}{\sqrt{2\pi}} \delta_G(X) \times \int G(Q') V_{nm}^{(\mathrm{mic})}(-\sqrt{2}K') e^{-iQ'X'} dQ' . \quad (5.27)$$

The nonlocal interactions in an inhomogeneous medium would therefore have a simple Gaussian dependence. However, K' is not independent of Q:

$$K' = \frac{\pi}{\sqrt{2}a} \left\{ \mathrm{erf}\left[\frac{(Q+Q')a}{2\sqrt{2\pi}} \right] - \mathrm{erf}\left[\frac{(Q-Q')a}{2\sqrt{2\pi}} \right] \right\} . \quad (5.28)$$

The expression (5.27) is therefore not strictly correct. It is nonetheless a good approximation to (5.26), since the integrand of (5.26) is of appreciable magnitude only for small Q and Q'. In this limit K' is approximately independent of Q, since to lowest order (5.28) reduces to $K' = Q'$, with the first corrections being proportional to $(Q')^3$ and $Q'Q^2$. Thus K' is only a weak function of Q, and (5.27) correctly describes the dominant nonlocal behavior of the potential-energy operator in an inhomogeneous medium.

Therefore the metacontinuum transformation yields the same kind of benefits for nonlocal operators that it does for functions. The above discussion treated only the multiband Hamiltonian $H_{nm}(x,x')$, but it is easy to see that the same arguments can be applied to expansion (4.11) for $H(x,x')$; the only difference would be an increase in complexity of the final results.

D. Effects of translation

Despite its obvious advantages, there is an important limitation to the metacontinuum theory having to do with its properties under translation. Examples (5.16)–(5.18) dealt with functions centered at $x=0$, but suppose we now consider what happens when a function $S(x)$ is translated by a distance x_0:

$$R(x) = S(x-x_0) . \quad (5.29)$$

The Fourier transforms of these two functions are related by

$$R(k) = e^{-ikx_0} S(k) , \quad (5.30)$$

i.e., by a phase shift which is linear in k. However, the phase shift between the metacontinuum functions is nonlinear in q:

$$r(q) = e^{-ikx_0} s(q) \neq e^{-iqx_0} s(q) . \quad (5.31)$$

The metacontinuum function $r(x)$ is therefore not the translation of $s(x)$:

$$r(x) \neq s(x-x_0) . \quad (5.32)$$

In fact, the properties of the mapping $k(q)$ show that although the small-q components of r are translated by nearly x_0, its large-q components are translated very little.

A specific example of this behavior is shown in Fig. 5, which compares $M[\delta_B(x-10a)]$ with $\delta_G(x-10a)$. These two functions are obviously not the same, with the phase distortion in the former leading to rapid oscillations at $x=0$ that become progressively slower near $x=10a$. The metacontinuum transformation therefore yields no advantages for translated functions comparable to those seen for functions centered at the origin.

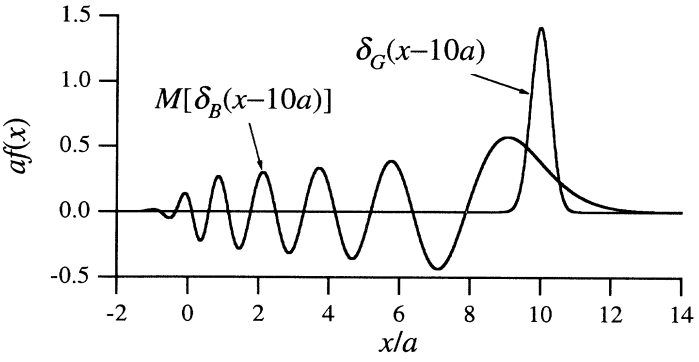

FIG. 5. Effect of transforming a translated function, showing that $M[F(x-x_0)] \neq f(x-x_0)$.

This means that the mapping (5.1) produces results that are directly useful only when it is applied to a single heterojunction. To treat extended structures such as superlattices, we must break the structure up into a series of independent units, each containing only a single junction, and apply the metacontinuum transformation to each of these units individually. For this step to be valid the interfaces must be sufficiently far apart that they are truly physically independent. In other words, the interface separation must be at least twice the interaction range l of the nonlocal Hamiltonian, where l is the minimum distance for which the condition

$$h(x,x') \cong 0 \quad \text{for } |x-x'| \geq l \tag{5.33a}$$

is satisfied for all x. To be more specific, if the parameter $\eta \ll 1$ is deemed negligible, then l is the minimum distance for which the condition

$$\frac{\int_{-\infty}^{x-l} |h(x,x')|dx' + \int_{x+l}^{\infty} |h(x,x')|dx'}{\int_{x-l}^{x+l} |h(x,x')|dx'} \leq \eta \tag{5.33b}$$

is satisfied for all x. (Note that this definition works only in the metacontinuum because the area under $|\delta_B(x)|$ diverges logarithmically, hence the quasicontinuum range l is always infinite.) If the separation between interfaces is less than $2l$, we can still treat them as independent, but this may incur errors of order η or larger due to the neglect of interference effects. The application of the metacontinuum theory to extremely short-period superlattices (period $<4l$) is therefore questionable, although it may turn out to be a good approximation in practice. No such restriction applies to thin quantum wells, however, since there the interfaces need not be separated; a transformation applied to the entire well yields results similar to those found at a single heterojunction.

To get an idea of how large the parameter l is in a typical system, note that many compound semiconductors are well described by an empirical tight-binding model that includes nearest-neighbor interactions only.[42] In these cases a good choice for l would be the second-neighbor distance. This may not be true in all materials, of course, so it is wise to keep in mind the limitations of the theory when applying it to different systems.

In general, if one wishes to apply the metacontinuum transformation to an interface located at $x = x_i$, then the change of variables (5.6) must be modified as follows:

$$\begin{aligned} f(q) &\equiv G(q)e^{i(k-q)x_i}F(k) , \\ h(q,q') &\equiv G(q)e^{i(k-q)x_i}H(k,k')G(q')e^{-i(k'-q')x_i} . \end{aligned} \tag{5.34}$$

This phase shift effectively translates the quasicontinuum functions from x_i back to the origin, applies transformation (5.6), and then returns the metacontinuum functions to the interface location x_i.

The remainder of this paper deals explicitly only with single heterojunctions. Structures such as quantum wells and superlattices are assumed to be obtained by superposition of the basic heterojunction results (with the exception of the thin quantum wells discussed above). The theory is exact for a single junction, with errors due to the neglect of interference between junctions in a superlattice dropping off exponentially with increasing interface separation.

E. Equation of motion

We may turn now to the question of what effect the metacontinuum transformation has on the equation of motion. It is not difficult to see from definitions (5.4) and (5.34) that the q-space equation of motion is formally identical to the k-space equation (4.2):

$$\int h(q,q')f(q')dq' = Ef(q) . \tag{5.35}$$

The Fourier-transform relations (5.7) and (5.8) ensure that this is true in x space as well:

$$\int h(x,x')f(x')dx' = Ef(x) . \tag{5.36}$$

Therefore, to reduce (5.36) to a local differential equation, we need only apply the machinery of the local transformation developed in Sec. IV—but with much greater hope of success this time, since the metacontinuum functions are not band limited.

The equations given in Sec. IV will for the most part not be rewritten here, since they can be obtained easily from the substitution $(F,H,k,K) \rightarrow (f,h,q,Q)$. The first major difference appears in Eq. (4.9), which becomes

$$h(q,q') = \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} (-qq')^n h_{2n}(q-q') . \tag{5.37}$$

Note that the factors $B(k)$ restricting the expansion to the quasicontinuum are no longer present. In essence, this is because $B(k)$ has been replaced with $G(q)$, which has then been absorbed into the definition of $h(q,q')$. Another difference occurs in (4.10):

$$h_{2n}(q-q')=\frac{2^n}{\sqrt{2\pi}}\int\int e^{iQ'X'}\hat{h}(X,X')X^{2n}\sum_{j=n}^{\infty}\frac{j!}{n!(j-n)!}\left[\frac{(-iQ'X)^{2(j-n)}}{(2j)!}\right]dX\,dX' . \tag{5.38}$$

The difference here is that $\hat{H}(X,X')$ has Gibbs oscillations as a function of X, while $\hat{h}(X,X')$ does not. Equation (4.10) will therefore suffer from the convergence problems discussed in Sec. II, none of which occur in (5.38). Explicit formulas for $h_{2n}(x)$ may be found in Appendix A.

Finally, since q is allowed to range over the entire real axis, the quasicontinuum δ functions in (4.11) are replaced with Dirac δ functions:

$$h(x,x')=\sum_{n=0}^{\infty}\int\delta^{(n)}(x-x'')h_{2n}(x'')\delta^{(n)}(x''-x')dx'' . \tag{5.39}$$

The metacontinuum equation of motion corresponding to (4.12) is therefore local, as expected:

$$\sum_{n=0}^{\infty}\frac{d^n}{dx^n}\left[h_{2n}(x)\frac{d^nf}{dx^n}\right]=Ef(x) . \tag{5.40}$$

The transformation to the metacontinuum has thus eliminated both of the major problems (Gibbs oscillations and nonlocal effects) associated with the quasicontinuum.

Equation (5.40) is exactly equivalent to the Schrödinger equation; it represents the central achievement of this paper. Before proceeding any further, we should stop and examine the physical meaning of this equation, in particular the meaning of the coefficients $h_{2n}(x)$. Although the above discussion has focused on the effect of the metacontinuum transformation on interface functions such as δ_B and θ_B, there is also an effect on the bulk coefficients which must be understood if Eq. (5.40) is to be interpreted and used correctly.

In a bulk crystal, nonlocal operators are dependent only on the relative separation of x and x', not on their absolute positions, so they may be reduced to functions of a single variable:

$$H(x,x')=H(x-x'),\quad h(x,x')=h(x-x') . \tag{5.41}$$

This means that in Fourier space, all operators are local:

$$H(k,k')=H(k)\delta(k-k'),\quad h(q,q')=h(q)\delta(q-q') , \tag{5.42}$$

where

$$H(k)=\int H(x)e^{-ikx}dx,\quad h(q)=\int h(x)e^{-iqx}dx . \tag{5.43}$$

As a consequence, the Fourier-space equations of motion (4.2) and (5.35) reduce to

$$E=H(k)=h(q) ; \tag{5.44}$$

hence the functions $H(k)$ and $h(q)$ are nothing more than the energy-versus-wave-vector relation expressed in k space and q space, respectively.

A band diagram illustrating this concept is presented in Fig. 6, where the dispersion relation $H(k)=\frac{1}{2}[1-\cos(ka)]$ was taken from the nearest-neighbor tight-binding theory.[43] This example shows that $h(q)$ follows $H(k)$ very closely near the zone center, but is warped near $q=\pi/a$ as a consequence of the mapping shown in Fig. 2. The coefficients H_{2n} and h_{2n} in the equations of motion (4.12) and (5.40) are simply the Taylor-series coefficients for these two dispersion relations:

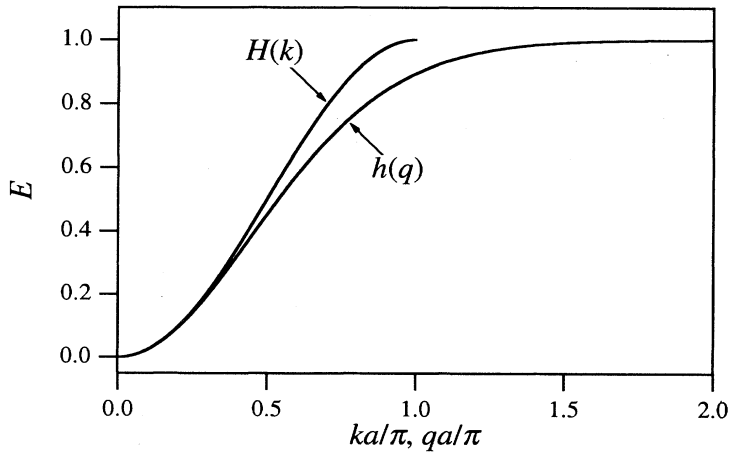

FIG. 6. Energy-band dispersion relation in the quasicontinuum and the metacontinuum.

$$H(k)=\sum_{n=0}^{\infty}(ik)^{2n}H_{2n},\quad h(q)=\sum_{n=0}^{\infty}(iq)^{2n}h_{2n}\ . \tag{5.45}$$

By inserting the power-series expansion (5.3) and equating coefficients of like powers of q on each side of (5.44), one finds that H_{2n} and h_{2n} are related by

$$h_0=H_0,\quad h_2=H_2,\quad h_4=H_4+\frac{a^2}{6\pi}H_2\ , \tag{5.46}$$

and so on. Therefore the first two coefficients, which represent the zone-center energy and the effective mass ($m^*=-\hbar^2/2H_2$), are exactly the same in the metacontinuum and quasicontinuum. The higher-order dispersion terms are somewhat modified in the metacontinuum because of the mapping-related distortion shown in Fig. 6, but as long as it is recognized that the coefficients are to be taken from $h(q)$ and not $H(k)$, there is no opportunity for misunderstanding.

The equation of motion (5.40) may therefore be interpreted physically as a Taylor-series expansion of the dispersion relation $E=h(q)$, with q replaced by the differential operator $-id/dx$, in which explicit account is taken of the spatial variation in band structure arising from any inhomogeneities. With this equation in hand, we can now proceed to derive approximate equations of motion for long-wavelength modes in a particularly simple and straightforward manner.

VI. FINITE-ORDER APPROXIMATIONS

A. Elimination of the higher-order differentials

If we restrict our attention to envelope functions that are slowly varying in the bulk regions of the heterostructure, the approximation needed to reduce Eq. (5.40) to the standard type of phenomenological effective-mass model[10,21] is obviously the elimination of all differential operators higher than, say, the second order. There can be no question of the validity of this step in the bulk, but it seems rather difficult to justify in the neighborhood of a heterojunction in view of the rapid variations that occur in the material properties and (to a lesser degree) the envelope functions. The approach taken here will be to demonstrate that both Eq. (5.40) and its finite-order truncation yield the same macroscopic boundary conditions, hence the higher-order terms in (5.40) have no effect on the macroscopic behavior of the envelopes.

To do this we must first define what is meant by the term "macroscopic boundary conditions." The standard approach to deriving interface boundary conditions from a differential equation[44] is to integrate the equation from $x=x_i-\varepsilon$ to $x=x_i+\varepsilon$ (where x_i is the interface location) and then take the limit as $\varepsilon\to 0$. If this procedure is applied to Eq. (5.40) we obtain nothing but the tautology $0=0$, since it is already known that every function in this equation is infinitely differentiable. However, if we instead take ε to be a *finite* number of the order of the lattice constant, the integral provides a measure of the impact of the interface on the behavior of the envelopes, since Fig. 3 shows that the interface exerts its influence over distances of this range.

How large, exactly, should the parameter ε be? The precise value is not too important, but in general it should be no less than the nonlocal interaction range l defined in Eq. (5.33). If a nearest-neighbor tight-binding model is valid,[42] for example, then l would be the second-neighbor distance. If no specific model has been adopted but the bulk energy-band structure is known, then Eqs. (5.43) and (5.44) show that l may be determined from a Fourier analysis of the dispersion relation. In any event, Fig. 3 shows that the minimum acceptable value for ε in the metacontinuum formalism is the lattice constant a.

With the parameter ε chosen according to these criteria, we may divide the space surrounding a heterojunction into two regions: the interface region $|x-x_i|<\varepsilon$, where the material parameters h_{2n} undergo rapid changes; and the bulk regions $|x-x_i|\geq\varepsilon$, where these coefficients are virtually independent of x. (Such a separation is easy to perform in the metacontinuum but very difficult in the quasicontinuum, as should be apparent from Figs. 1 and 3.) Then if we integrate the equation of motion (5.40) across the interface region, the following result is obtained:

$$\int_{x_i-\varepsilon}^{x_i+\varepsilon}[h_0(x)-E]f(x)dx + \sum_{n=1}^{\infty}\frac{d^{n-1}}{dx^{n-1}}\left[h_{2n}(x)\frac{d^nf}{dx^n}\right]\Bigg|_{x_i-\varepsilon}^{x_i+\varepsilon}=0\ . \tag{6.1}$$

Note that the interface properties appear *only in the first term.* Since the spatial dispersion terms in (5.40) are exact differentials, they appear in (6.1) only in terms of their values at the edges of the bulk regions. Therefore, since ε is small on the macroscopic scale (where macroscopic means large compared to l), Eq. (6.1) can be interpreted as a *macroscopic boundary condition* relating the behavior of f in one bulk region $x\leq x_i-\varepsilon$ to its behavior in the other bulk region $x\geq x_i+\varepsilon$.

If the envelope function f is slowly varying in the bulk regions, where the words "slowly varying" mean

$$\left|\frac{df}{dx}\right|\ll\frac{1}{l}|f| \tag{6.2}$$

(provided f is not too close to a zero crossing), then the series (6.1) will converge quite rapidly. Suppose that we decide for this reason to approximate (6.1) by its $2N$th-order truncation, where $N\geq 1$:

$$\int_{x_i-\varepsilon}^{x_i+\varepsilon}[h_0(x)-E]f(x)dx + \sum_{n=1}^{N}\frac{d^{n-1}}{dx^{n-1}}\left[h_{2n}(x)\frac{d^nf}{dx^n}\right]\Bigg|_{x_i-\varepsilon}^{x_i+\varepsilon}=0\ . \tag{6.3}$$

This boundary condition, however, is just that obtained from integrating the truncated differential equation

$$\sum_{n=0}^{N}\frac{d^n}{dx^n}\left[h_{2n}(x)\frac{d^nf}{dx^n}\right]=Ef(x)\ . \tag{6.4}$$

Thus, despite the *apparent* importance of the higher-

order derivatives at an interface, they have no significant influence on the macroscopic behavior of the envelopes. In effect, we can treat the differential equation (5.40) as if it contained only slowly varying quantities—provided that any rapid variations are limited to regions small on the macroscopic scale. This explains why the standard phenomenological theory,[10,21] in which equations such as (6.4) are adopted without any rigorous justification, often gives such good results.

The above argument establishes the validity of using finite-order differential equations to describe a heterostructure. Equation (6.4) is not, however, especially well suited for practical calculations, since the metacontinuum coefficients $h_{2n}(x)$ have a rather complicated functional dependence within the interface region. What we need for practical use is an envelope-function theory that effectively shrinks the interface region down to a single point ($x=x_i$), so that the material coefficients are constant everywhere except at this point. The single-point connection rules in this simpler theory should be fully equivalent to the distributed boundary condition (6.3).

B. Single-point connection rules

To achieve this goal, we need to rewrite the macroscopic boundary condition so that it involves $f(x)$ and its derivatives only at the interface location x_i. The exact expression (6.1) rather than its approximation (6.3) will be used here so as to obtain a more general result; the finite-order approximation will then be reintroduced at the end of the analysis. Consider first the interface integral. We can separate the coefficient h_0 into distinct bulk and interface parts via the definition

$$h_0(x)=h_{0b}(x)+h_{0i}(x) , \tag{6.5}$$

where h_{0b} is a function that steps abruptly between the two bulk values of h_0,

$$\begin{aligned} &h_{0b}(x)=h_{0-}\theta(x_i-x)+h_{0+}\theta(x-x_i) , \\ &h_{0\pm}=h_0(x_i\pm\varepsilon) , \end{aligned} \tag{6.6}$$

and h_{0i} is zero outside the interface region [to within terms of order η, where η is the parameter given in (5.33b)]. We may also use a Taylor series to represent $f(x)$ at any position in terms of its derivatives at x_i:

$$f(x)=\sum_{n=0}^{\infty}\frac{1}{n!}(x-x_i)^n f^{(n)}(x_i) . \tag{6.7}$$

This expansion converges absolutely and uniformly for all $|x|<\infty$, since $f(x)$ is an entire function. The interface integral in (6.1) may therefore be rewritten as

$$\int_{x_i-\varepsilon}^{x_i+\varepsilon}[h_0(x)-E]f(x)dx=\sum_{n=0}^{\infty}\tilde{h}_{in}(\varepsilon)f^{(n)}(x_i) , \tag{6.8}$$

in which

$$\begin{aligned} \tilde{h}_{in}(\varepsilon)=&\frac{1}{n!}\int(x-x_i)^n h_{0i}(x)dx \\ &+\frac{\varepsilon^{n+1}}{(n+1)!}[(h_{0+}-E)+(-1)^n(h_{0-}-E)] . \end{aligned} \tag{6.9}$$

This coefficient is a functional of $h_{0i}(x)$ and a function of both the interface width ε and (for n even) the energy E. The limits on the integral in (6.9) have been omitted since $h_{0i}(x)$ is zero outside the interface region.

The macroscopic boundary condition (6.1) is now of the form

$$\begin{aligned} \sum_{n=0}^{\infty}\tilde{h}_{in}(\varepsilon)f^{(n)}(x_i)+&\sum_{n=1}^{\infty}h_{2n+}f^{(2n-1)}(x_i+\varepsilon) \\ &-\sum_{n=1}^{\infty}h_{2n-}f^{(2n-1)}(x_i-\varepsilon)=0 . \end{aligned} \tag{6.10}$$

The second and third terms in this equation have been simplified by noting that $h_{2n}(x)$ is constant (to order η) for $|x-x_i|\geq\varepsilon$. We may therefore treat these terms using the same Taylor-series expansion method as before:

$$f^{(m)}(x_i\pm\varepsilon)=\sum_{n=0}^{\infty}\frac{(\pm\varepsilon)^n}{n!}f^{(n+m)}(x_i) . \tag{6.11}$$

In the second term, for example, we have

$$\begin{aligned} &\sum_{n=1}^{\infty}h_{2n+}f^{(2n-1)}(x_i+\varepsilon) \\ &\quad=\sum_{n=1}^{\infty}h_{2n+}\sum_{j=0}^{\infty}\frac{\varepsilon^j}{j!}f^{(j+2n-1)}(x_i) \\ &\quad=\sum_{n=1}^{\infty}f^{(n)}(x_i)\sum_{j=1}^{(n+1)/2}\frac{\varepsilon^{n+1-2j}}{(n+1-2j)!}h_{2j+} . \end{aligned} \tag{6.12}$$

The notation in the last line implies a sum over all integers j that fall in the range $1\leq j\leq(n+1)/2$, even though the upper limit is not an integer if n is even.

If the third term in (6.10) is also rearranged in this manner, the macroscopic boundary condition (6.1) can be cast in the following form:

$$\begin{aligned} &\sum_{n=0}^{\infty}h_{in}(\varepsilon)f^{(n)}(x_i) \\ &\qquad+\sum_{n=1}^{\infty}\frac{d^{n-1}}{dx^{n-1}}\left[h_{2nb}(x)\frac{d^n f}{dx^n}\right]\Bigg|_{x_i-0}^{x_i+0}=0 , \end{aligned} \tag{6.13}$$

where

$$\begin{aligned} h_{in}(\varepsilon)=\tilde{h}_{in}(\varepsilon)+&\sum_{j=1}^{n/2}\frac{\varepsilon^{n+1-2j}}{(n+1-2j)!} \\ &\times[h_{2j+}+(-1)^n h_{2j-}] . \end{aligned} \tag{6.14}$$

This is a very useful form for the boundary condition because it reduces the net effect of the interface region to an action at a *single point.* In other words, the boundary condition (6.13) may be obtained by replacing the metacontinuum material coefficients in (5.40) with Heaviside step functions and Dirac δ functions according to the prescription

$$\begin{aligned} &h_0(x)\rightarrow h_{0b}(x)+\sum_{n=0}^{\infty}(-1)^n h_{in}(\varepsilon)\delta^{(n)}(x-x_i) , \\ &h_{2n}(x)\rightarrow h_{2nb}(x)\quad(n\geq 1) , \end{aligned} \tag{6.15}$$

as can be seen by integrating (5.40) across an infinitesimal region centered on the point $x=x_i$. Although (6.13) has been written in a very suggestive form, this infinite-order boundary condition still reduces to the continuity of f and all of its derivatives. Its true significance becomes apparent only in a finite-order theory.

If we truncate the equation of motion to terms of order $2N$ as in (6.4), then the truncation of (6.13) can be interpreted as specifying the discontinuity in $f^{(2N-1)}(x)$ at $x=x_i$. Therefore, since $f^{(2N-1)}(x_i)$ is mathematically undefined, we must limit the δ-function series in (6.15) to terms of order $2N-2$. The metacontinuum material coefficients in the equation of motion (6.4) are therefore to be replaced according to the rule

$$h_0(x)\rightarrow h_{0b}(x)+\sum_{n=0}^{2N-2}(-1)^n h_{in}(\varepsilon)\delta^{(n)}(x-x_i)\ ,$$
$$h_{2n}(x)\rightarrow h_{2nb}(x)\quad (1\le n\le N)\ . \tag{6.16}$$

This means that the boundary conditions at $x=x_i$ require the continuity of f and its first $2N-2$ derivatives, with a discontinuity in $f^{(2N-1)}(x)$ given by

$$\sum_{n=0}^{2N-2} h_{in}(\varepsilon)f^{(n)}(x_i) + \sum_{n=1}^{N}\frac{d^{n-1}}{dx^{n-1}}\left[h_{2nb}(x)\frac{d^n f}{dx^n}\right]\Bigg|_{x_i-0}^{x_i+0}=0\ . \tag{6.17}$$

There may appear to be a slightly inconsistency here, since the second (bulk) series in (6.17) includes derivatives of order $2N-1$, whereas the first (interface) series runs only to order $2N-2$. However, this problem can be eliminated through a proper choice of the interface location x_i, as shown below.

Ordinarily one chooses the interface location based on the geometry of the lattice; for example, at an InAs/GaSb interface, x_i would be placed at the midpoint of the interface cation and anion, while at a GaAs/AlAs interface, x_i would be placed at the interfacial As ion. However, as we have seen, the interface is not a point but a finite region, so we have the flexibility to place x_i anywhere within this region that we choose. It is most convenient to choose x_i so that the following condition is satisfied:

$$h_{i(2N-1)}(\varepsilon)=0\ . \tag{6.18}$$

With this choice we see that the boundary condition (6.17) incorporates derivatives through order $2N-1$ in both the interface and the bulk series.

The results presented here may be used as the basis for any finite-order theory, but the second-order approximation ($N=1$) is used most frequently in practice, so it is worthwhile to take a closer look at this special case. The second-order equation of motion is

$$h_0(x)f(x)+\frac{d}{dx}\left[h_{2b}(x)\frac{df}{dx}\right]=Ef(x)\ , \tag{6.19}$$

in which the band-edge potential $h_0(x)$ consists of step-function bulk term and a δ-function interface term:

$$h_0(x)=h_{0b}(x)+h_{i0}\delta(x-x_i)\ . \tag{6.20}$$

The weight of the δ-function potential is defined by

$$h_{i0}=\int h_{0i}(x)dx+\varepsilon(h_{0+}+h_{0-}-2E)\ , \tag{6.21}$$

which has two components: a constant term equal to the area under the interface part of $h_0(x)$, and an energy-dependent term proportional to the deviation of the energy from the average of the two bulk potentials. The first term was to be expected, but the second term is a bit surprising, since it seems that by including this term here we are counting the effect of the bulk function $h_{0b}(x)-E$ twice. It will be shown in a moment that this is not actually so, but the reader can convince himself or herself of the need for the energy-dependent term by performing a very simple exercise: calculating the ground-state energy of a square quantum well and comparing it to the result obtained from the δ-function potential (6.21), both with and without the second term. This calculation shows that the neglect of the energy-dependent term gives a poor result unless the quantum well is very deep and narrow to begin with. Another example leading to the same conclusion is provided in the companion paper on phonons.[36]

The boundary conditions derived from (6.19) require the continuity of $f(x)$ along with a slope discontinuity that is determined primarily by the interface term (6.21):

$$h_{i0}f(x_i)+h_{2+}f'(x_i+0)-h_{2-}f'(x_i-0)=0\ . \tag{6.22}$$

Note that in a second-order theory, the slope of f is constant over distances of order ε, so we have $f'(x_i+0)=f'(x_i+\varepsilon)$ and $f'(x_i-0)=f'(x_i-\varepsilon)$. The inclusion of bulk effects in (6.21) is therefore correct to within terms of the second order. (The same conclusion holds in a general $2N$th-order theory, where $f^{(2N-1)}$ is constant over distances of order ε.)

The interface position is chosen according to requirement (6.18):

$$x_i=\frac{\int x h_{0i}(x)dx+\frac{1}{2}\varepsilon^2(h_{0+}-h_{0-})}{\int h_{0i}(x)dx}\ . \tag{6.23}$$

If we assume that $h_0(x)$ is of the simple form shown in Fig. 7, i.e., constant within the interface region (which

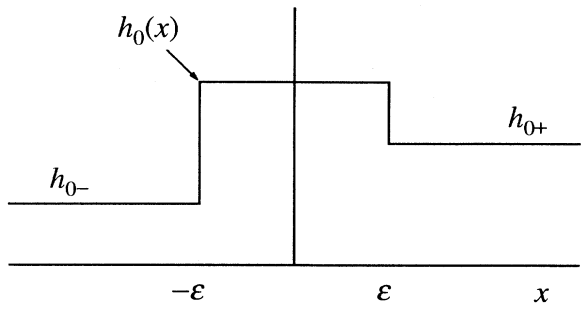

FIG. 7. A simple model for the interface dependence of the coefficient $h_0(x)$.

obviously cannot be exact, since metacontinuum functions are not discontinuous), then an evaluation of the integral (6.23) yields

$$x_i = 0 \ . \tag{6.24}$$

In other words, for this simple case, the interface position that satisfies criterion (6.18) is no different from the one that we would have chosen anyway on the basis of geometrical considerations. A more accurate description of $h_0(x)$ may not yield precisely this result, but it is reasonable to assume that the interface displacement will be small in general (i.e., small compared to ε), so there is little to be lost by ignoring this effect entirely and choosing x_i to be at the geometrical interface.

The second-order equation of motion (6.19) is of the same form as that found in the widely used BenDaniel-Duke theory,[4,21] with the exception of the δ-function interface potential in (6.20). This term is usually neglected in effective-mass calculations on heterostructures (although it has been treated in many of the more complex microscopic-based theories[8,9,13,19,22,26,27,30]), and it is natural to wonder how much error arises from this approximation. The main question is whether the δ-function potential is strong enough to generate interface modes of the form shown in Fig. 4; if it is not, then it can safely be discarded, since no qualitative error will arise from its neglect. It is shown elsewhere[19,45] that this is the case in many common III-V compound heterostructures, in which the media are weakly polar and chemically similar (excluding, of course, structures in which the single-band treatment is not valid). No explicit examples will be worked out in this paper, however. Readers interested in seeing practical applications of the present theory are referred to the accompanying paper on phonons,[36] where the example of optical phonons in an InAs/GaSb superlattice is treated in some detail. (Note that the discrete theory of lattice dynamics is mathematically the same as the tight-binding theory for electrons.) For this case, the δ-function potential cannot be neglected, since it generates mechanical interface modes that lie well outside the bulk optical spectra of both InAs and GaSb.

One of the more interesting effects of the δ-function potential will occur in the multiband theory, since this potential can generate a coupling between bands that would not otherwise exist. As an example, consider the Γ-X coupling that is known to exist in GaAs/AlAs heterostructures.[22,26,46] The present single-band theory is in principle applicable to this situation, but would not be useful in practice because no finite-order approximation would be valid. For practical calculations it is better to use the method suggested by Burt,[19] in which the basis functions U_n are chosen to have a period of $2a$ rather than a. This effectively cuts the Brillouin zone in half, folding the zone boundary over onto the zone center, so the Γ-X coupling can be treated as a two-band zone-center problem. The simplest approximation that includes any Γ-X coupling would retain only the diagonal effective-mass terms and the off-diagonal δ-function potential, thereby coupling the slope of the Γ envelope to the amplitude of the X envelope and vice versa. This is just the model used by Liu[46] to describe resonant tunneling in GaAs/AlAs barrier structures.

A similar effect would occur in the valence bands of zinc-blende heterostructures. The standard effective-mass treatment of this problem[10,12,47] uses only step-function material parameters, hence there is no coupling between the heavy- and light-hole bands at $\mathbf{k}_{\|}=\mathbf{0}$. However, Edwards and Inkson[48] have shown, using empirical pseudopotential calculations, that such coupling does in fact exist due to the change in zone-center Bloch functions at the interface. The addition of a δ-function coupling potential to the standard effective-mass model may be sufficient to reproduce their results, although this is yet to be investigated.

Another interesting property of the present second-order theory is that it can describe nonparabolic effects through the energy dependence of the coefficients h_0 and h_2. This is a considerable advantage, since the nonparabolic second-order theory is accurate over a much wider range of energy than the usual parabolic theory, hence excellent results can often be obtained without the need to include any fourth- or higher-order corrections. (For a specific example, see the phonon calculations in Ref. 36.) The energy dependence of the material coefficients arises primarily from the use of Luttinger-Kohn rather than Wannier-Slater basis functions in Eq. (3.2). A simple method for calculating these nonparabolic effects is described in Appendix B.

VII. SUMMARY AND CONCLUSIONS

This paper has presented a method for deriving effective-mass equations which leads to an exact effective-mass theory for electrons in heterostructures. The derivation started from the Schrödinger equation. The Burt envelope-function theory was used to obtain a set of exact integral envelope-function equations; the interband coupling was then eliminated using a Green-function decoupling technique that allowed the correct form of the solution to be established without the need for any detailed calculations. The resulting single-band integral equation was exact, but its mathematical properties were found to be less than ideal, since it was unavoidably nonlocal and the perturbations arising at an interface were very long range. The source of these troubles was the finite bandwidth of the envelope functions. The problem was resolved using a transformation of variables in Fourier space that mapped the Brillouin zone onto an unrestricted space, thereby allowing the wave vector to range over all real values. The equation of motion in this metacontinuum space had the desired form of a local infinite-order differential equation.

Only two assumptions were needed to obtain this equation. The first was that the Green function used to eliminate the interband coupling existed over the entire energy range of interest. This assumption is valid if no band degeneracy occurs within this range, either within a given medium or across the heterojunction. If a degeneracy does occur, then it is necessary to move to a multiband formalism, which increases the algebraic complexity of the theory but involves no new concepts.

The second assumption was that the separation between heterojunctions in a superlattice was no less than twice the interaction range l of the nonlocal Hamiltonian. This restriction is necessary since the transformation of variables gives useful results only when it is applied to a single heterojunction (or thin quantum well), so we must be able to break the heterostructure down into a set of physically independent units. The theory is therefore exact only for a single heterojunction (or thin quantum well), but the error due to the neglect of interference between junctions drops off exponentially with increasing junction separation, so this becomes a limiting factor only in the case of extremely short-period superlattices.

Several auxiliary assumptions were used to simplify the differential equation by eliminating its odd-order terms, but these were not essential to the derivation. The argument was based on time-reversal symmetry as applied to the Γ-point basis functions in a zinc-blende crystal, hence it is not limited to the one-dimensional case; it does, however, require the neglect of spin-orbit coupling.

Macroscopic boundary conditions were derived by integrating the differential equation across the (finite) interface region and then translating the results back to a central point chosen to represent the interface. This effectively replaces the spatial dispersion coefficients with abrupt step functions, and the band-edge potential with an abrupt step plus a series of terms proportional to the Dirac δ function and its derivatives. The connection rules in a theory of $2N$ require the continuity of the envelope $f(x)$ and its first $2N-2$ derivatives, with a discontinuity in $f^{(2N-1)}(x)$ determined by the δ-function terms and the change in bulk properties across the interface. The second-order theory reduces to the well-known BenDaniel-Duke model,[4,21] but with a δ-function potential at the interface. (It is interesting to note that many pronouncements in the literature regarding the failure of effective-mass theory in a given situation are due not to any failure of the second-order differential equation, nor to the assumed continuity of the envelope functions, but rather to the assumption that the material parameters vary in a stepwise fashion across the interface.) Nonparabolic effects are included in the second-order theory through the energy dependence of the material parameters.

To obtain a broader perspective on the present theory, it is helpful to place it in the context of previous work. The traditional approach to heterostructures applies a different envelope-function expansion in each medium, invariably choosing the microscopic basis functions to be the local band-edge Bloch functions. In this approach the envelopes are always discontinuous at an interface, so any theory that assumes continuity of the envelopes is viewed as an approximation. However, it has been shown by Smith and Mailhiot,[14] Burt,[15–20] Elçi,[30] and others that significant mathematical advantages can be gained by using the *same* basis functions throughout the structure, not the least of which is the automatic continuity of the envelopes at an interface. This continuity is *not* an approximation (although it is often misinterpreted as such[49,50]), since the change in band-edge Bloch functions is included through the off-diagonal matrix elements in the Hamiltonian. Perhaps the most important conclusion that can be drawn from this change of basis is that the concept of an envelope function need not be tied rigidly to the original Wannier-Slater or Luttinger-Kohn definition; it can instead be adapted at will to meet the needs of the problem at hand.

The present work is merely another step along the same path. By applying a judiciously chosen transformation of variables, the approximations in the Burt theory involving the neglect of Gibbs oscillations and nonlocal effects can be completely eliminated. Although the definition of the metacontinuum envelopes may seem unfamiliar at first, the material parameters and envelope functions that result from this definition are actually very closely related to those found in the standard effective-mass theory, the main difference being the smoothing of abrupt transitions at an interface. The only real cost to this transformation is a slight change in interpretation of the band structure (since the Brillouin zone is now infinite); however, this change affects only fourth- and higher-order terms, so in the usual second-order approximation it need not even be considered.

ACKNOWLEDGMENTS

I am grateful to Brian Ridley, Mike Burt, Lester Eastman, Frank Wise, Sean O'Keefe, and Robert Spencer for helpful discussions. This work was supported by AT&T Bell Laboratories and by Rome Laboratory of the U.S. Air Force under Contract No. F30602-93-C-0249.

APPENDIX A: MATERIAL COEFFICIENTS

The coefficients $h_{2n}(x)$ in the differential series (5.40) are simply the Fourier transforms of the $h_{2n}(q)$ given in (5.38). The first two terms are easy to evaluate directly:

$$h_0(x)=\int h(x,x')dx' ,$$
$$h_2(x)=\int\int (x''-x')\theta(x-x')h(x',x'')dx'dx'' , \tag{A1}$$

where $\theta(x)$ is the Heaviside step function (2.7). However, simple solutions for the fourth- and higher-order coefficients are rather difficult to obtain in this manner because of the complicated nature of the definition (5.38). It is easier in these cases to take a more indirect route, namely multiplying the expansion (5.39) by $(x'-x)^m$ and integrating with respect to x'. This yields the relation

$$\bar{h}_m(x)=\sum_{n=m/2}^{m}\frac{n!}{(m-n)!(2n-m)!}\frac{d^{2n-m}h_{2n}}{dx^{2n-m}} , \tag{A2}$$

where $\bar{h}_m$ is a related coefficient defined by

$$\bar{h}_m(x)=\frac{1}{m!}\int (x'-x)^m h(x,x')dx' . \tag{A3}$$

An explicit evaluation of (A2) for terms $\bar{h}_0$ through $\bar{h}_4$ gives

$$\bar{h}_0=h_0,\bar{h}_1=\frac{dh_2}{dx},\bar{h}_2=\bar{h}_2=h_2+\frac{d^2h_4}{dx^2},$$
$$\bar{h}_3=2\frac{dh_4}{dx}+\frac{d^3h_6}{dx^3},\bar{h}_4=h_4+3\frac{d^2h_6}{dx^2}+\frac{d^4h_8}{dx^4}. \quad \text{(A4)}$$

Note that in each successive equaiton, only one additional coefficient is introduced on the right-hand side. These equations may therefore be used to calculate the coefficients $h_{2n}(x)$ by direct integration. The first two terms obtained in this way are, of course, identical to (A1).

APPENDIX B: NONPARABOLIC EFFECTS

In general the basis functions U_n used in the present paper do not diagonalize the crystal Hamiltonian for a given bulk medium, so the calculation of the bulk material parameters h_{2n} is slightly more complicated than in the usual diagonal case. For example, the band-edge potential h_0 is defined in (A1) and (3.14), so for bulk media we have

$$h_0=w_{ss}^{[0]}+\sum_{r,r'}w_{sr}^{[0]}g_{rr'}^{[0]}w_{r's}^{[0]}. \quad \text{(B1)}$$

Here the bracketed superscripts on the matrix operators have the same meaning as the subscript on h_0. Note that the second term is energy dependent, since the zeroth-order matrix $g_{rr'}^{[0]}$ is defined by $g^{[0]}=(E1-\bar{w}^{[0]})^{-1}$. This energy dependence vanishes if the U_n are chosen to diagonalize the Hamiltonian (so that $w_{nm}[0]=E_n\delta_{nm}$, where E_n is the zone-center energy of the nth state); in this case we have simply the standard result from the $\mathbf{k}\cdot\mathbf{p}$ theory:

$$h_0=w_{ss}^{[0]}=E_s. \quad \text{(B2)}$$

Expressions (B1) and (B2) always coincide at the zone-center energy $E=E_s$, but they differ away from zone center due to the energy dependence of (B1).

To calculate the effective mass $m^*\equiv-\hbar^2/2h_2$, we may note from Eqs. (5.43) and (5.45) that the bulk material coefficients h_n are given by

$$h_n=\frac{1}{n!}\int(-x)^n h(x)dx, \quad \text{(B3)}$$

where $h(x)$ is defined in (5.41). Equation (B3) may be used in conjunction with (3.14) to find the following expression for the bulk effective mass:

$$\frac{m}{m^*}=1+\frac{2}{m}\sum_{r,r'}p_{sr}g_{rr'}^{[0]}p_{r's}-\frac{4}{\hbar}\sum_{r,r'}\mathrm{Im}(p_{sr}g_{rr'}^{[1]}w_{r's}^{[0]})-\frac{2m}{\hbar^2}\sum_{r,r'}w_{sr}^{[0]}g_{rr'}^{[2]}w_{r's}^{[0]}. \quad \text{(B4)}$$

The only unknown quantities in (B4) are the first- and second-order matrices $g_{rr'}^{[1]}$ and $g_{rr'}^{[2]}$, which may be calculated in terms of the lower-order parameters using a matrix power-series expansion:[45]

$$g_{rr'}^{[1]}=-\frac{i\hbar}{m}[g^{[0]}pg^{[0]}]_{rr'},$$
$$g_{rr'}^{[2]}=-\frac{\hbar^2}{2m}[g^{[0]}g^{[0]}]_{rr'}-\frac{\hbar^2}{m^2}[g^{[0]}pg^{[0]}pg^{[0]}]_{rr'}. \quad \text{(B5)}$$

Thus for the general nondiagonal case, we can use (B4) and (B5) to determine the energy dependence of m^* using only the matrices p and $w^{[0]}$.

If $w^{[0]}$ is diagonal, then so is $g^{[0]}$, while $g^{[1]}$ and $g^{[2]}$ vanish entirely. Equation (B4) therefore reduces to the familiar $\mathbf{k}\cdot\mathbf{p}$ expression

$$\frac{m}{m^*}=1+\frac{2}{m}\sum_r\frac{p_{sr}p_{rs}}{E-E_r}. \quad \text{(B6)}$$

Nonparabolic effects are therefore included in the present model in even the simplest second-order approximation. This capability arises from the use of Luttinger-Kohn basis functions in the original envelope-function definition (3.2); if Wannier-Slater envelope functions had been used,[19] the energy E in (B6) would be replaced by its zone-center value E_s.

The diagonal expressions (B2) and (B6) are much simpler than their nondiagonal counterparts (B1) and (B4), since no matrix calculations are needed. Furthermore, the parameters E_n and p_{rs} in (B2) and (B6) are readily available in tables of bulk experimental data, whereas a calculation of the matrices p and $w^{[0]}$ in (B1) and (B4) requires knowledge of the microscopic basis functions U_n for the band s and all of the remote bands of interest. It is therefore preferable to use the diagonal formulas whenever possible, but this choice is strictly permissible in only one of the bulk media making up the heterostructure. In the remaining media, if strict accuracy is desired, one has no choice but to use the complicated nondiagonal formulas.

Such rigor is usually unnecessary, however, since the diagonal expressions provide an approximation good enough for most practical purposes. Equations (B1) and (B2) coincide at zone center, and although (B4) and (B6) do not [since the energy-dependent part of (B1) includes second-order effects that are conventionally described by an effective mass], the difference is often quite small[19,20] and can usually be neglected. The reason for this is that the off-diagonal bulk matrix elements $w_{nm}^{[0]}$ are nonvanishing only between states of the same symmetry, which are always well separated in energy, typically by several eV. Therefore, even though the magnitude of the off-diagonal terms is significant (typically several hundred meV),[19] the fractional difference between (B4) and (B6) at zone center is small because it varies as the square of the ratio of these two energies. (For specific numerical examples, see Burt.[19,20])

*Present address: Department of Physics, University of Essex, Wivenhoe Park, Colchester CO4 35Q, United Kingdom.

[1]G. H. Wannier, Phys. Rev. **52**, 191 (1937); J. C. Slater, *ibid.* **76**, 1592 (1949); E. N. Adams, *ibid.* **85**, 41 (1952); J. Chem. Phys. **21**, 2013 (1953).
[2]J. M. Luttinger and W. Kohn, Phys. Rev. **97**, 869 (1955).
[3]W. A. Harrison, Phys. Rev. **123**, 85 (1961).
[4]D. J. BenDaniel and C. B. Duke, Phys. Rev. **152**, 683 (1966).
[5]L. J. Sham and M. Nakayama, Phys. Rev. B **20**, 734 (1979).
[6]G. Bastard, Phys. Rev. B **24**, 5693 (1981); *ibid.* **25**, 7584 (1982); R. I. Taylor and M. G. Burt, Semicond. Sci. Technol. **2**, 485 (1987).
[7]S. R. White and L. J. Sham, Phys. Rev. Lett. **47**, 879 (1981).
[8]T. Ando and S. Mori, Surf. Sci. **113**, 124 (1982).
[9]Q.-G. Zhu and H. Kroemer, Phys. Rev. B **27**, 3519 (1983).
[10]M. Altarelli, Phys. Rev. B **28**, 842 (1983); in *Heterojunctions and Semiconductor Superlattices,* edited by G. Allan, G. Bastard, N. Boccara, M. Lannoo, and M. Voos (Springer, Berlin, 1986), p. 12.
[11]R. A. Morrow and K. R. Brownstein, Phys. Rev. B **30**, 678 (1984); R. A. Morrow, *ibid.* **35**, 8074 (1987); **36**, 4836 (1987).
[12]M. F. H. Schuurmans and G. W. 't Hooft, Phys. Rev. B **31**, 8041 (1985); R. Eppenga, M. F. H. Schuurmans, and S. Colak, *ibid.* **36**, 1554 (1987).
[13]A. Ishibashi, Y. Mori, K. Kaneko, and N. Watanabe, J. Appl. Phys. **59**, 4087 (1986).
[14]D. L. Smith and C. Mailhiot, Phys. Rev. B **33**, 8345 (1986); **33**, 8360 (1986); Rev. Mod. Phys. **62**, 173 (1990); J. Vac. Sci. Technol. B **8**, 793 (1990).
[15]M. B. Burt, Semicond. Sci. Technol. **2**, 460 (1987); **2**, 701(E) (1987).
[16]M. G. Burt, Semicond. Sci. Technol. **3**, 739 (1988).
[17]M. G. Burt, in *Band Structure Engineering in Semiconductor Microstructures,* edited by R. A. Abram and M. Jaros (Plenum, New York, 1989), p. 99.
[18]M. G. Burt, Semicond. Sci. Technol. **3**, 1224 (1988).
[19]M. G. Burt, J. Phys. Condens. Matter **4**, 6651 (1992).
[20]M. G. Burt, Phys. Rev. B **50**, 7518 (1994).
[21]G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures* (Wiley, New York, 1988); G. Bastard, J. A. Brum, and R. Ferreira, in *Solid State Physics,* edited by H. Ehrenreich and D. Turnbull (Academic, New York, 1991), Vol. 44, p. 229.
[22]W. Trzeciakowski, Phys. Rev. B **38**, 4322 (1988); **38**, 12 493 (1988).
[23]G. T. Einevoll and P. C. Hemmer, J. Phys. C **21**, L1193 (1988); G. T. Einevoll, P. C. Hemmer, and J. Thomsen, Phys. Rev. B **42**, 3485 (1990); G. T. Einevoll, *ibid.* **42**, 3497 (1990).
[24]K. Young, Phys. Rev. B **39**, 13 434 (1989).
[25]U. Ekenberg, Phys. Rev. B **40**, 7714 (1989).
[26]T. Ando, S. Wakahara, and H. Akera, Phys. Rev. B **40**, 11 609 (1989); T. Ando and H. Akera, *ibid.* **40**, 11 619 (1989).
[27]G. F. Karavaev and Yu. S. Tikhodeev, Fiz. Tekh. Poluprovodn. **25**, 1237 (1991) [Sov. Phys. Semicond. **25**, 745 (1991)].
[28]C. Aversa and J. E. Sipe, Phys. Rev. B **47**, 6590 (1993).
[29]M. R. Geller and W. Kohn, Phys. Rev. Lett. **70**, 3103 (1993).
[30]A. Elçi, Phys. Rev. B **49**, 7432 (1994); **50**, 8882 (1994).
[31]G. T. Einevoll and L. J. Sham, Phys. Rev. B **49**, 10 533 (1994).
[32]T. L. Li and K. J. Kuhn, Phys. Rev. B **50**, 8589 (1994).
[33]C. Trallero-Giner, F. García-Moliner, V. R. Velasco, and M. Cardona, Phys. Rev. B **45**, 11 944 (1992).
[34]I. A. Kunin, *Elastic Media with Microstructure* (Springer, Berlin, 1982).
[35]R. E. Ziemer, W. H. Tranter, and D. R. Fannin, *Signals and Systems: Continuous and Discrete* (Macmillan, New York, 1983), p. 292.
[36]B. A. Foreman, following paper, Phys. Rev. B **52**, 12 260 (1995).
[37]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1964).
[38]E. O. Kane, in *Physics of III-V Compounds,* edited by R. K. Willardson and A. C. Beer, Semiconductors and Semimetals Vol. 1 (Academic, New York, 1966), p. 75.
[39]G. L. Bir and G. E. Pikus, *Symmetry and Strain-Induced Effects in Semiconductors* (Wiley, New York, 1974).
[40]R. E. Ziemer, W. H. Tranter, and D. R. Fannin, *Signals and Systems: Continuous and Discrete* (Ref. 35), pp. 355–363.
[41]I. M. Gel'fand and G. E. Shilov, *Generalized Functions* (Academic, New York, 1964), Vol. 1.
[42]J. N. Schulman and Y.-C. Chang, Phys. Rev. B **31**, 2056 (1985).
[43]C. Kittel, *Introduction to Solid State Physics,* 6th ed. (Wiley, New York, 1986).
[44]B. Friedman, *Principles and Techniques of Applied Mathematics* (Wiley, New York, 1956), p. 174.
[45]B. A. Foreman, Ph.D. dissertation, Cornell University, 1995.
[46]H. C. Liu, Appl. Phys. Lett. **51**, 1019 (1987).
[47]B. A. Foreman, Phys. Rev. B **48**, 4964 (1993).
[48]G. Edwards and J. C. Inkson, Solid State Commun. **89**, 595 (1994); **91**, 84(E) (1994).
[49]P. von Allmen, Phys. Rev. B **46**, 15 377 (1992).
[50]T. Yamanaka, H. Kamada, Y. Yoshikuni, W. W. Lui, S. Seki, and K. Yokoyama, J. Appl. Phys. **76**, 2347 (1994).