Quantum Mechanics

An Introduction for Device Physicists and Electrical Engineers

David K Ferry

Department of Electrical Engineering,
Arizona State University, Tempe, AZ 85287-5706, USA

Institute of Physics Publishing
Bristol and Philadelphia

Preface

Most treatments of quantum mechanics have begun from the historical basis of the application to nuclear and atomic physics. This generally leaves the important topics of quantum wells, tunnelling, and periodic potentials until late in the course. This puts the person interested in solid-state electronics and solid-state physics at a disadvantage, relative to their counterparts in more traditional fields of physics and chemistry. While there are a few books that have departed from this approach, it was felt that there is a need for one that concentrates primarily upon examples taken from the new realm of artificially structured materials in solid-state electronics. Quite frankly, we have found that students are often just not prepared adequately with experience in those aspects of quantum mechanics necessary to begin to work in small structures (what is now called mesoscopic physics) and nanoelectronics, and that it requires several years to gain the material in these traditional approaches. Students need to receive the material in an order that concentrates on the important aspects of solid-state electronics, and the modern aspects of quantum mechanics that are becoming more and more used in everyday practice in this area. That has been the aim of this text. The topics and the examples used to illustrate the topics have been chosen from recent experimental studies using modern microelectronics, heteroepitaxial growth, and quantum well and superlattice structures, which are important in today's rush to nanoelectronics.

At the same time, the material has been structured around a senior-level course that we offer at Arizona State University. Certainly, some of the material is beyond this (particularly chapter 9), but the book could as easily be suited to a first-year graduate course with this additional material. On the other hand, students taking a senior course will have already been introduced to the ideas of wave mechanics with the Schrödinger equation, quantum wells, and the Krönig–Penney model in a junior-level course in semiconductor materials. This earlier treatment is quite simplified, but provides an introduction to the concepts that are developed further here. The general level of expectation on students using this material is this prior experience plus the linear vector spaces and electromagnetic field theory to which electrical engineers have been exposed.

I would like to express thanks to my students who have gone through the course, and to Professors Joe Spector and David Allee, who have read the manuscript completely and suggested a great many improvements and changes.

David K Ferry
Tempe, AZ, 1992

1

Waves and particles

1.1 INTRODUCTION

Science has developed through a variety of investigations more or less over the time scale of human existence. On this scale, quantum mechanics is a very young field, existing essentially only since the beginning of this century. Even our understanding of classical mechanics has existed for a comparatively long period—roughly having been formalized with Newton's equations published in his *Principia Mathematica*, in April 1686. In fact, we have just celebrated more than 400 years of classical mechanics.

In contrast with this, the ideas of quantum mechanics emerged in about 1913 when Bohr and Sommerfeld developed a model of atomic structure to explain the discrete absorption and emission lines that were seen experimentally. However, much of the physics entailed in this picture of 'quantization' of the energy levels was quite *ad hoc* and could not be justified, although in the end the ideas proved correct—it was only some of the mathematical details that needed changing. However, quantum mechanics, as we currently know it, really entered the physics scene in the period after the First World War. The basic work of Schrödinger and Heisenberg led to different, but equivalent, formulations of the quantum principles that were important in physical systems. Today, there is a consensus (but not a complete agreement) as to the general understanding of the quantum principles. In essence, quantum mechanics is the mathematical description of physical systems with non-commuting operators; for example, the ordering of the operators is very important. The engineer is familiar with such an ordering dependence through the use of matrix algebra, where in general the order of two matrices is important; that is $AB \neq BA$. In quantum mechanics, the ordering of various *operators* is important, and it is these operators that do not commute. There are two additional, and quite important, postulates. These are *complementarity* and the *correspondence principle*.

Complementarity refers to the duality of waves and particles. That is, for both electrons and light waves, there is a duality between a treatment in terms of waves and a treatment in terms of particles. The wave treatment generally is described by a field theory with the corresponding operator effects introduced into the wave amplitudes. The particle is treated in a manner similar to the classical particle dynamics treatment with the appropriate operators properly introduced. In the next two sections, we will investigate two of the operator effects.

On the other hand, the *correspondence principle* relates to the limiting approach to the well known classical mechanics. It will be found that Planck's constant, $h = 2\pi\hbar$, appears in all results that truly reflect quantum mechanical behaviour. As we allow $h \to 0$, the classical results must be obtained. That is, the true quantum effects must vanish as we take this limit. Now, we really don't vary the value of such a fundamental constant, but the correspondence principle asserts that if we were to do so, the classical results would be recovered. What this means is that the quantum effects are modifications of the classical properties. These effects may be small or large, depending upon a number of factors such as time scales, size scales and energy scales. The value of Planck's constant is quite small, 6.6025×10^{-34} J s, but one should not assume that the quantum effects are small. For example, quantization is found to affect the operation of modern metal–oxide–semiconductor (MOS) transistors and to be the fundamental property of devices such as a tunnel diode.

1.2 LIGHT AS PARTICLES—THE PHOTOELECTRIC EFFECT

One of the more interesting examples of the principle of complementarity is that of the photoelectric effect. It was known that when light was shone upon the surface of a metal, or some other conducting medium, electrons could be emitted from the surface provided that the frequency of the incident light was sufficiently high. The curious effect is that the velocity of the emitted electrons depends only upon the wavelength of the incident light, and *not upon the intensity of the radiation*. In fact, the energy of the emitted particles varies inversely with the wavelength of the light waves. On the other hand, the *number* of emitted electrons does depend upon the intensity of the radiation, and not upon its wavelength. Today, of course, we do not consider this surprising at all, but this is after it has been explained in the Nobel-prize-winning work of Einstein. What Einstein concluded was that the explanation of this phenomenon required a treatment of light in terms of its 'corpuscular' nature; that is, we need to treat the light wave as a beam of particles impinging upon the surface of the metal. In fact, it is important to describe the energy of the individual light particles, which we call *photons*, using the relation (Einstein 1905)

$$\mathcal{E} = h\nu = \hbar\omega. \tag{1.1}$$

The photoelectric effect can be understood through consideration of figure 1.1. However, it is essential to understand that we are talking about the flow of 'particles' as directly corresponding to the wave intensity of the light wave.

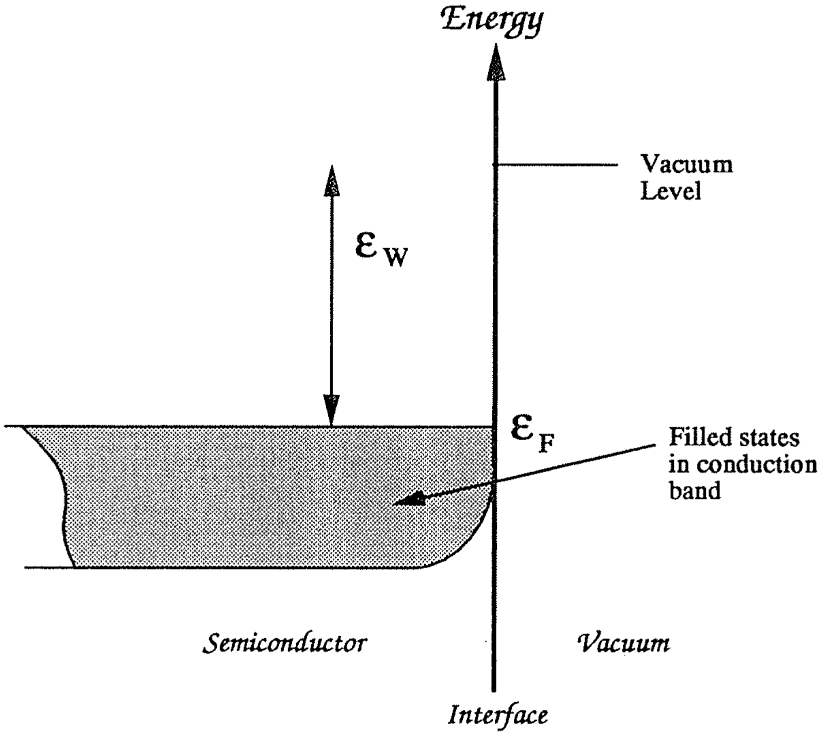


Figure 1.1 The energy bands for the surface of a metal. An incident photon with an energy greater than the work function, $\mathcal{E}_W$, can cause an electron to be raised from the Fermi energy, $\mathcal{E}_F$, to above the vacuum level, whereby it can be photoemitted.

Where the intensity is 'high', there is a high density of photons. Conversely, where the wave amplitude is weak, there is a low density of photons.

A metal is characterized by a work function $\mathcal{E}_W$, which is the energy required to raise an electron from the Fermi energy to the vacuum level, from which it can be emitted from the surface. Thus, in order to observe the photoelectric effect, or photoemission as it is now called, it is necessary to have the energy of the photons greater than the work function, or $\mathcal{E} > \mathcal{E}_W$. The excess energy, that is the energy difference between that of the photon and the work function, becomes the kinetic energy of the emitted particle. Since the frequency of the photon is inversely proportional to the wavelength, the kinetic energy of the emitted particle varies inversely as the wavelength of the light. As the intensity of the light wave in increased, the number of incident photons increases, and therefore the number of emitted electrons increases. However, the momentum of each emitted electron depends upon the properties of a single photon, and therefore is independent of the intensity of the light wave.

A corollary of the acceptance of light as particles is that there is a momentum associated with each of the particles. It is well known in field theory that there is a momentum associated with the (massless) wave, which is given by $p = h\nu/c$, which leads immediately to the relationship

$$p = \frac{h\nu}{c} = \frac{h}{\lambda}. \tag{1.2}$$

Here, we have used the magnitude, rather than the vector description, of the momentum. It then follows that

$$p = \frac{h}{\lambda} = \hbar k \tag{1.3}$$

a relationship that is familiar both to those accustomed to field theory and to those familiar with solid-state theory.

It is finally clear from the interpretation of light waves as particles that there exists a relationship between the 'particle' energy and the frequency of the wave, and a connection between the momentum of the 'particle' and the wavelength of the wave. The two equations (1.1) and (1.3) give these relationships. The form of (1.2) has usually been associated with de Broglie, and the wavelength corresponding to the particle momentum is usually described as the *de Broglie wavelength*. However, it is worth noting that de Broglie (1939) referred to the set of equations (1.1) and (1.3) as the Einstein relations! In fact, de Broglie's great contribution was the recognition that atoms localized in orbits about a nucleus must possess these same wave-like properties. Hence, the electron orbit must be able to incorporate an exact integer number of wavelengths, given by (1.3) in terms of the momentum. This then leads to quantization of the energy levels.

1.3 ELECTRONS AS WAVES

In the previous section, we discussed how in many cases it is clearly more appropriate, and indeed necessary, to treat electromagnetic waves as the flow of particles, which in turn are termed photons. By the same token, there are times when it is clearly advantageous to describe particles, such as electrons, as waves. In the correspondence between these two viewpoints, it is important to note that the varying intensity of the wave reflects the presence of a varying number of particles; the particle density at a point x, at time t, reflects the varying intensity of the wave at this point and time. For this to be the case, it is important that quantum mechanics describe both the wave and particle pictures through the principle of superposition. That is, the amplitude of the composite wave is related to the sum of the amplitudes of the individual waves corresponding to each of the particles present. Note that it is the amplitudes, and not the intensities, that are summed, so there arises the real possibility for *interference* between the waves of individual particles. Thus, for the presence of two (non-interacting) particles at a point x, at time t, we may write the composite wave function as

$$\Psi(x, t) = \Psi_1(x, t) + \Psi_2(x, t). \tag{1.4}$$

This composite wave may be described as a *probability wave*, in that the square of the magnitude describes the probability of finding an electron at a point.

It may be noted from (1.3) that the momentum of the particles goes immediately into the so-called *wave vector* $\boldsymbol{k}$ of the wave. A special form of (1.4) is

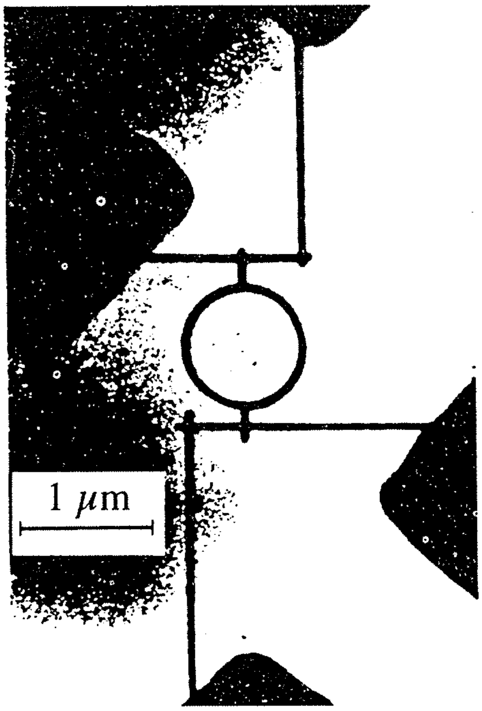


Figure 1.2 Transmission electron micrograph of a large-diameter (820 nm) polycrystalline Au ring. The lines are about 40 nm wide and about 38 nm thick. (After Washburn and Webb (1986), by permission.)

$$\Psi(x,t) = A\mathrm{e}^{\mathrm{i}(k_1 x - \omega t)} + B\mathrm{e}^{\mathrm{i}(k_2 x - \omega t)} \tag{1.5}$$

where it has been assumed that the two components may have different momenta (but we have taken the energies equal). For the moment, the time-independent steady state will be considered, so the time-varying parts of (1.5) will be suppressed as we will talk only about steady-state results of phase interference. It is known, for example, that a time-varying magnetic field that is enclosed by a conducting loop will induce an electric field (and voltage) in the loop through Faraday's law. Can this happen for a time-independent magnetic field? The classical answer is, of course, no, and Maxwell's equations give us this answer. But do they in the quantum case where we can have the interference between the two waves corresponding to two separate electrons?

For the experiment, we consider a loop of wire. Specifically, the loop is made of Au wire deposited on a Si_3N_4 substrate. Such a loop is shown in figure 1.2, where the loop is about 820 nm in diameter, and the Au lines are 40 nm wide (Webb *et al* 1985). The loop is connected to an external circuit through Au leads (also shown), and a magnetic field is threaded through the loop.

To understand the phase interference, we proceed by assuming that the electron waves enter the ring at a point described by $\phi = -\pi$. For the moment, assume that the field induces an electric field in the ring (the time variation will in the end cancel out, and it is not the electric field *per se* that causes the effect, but this approach allows us to describe the effect). Then, for one electron passing through the upper side of the ring, the electron is accelerated by the field, as it moves *with* the field, while on the other side of the ring the electron is decelerated by the field as it moves *against* the field. The field enters through Newton's law, and

$$k = k_0 - \frac{e}{\hbar}\int E\,\mathrm{d}t. \tag{1.6}$$

If we assume that the initial wave vector is the same for both electrons, then the phase difference at the output of the ring is given by taking the difference of the integral over momentum in the top half of the ring (from an angle of π down to 0) and the integral over the bottom half of the ring (from $-\pi$ up to 0):

$$\begin{aligned}\Delta\phi &= -\frac{e}{\hbar}\int \mathrm{d}t\left(\int_{\pi}^{0} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l} + \int_{-\pi}^{0} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l}\right) = -\frac{e}{\hbar}\int \mathrm{d}t\int_{0}^{2\pi} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l} \\ &= -\frac{e}{\hbar}\int \mathrm{d}t\int \nabla\times\boldsymbol{E}\cdot\boldsymbol{n}\,\mathrm{d}A = \frac{e}{\hbar}\int \boldsymbol{B}\cdot\boldsymbol{n}\,\mathrm{d}A = 2\pi\frac{\Phi}{\Phi_0}\end{aligned} \tag{1.7}$$

where $\Phi_0 = h/e$ is the quantum unit of flux, and we have used Maxwell's equations to replace the electric field by the time derivative of the magnetic flux density. Thus, a *static* magnetic field coupled through the loop creates a phase difference between the waves that traverse the two paths. This effect is the Aharonov–Bohm (1959) effect.

In figure 1.3(*a*), the conductance through the ring of figure 1.2 is shown. There is a strong oscillatory behaviour as the magnetic field coupled by the ring is varied. The curve of figure 1.3(*b*) is the Fourier transform (with respect to magnetic field) of the conductance and shows a clear fundamental peak corresponding to a 'frequency' given by the periodicity of Φ_0. There is also a weak second harmonic evident in the Fourier transform, which may be due to weak non-linearities in the ring (arising from variations in thickness, width etc) or to other physical processes (some of which are understood).

The coherence of the electron waves is a clear requirement for the observation of the Aharonov–Bohm effect, and this is why the measurements are done at such low temperatures. It is important that the size of the ring be smaller than some characteristic coherence length, which is termed the inelastic mean free path (where it is assumed that it is inelastic collisions between the electrons that destroy the phase coherence). Nevertheless, the understanding of this phenomenon depends upon the ability to treat the electrons as waves, and, moreover, the phenomenon is only found in a temperature regime where the phase coherence is maintained. At higher temperatures, the interactions between the electrons in the metal ring become so strong that the phase is *randomized*, and any possibility of phase interference effects is lost. Thus the quantum interference is only observable on size and energy scales (set by the coherence length and the temperature, respectively) such that the quantum interference is quite significant. As the temperature is raised, the phase is randomized by the collisions, and normal classical behaviour is recovered. This latter may be

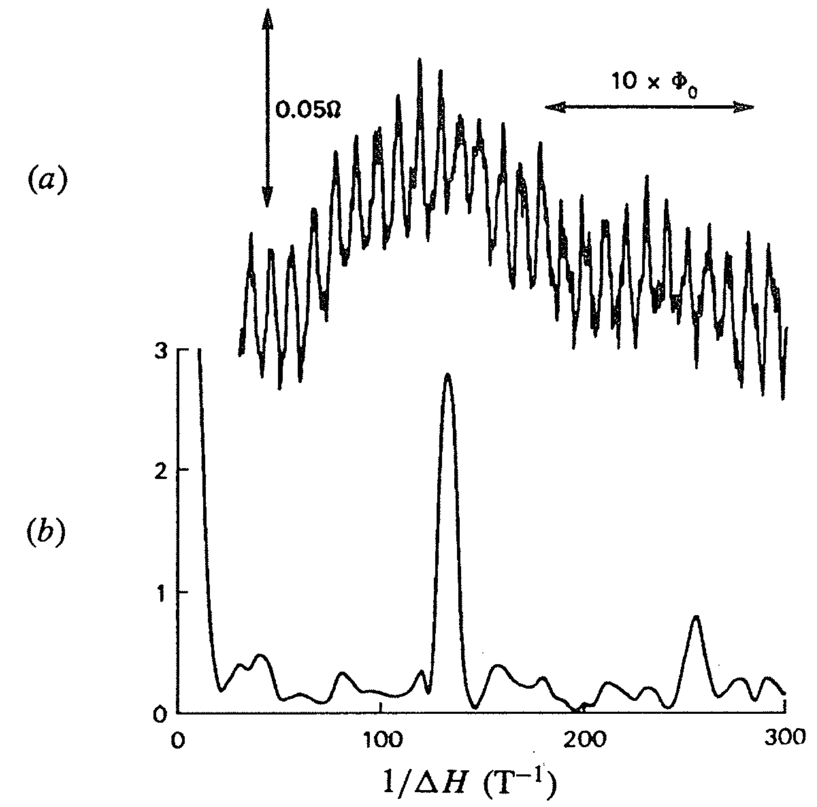


Figure 1.3 Conductance through the ring of figure 1.2. In (*a*), the conductance oscillations are shown at a temperature of 0.04 K. The Fourier transform is shown in (*b*) and gives clearly evidence of the dominant h/e period of the oscillations. (After Washburn and Webb (1986), by permission.)

described by requiring that the two waves used above add in intensity, and not in amplitude as we have done. The addition of intensities 'throws away' the phase variables and precludes the possibility of phase interference between the two paths.

Which is the proper interpretation to use for a general problem: particle or wave? The answer is not an easy one to give. Rather, the proper choice depends largely upon the particular quantum effect being investigated. Thus one chooses the approach that yields the answer with minimum effort. Nevertheless, the great majority of work actually has tended to treat the quantum mechanics via the wave mechanical picture, as embodied in the Schrödinger equation (discussed in the next chapter). One reason for this is the great wealth of mathematical literature dealing with boundary value problems, as the time-independent Schrödinger equation is just a typical wave equation. Most such problems actually lie in the formulation of the proper boundary conditions, and then the imposition of non-commuting variables. Before proceeding to this, however, we diverge to continue the discussion of position and momentum as variables and operators.

1.4 POSITION AND MOMENTUM

For the remainder of this chapter, we want to concentrate on just what properties we can expect from this wave that is supposed to represent the particle (or particles). Do we represent the particle simply by the wave itself? No, because the wave is a complex quantity, while the charge and position of the particle are real quantities. Moreover, the wave is a distributed quantity, while we expect the particle to be relatively localized in space. This suggests that we relate the *probability* of finding the electron at a position x to the square of the magnitude of the wave. That is, we say that

$$|\Psi(x,t)|^2 \tag{1.8}$$

is the probability of finding an electron at point x at time t. Then, it is clear that the wave function must be normalized through

$$\int_{-\infty}^{\infty} |\Psi(x,t)|^2 \, \mathrm{d}x = 1. \tag{1.9}$$

While (1.9) extends over all space, the appropriate volume is that of the system under discussion. This leads to a slightly different normalization for the plane waves utilized in section 1.3 above. Here, we use *box normalization* (the term 'box' refers to the three-dimensional case):

$$\lim_{L\to\infty} \int_{-L/2}^{L/2} |\Psi(x,t)|^2 \, \mathrm{d}x = 1. \tag{1.10}$$

This normalization keeps constant total probability and recognizes that, for a uniform probability, the amplitude must go to zero as the volume increases without limit.

1.4.1 Expectation of the position

With the normalizations that we have now introduced, it is clear that we are equating the square of the magnitude of the wave function with a probability density function. This allows us to compute immediately the expectation value, or average value, of the position of the particle with the normal definitions introduced in probability theory. That is, the average value of the position is given by

$$\langle x \rangle = \int_{-\infty}^{\infty} x |\Psi(x,t)|^2 \, \mathrm{d}x = \int_{-\infty}^{\infty} \Psi^*(x,t) x \, \Psi(x,t) \, \mathrm{d}x. \tag{1.11}$$

In the last form, we have split the wave function product into its two components and placed the position *operator* between the complex conjugate of the wave function and the wave function itself. This is the standard notation, and designates that we are using the concept of an inner product of two functions to describe the average. If we use (1.9) to define the inner product of the wave function and its complex conjugate, then this may be described in the short-hand notation

$$(\Psi, \Psi) = \int_{-\infty}^{\infty} \Psi^*(x,t)\Psi(x,t)\,\mathrm{d}x = 1 \tag{1.12}$$

and

$$\langle x \rangle = (\Psi, x\Psi). \tag{1.13}$$

We say at this point that we have described the wave function corresponding to the particle in the *position representation*. That is, the wave function is a function of the position and the time, and the square of the magnitude of this function describes the probability density function for the position. The position operator itself, x, operates on the wave function to provide a new function, so the inner product of this new function with the original function gives the average value of the position. Now, if the position variable x is to be interpreted as an operator, and the wave function in the position representation is the natural function to use to describe the particle, then it may be said that the wave function $\Psi(x,t)$ has an *eigenvalue* corresponding to the operator x. This means that we can write the operation of x on $\Psi(x,t)$ as

$$x\Psi(x,t) = \underline{x}\Psi(x,t) \tag{1.14}$$

where $\underline{x}$ is the eigenvalue of x operating on $\Psi(x,t)$. It is clear that the use of (1.14) in (1.13) means that the eigenvalue $\underline{x} = \langle x \rangle$.

We will see later that one may decompose the overall wave function into an expansion over a complete orthonormal set of basis functions, just like a Fourier series expansion in sines and cosines. Each member of the set has a well defined eigenvalue corresponding to an operator if the set is the proper basis set with which to describe the effect of that operator. Thus, the present use of the position representation means that our functions are the proper functions with which to describe the action of the position operator, which does no more than determine the expectation value of the position of our particle.

Consider the wave function shown in figure 1.4. Here, the real part of the wave function is plotted, as the wave function itself is in general a complex quantity. However, it is clear that the function is peaked about some point x_{peak}. While it is likely that the expectation value of the position is very near this point, this cannot be discerned exactly without actually computing the action of the position operator on this function and computing the expectation value, or inner product, directly. This circumstance arises from the fact that we are now dealing with probability functions, and the expectation value is simply the most likely position in which to find the particle. On the other hand, another quantity is evident in figure 1.4, and this is the width of the wave function, which relates to the standard deviation of the wave function. Thus, we can define

$$(\Delta x)^2 = (\Psi, (x - \langle x \rangle)^2\Psi). \tag{1.15}$$

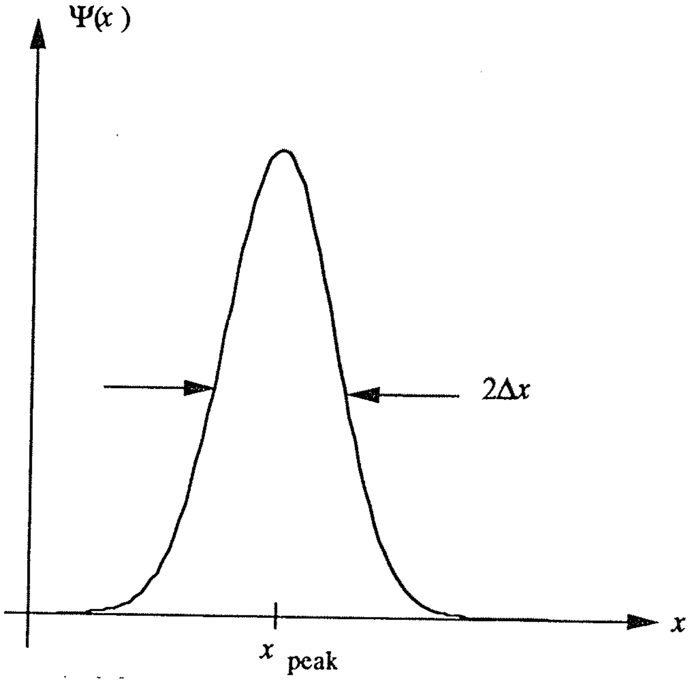


Figure 1.4 The positional variation of a typical wave function.

The quantity Δx relates to the uncertainty in finding the particle at the position $\langle x \rangle$. It is clear that if we want to use a wave packet that describes the position of the particle *exactly*, then Δx must be made to go to zero. Such a function is the Dirac delta function familiar from circuit theory (the impulse function). Here, though, we use a delta function in position rather than in time; for example, we describe the wave function through

$$\Psi(x, 0) = \delta(x - x_{\text{peak}}). \tag{1.16}$$

The time variable has been set to zero here for convenience, but it is easy to extend (1.16) to the time-varying case. Clearly, equation (1.16) describes the wave function under the condition that the position of the particle is known absolutely! We will examine in the following paragraphs some of the limitations this places upon our knowledge of the dynamics of the particle.

1.4.2 Momentum

The wave function shown in figure 1.4 contains variations in space, and is not a uniform quantity. In fact, if it is to describe a localized particle, it must vary quite rapidly in space. It is possible to Fourier transform this wave function in order to get a representation that describes the spatial frequencies that are involved. Then, the wave function in this figure can be written in terms of the spatial frequencies as an inverse transform:

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}kx} \, \mathrm{d}k. \tag{1.17}$$

The quantity $\phi(k)$ represents the Fourier transform of the wave function itself. Here, k is the spatial frequency. However, this k is precisely the same k as appears in (1.3). That is, the spatial frequency is described by the *wave vector* itself, which in turn is related to the momentum through (1.3). For this reason, $\phi(k)$ is called the *momentum wave function*. A description of the particle in momentum space is made using the Fourier-transformed wave functions, or momentum wave functions. Consequently, the average value of the momentum for our particle, the expectation value of the operator p, may be evaluated using these functions. In essence, we are saying that the proper basis set of functions with which to evaluate the momentum is that of the momentum wave functions. Then, it follows that

$$\langle p \rangle = \hbar(\phi, k\phi) = \int_{-\infty}^{\infty} \phi^* p \phi \, \mathrm{d}p. \tag{1.18}$$

Suppose, however, that we are using the position representation wave functions. How then are we to interpret the expectation value of the momentum? The wave functions in this representation are functions only of x and t. To evaluate the expectation value of the momentum operator, it is necessary to develop the operator corresponding to the momentum in the position representation. To do this, we use (1.18) and introduce the Fourier transforms corresponding to the functions ϕ. Then, we may write (1.18) as

$$\begin{aligned}
\langle p \rangle &= \frac{\hbar}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k \int_{-\infty}^{\infty} \mathrm{d}x' \, \Psi^*(x') \mathrm{e}^{\mathrm{i}kx'} k \int_{-\infty}^{\infty} \mathrm{d}x \, \Psi(x) \mathrm{e}^{-\mathrm{i}kx} \\
&= \frac{\hbar}{2\mathrm{i}\pi} \int_{-\infty}^{\infty} \mathrm{d}k \int_{-\infty}^{\infty} \mathrm{d}x' \, \Psi^*(x') \mathrm{e}^{\mathrm{i}kx'} \int_{-\infty}^{\infty} \mathrm{d}x \, \mathrm{e}^{-\mathrm{i}kx} \frac{\partial}{\partial x} \Psi(x) \\
&= -\mathrm{i}\hbar \int_{-\infty}^{\infty} \mathrm{d}x' \int_{-\infty}^{\infty} \mathrm{d}x \, \Psi^*(x') \, \delta(x - x') \frac{\partial}{\partial x} \Psi(x) \\
&= -\mathrm{i}\hbar \int_{-\infty}^{\infty} \mathrm{d}x \, \Psi^*(x) \frac{\partial}{\partial x} \Psi(x).
\end{aligned} \tag{1.19}$$

In arriving at the final form of (1.19), an integration by parts has been done from the first line to the second (the evaluation at the limits is assumed to vanish), after replacing k by the partial derivative. The third line is achieved by recognizing the δ-function:

$$\delta(x - x') = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k \, \mathrm{e}^{\mathrm{i}k(x-x')}. \tag{1.20}$$

Thus, in the position representation, *the momentum operator* is given by the functional operator

$$p = -\mathrm{i}\hbar \frac{\partial}{\partial x}. \tag{1.21}$$

1.4.3 Non-commuting operators

The description of the momentum operator in the position representation is that of a differential operator. This means that the operators corresponding to the position and to the momentum will not commute, by which we mean that

$$[x, p] = xp - px \neq 0. \tag{1.22}$$

The left-hand side of (1.22) defines a quantity that is called the *commutator bracket*. However, by itself it only has implied meaning. The terms contained within the brackets are operators and must actually operate on some wave function. Thus, the role of the commutator can be explained by considering the inner product, or expectation value. This gives

$$-(\Psi, [x, p]\Psi) = -\mathrm{i}\hbar\left\{\left(\Psi, x\frac{\partial}{\partial x}\Psi\right) - \left(\Psi, \frac{\partial}{\partial x}x\Psi\right)\right\} = \mathrm{i}\hbar. \tag{1.23}$$

If variables, or operators, do not commute, there is an implication that these quantities cannot be measured simultaneously. Here again, there is another and deeper meaning. In the previous section, we noted that the operation of the position operator x on the wave function in the position representation produced an eigenvalue $\underline{x}$, which is actually the expectation value of the position. The momentum operator does not produce this simple result with the wave function of the position representation. Rather, the differential operator produces a more complex result. For example, if the differential operator were to produce a simple eigenvalue, then the wave function would be constrained to be of the form $\exp(\mathrm{i}px/\hbar)$ (which can be shown by assuming a simple eigenvalue form as in (1.14) with the differential operator and solving the resulting equation). This form is not integrable (it does not fit our requirements on normalization), and thus the same wave function cannot simultaneously yield eigenvalues for both position and momentum. Since the eigenvalue relates to the expectation value, which corresponds to the most likely result of an experiment, these two quantities cannot be simultaneously measured.

There is a further level of information that can be obtained from the Fourier transform pair of position and momentum wave functions. If the position is known, for example if we choose the delta function of (1.16), then the Fourier transform has unit amplitude everywhere; that is, the momentum has equal probability of taking on any value. Another way of looking at this is to say that since the position of the particle is completely determined, it is impossible to say anything about the momentum, as any value of the momentum is equally likely. Similarly, if a delta function is used to describe the momentum wave function, which implies that we know the value of the momentum exactly, then the position wave function has equal amplitude everywhere. This means that if the momentum is known, then it is impossible to say anything about the position, as all values of the latter are equally likely. As a consequence, if we want to describe both of these properties of the particle, the position wave function and its Fourier transform must be selected carefully to allow this to occur. Then there will be an uncertainty Δx in position, as indicated in figure 1.4, and there will e a corresponding uncertainty Δp in momentum.

To investigate the relationship between the two uncertainties, in position and momentum, let us choose a Gaussian wave function to describe the wave function in the position representation. Therefore, we take

$$\Psi(x) = \frac{1}{(2\pi)^{1/4}\sigma^{1/2}} \exp\left[-\frac{x^2}{4\sigma^2}\right]. \tag{1.24}$$

Here, the wave packet has been centred at $x_{\text{peak}} = 0$, and

$$\langle x \rangle = \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] x \, \mathrm{d}x = 0 \tag{1.25}$$

as expected. Similarly, the uncertainty in the position is found from (1.15) as

$$\begin{aligned}(\Delta x)^2 &= \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] x^2 \, \mathrm{d}x \\ &= \frac{\sigma^2}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] \mathrm{d}x = \sigma^2\end{aligned} \tag{1.26}$$

and $\Delta x = \sigma$.

The appropriate momentum wave function can now be found by Fourier transforming this position wave function. This gives

$$\begin{aligned}\phi(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Psi(x) \mathrm{e}^{-\mathrm{i}kx} \, \mathrm{d}x \\ &= \frac{1}{\sigma^{1/2}(2\pi)^{3/4}} \mathrm{e}^{-\sigma^2 k^2} \int_{-\infty}^{\infty} \exp\left[-\frac{(x - 2\mathrm{i}\sigma^2 k)^2}{4\sigma^2}\right] \mathrm{d}x \\ &= \left(\frac{2}{\pi}\right)^{1/4} \sqrt{\sigma}\, \mathrm{e}^{-\sigma^2 k^2}.\end{aligned} \tag{1.27}$$

We note that the momentum wave function is also centred about zero momentum. Then the uncertainty in the momentum can be found as

$$(\Delta p)^2 = \hbar^2 \sigma \sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-2\sigma^2 k^2} k^2 \, \mathrm{d}k = \frac{\hbar^2}{4\sigma^2}. \tag{1.28}$$

Hence, the uncertainty in the momentum is $\hbar/2\sigma$. We now see that the non-commuting operators x and p can be described by an uncertainty $\Delta x \Delta p = \hbar/2$. It turns out that our description in terms of the static Gaussian wave function is a *minimal-uncertainty* description, in that the product of the two uncertainties is a minimum.

The uncertainty principle describes the connection between the uncertainties in determination of the expectation values for two non-commuting operators. If we have two operators A and B, which do not commute, then the uncertainty relation states that

$$\Delta A \, \Delta B \geqslant \tfrac{1}{2}|\langle [A, B] \rangle| \tag{1.29}$$

where the angular brackets denote the expectation value, as above. It is easily confirmed that the position and momentum operators satisfy this relation. It is important to note that the basic uncertainty relation is only really valid for non-commuting operators. It has often been asserted for variables like energy (frequency) and time, but in the non-relativistic quantum mechanics that we are investigating here, time is not a dynamic variable and has no corresponding operator. Thus, if there is any uncertainty for these latter two variables, it arises from the problems of making measurements of the energy at different times—and hence is a measurement uncertainty and not one expected from the uncertainty relation (1.29).

1.4.4 Returning to temporal behaviour

While we have assumed that the momentum wave function is centred at zero momentum, this is not the general case. Suppose, we now assume that the momentum wave function is centred at a displaced value of k, given by k_0. Then, the entire position representation wave function moves with this average momentum, and shows an average velocity $v_0 = \hbar k_0/m$. We can expect that the peak of the position wave function, x_{peak}, moves, but does it move with this velocity? The position wave function is made up of a sum of a great many Fourier components, each of which arises from a different momentum. Does this affect the uncertainty in position that characterizes the half-width of the position wave function? The answer to both of these questions is yes, but we will try to demonstrate that these are the correct answers in this section.

Our approach is based upon the definition of the Fourier inverse transform (1.17). This latter equation expresses the position wave function $\Psi(x)$ as a summation of individual Fourier components, each of whose amplitudes is given by the value of $\phi(k)$ at that particular k. From the earlier work, we can extend each of the Fourier terms into a plane wave corresponding to that value of k, by introducing the frequency term via

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}(kx-\omega t)} \, \mathrm{d}k. \tag{1.30}$$

While the frequency term has not been shown with a variation with k, it must be recalled that each of the Fourier components may actually possess a slightly different frequency. If the main frequency corresponds to the peak of the momentum wave function, then the frequency can be expanded as

$$\omega(k) = \omega(k_0) + (k - k_0)\frac{\partial \omega}{\partial k}\bigg|_{k=k_0} + \dots . \tag{1.31}$$

The interpretation of the position wave function is now that it is composed of a group of closely related waves, all propagating in the same direction (we assume that $\phi(k) = 0$ for $k < 0$, but this is merely for convenience and is not critical to the overall discussion). Thus, $\Psi(x, t)$ is now defined as a *wave packet*. Equation (1.31) defines the *dispersion* across this wave packet, as it gives the gradual change in frequency for different components of the wave packet.

To understand how the dispersion affects the propagation of the wave functions, we insert (1.31) into (1.30), and define the difference variable $u = k - k_0$. Then, (1.30) becomes

$$\Psi(x, t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}(k_0 x - \omega_0 t)} \int_{-\infty}^{\infty} \phi(u + k_0) \mathrm{e}^{\mathrm{i}(ux - \omega' ut)} \, \mathrm{d}u \tag{1.32}$$

where ω_0 is the leading term in (1.31) and ω' is the partial derivative in the second term of (1.31). The higher-order terms of (1.31) are neglected, as the first two terms are the most significant. If u is factored out of the argument of the exponential within the integral, it is seen that the position variable varies as $x - \omega' t$. This is our guide as to how to proceed. We will reintroduce k_0 within the exponential, but multiplied by this factor, so that

$$\begin{aligned}\Psi(x, t) &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-ik_0(x-\omega' t)} \mathrm{e}^{\mathrm{i}(k_0 x - \omega_0 t)} \int_{-\infty}^{\infty} \phi(u + k_0) \mathrm{e}^{ik_0(x-\omega' t)} \mathrm{e}^{\mathrm{i}u(x-\omega' t)} \, \mathrm{d}k \\ &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\mathrm{i}(\omega_0 - \omega' k_0)t} \int_{-\infty}^{\infty} \phi(u + k_0) \mathrm{e}^{\mathrm{i}(u+k_0)(x-\omega' t)} \, \mathrm{d}k \\ &= \mathrm{e}^{-\mathrm{i}(\omega_0 - \omega' k_0)t} \Psi(x - \omega' t, t).\end{aligned} \tag{1.33}$$

The leading exponential provides a phase shift in the position wave function. This phase shift has no effect on the square of the magnitude, which represents the expectation value calculations. On the other hand, the entire wave function moves with a velocity given by ω'. This is not surprising. The quantity ω' is the partial derivative of the frequency with respect to the momentum wave vector, and hence describes the group velocity of the wave packet. Thus, the average velocity of the wave packet in position space is given by the group velocity

$$v_g = \omega' = \frac{\partial \omega}{\partial k}\bigg|_{k=k_0}. \tag{1.34}$$

This answers the first question: the peak of the position wave function remains the peak and moves with an average velocity defined as the group velocity of the wave packet. Note that this group velocity is defined by the frequency variation with respect to the wave vector. Is this related to the average momentum given by k_0? The answer again is affirmative, as we cannot let k_0 take on any arbitrary value. Rather, the peak in the momentum distribution must relate to the average motion of the wave packet in position space. Thus, we must impose a value on k_0 so that it satisfies the condition of actually being the average momentum of the wave packet:

$$v_{\mathrm{g}} = \frac{\hbar k_0}{m} = \frac{\partial \omega}{\partial k}. \tag{1.35}$$

If we integrate the last two terms of (1.35) with respect to the wave vector, we recover the other condition that ensures that our wave packet is actually describing the dynamic motion of the particles:

$$\mathcal{E} = \hbar\omega = \frac{\hbar^2 k^2}{2m} = \frac{p^2}{2m}. \tag{1.36}$$

It is clear that it is the group velocity of the wave packet that describes the average momentum of the momentum wave function and also relates the velocity (and momentum) to the energy of the particle.

Let us now turn to the question of what the wave packet looks like with the time variation included. We rewrite (1.30) to take account of the centred wave packet for the momentum representation to obtain

$$\Psi(x,t) = \sqrt{\frac{\sigma}{2\pi}}\left(\frac{2}{p}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0 x}\int_{-\infty}^{\infty} \mathrm{e}^{-\sigma^2 u^2 + \mathrm{i}ux - \mathrm{i}\omega t}\,\mathrm{d}u. \tag{1.37}$$

To proceed, we want to insert the above relationship between the frequency (energy) and average velocity:

$$\omega = \frac{\hbar k^2}{2m} = \frac{\hbar}{2m}(u + k_0)^2 = \frac{\hbar u^2}{2m} + u v_{\mathrm{g}} + \frac{\hbar k_0^2}{2m}. \tag{1.38}$$

If (1.38) is inserted into (1.37), we recognize a new form for the 'static' *effective* momentum wave function:

$$\phi(k) = \sqrt{\sigma}\left(\frac{2}{p}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0(x - v_{\mathrm{g}}t/2)} \exp\left[-u^2\left(\sigma^2 + \mathrm{i}\frac{\hbar t}{2m}\right)\right] \tag{1.39}$$

which still leads to $\langle p\rangle = 0$, and $\Delta p = \hbar/2\sigma$. We can then evaluate the position representation wave function by continuing the evaluation of (1.37) using the short-hand notation

$$\sigma' = \sqrt{\sigma^2 + \mathrm{i}\frac{\hbar t}{2m}} \tag{1.40a}$$

and

$$x' = x - v_{\mathrm{g}}t. \tag{1.40b}$$

This gives

$$\begin{aligned}\Psi(x',t) &= \sqrt{\frac{\sigma}{2\pi}}\left(\frac{2}{\pi}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0(x-v_{\mathrm{g}}t/2)} \int_{-\infty}^{\infty} \mathrm{e}^{-\sigma'^2u^2+\mathrm{i}ux'}\,\mathrm{d}u \\ &= \frac{\sqrt{\sigma}}{(2\pi)^{1/4}\sigma'}\,\mathrm{e}^{\mathrm{i}k_0(x-v_{\mathrm{g}}t/2)} \exp\left[-\left(\frac{x'}{2\sigma'}\right)^2\right].\end{aligned} \tag{1.41}$$

This has the exact form of the previous wave function in the position representation with one important exception. The exception is that the time variation has made this result unnormalized. If we compute the inner product now, recalling that the terms in σ' are complex, the result is

$$(\Psi,\Psi) = \frac{\sigma}{|\sigma'|} = \frac{1}{\sqrt{1+\hbar^2t^2/(4m^2\sigma^4)}} \equiv \frac{1}{S}. \tag{1.42}$$

With this normalization, it is now easy to show that the expectation value of the position is that found above:

$$\langle x\rangle = \frac{(\Psi, x\Psi)}{(\Psi,\Psi)} = v_{\mathrm{g}}t. \tag{1.43}$$

Similarly, the standard deviation in position is found to be

$$\langle(\Delta x)^2\rangle = \sigma^2S^2 = \sigma^2\left[1+\frac{\hbar^2t^2}{4m^2\sigma^4}\right]. \tag{1.44}$$

This means that the uncertainty in the two non-commuting operators x and p increases with time according to

$$\Delta x\Delta p = \frac{\hbar}{2}\sqrt{1+\frac{\hbar^2t^2}{4m^2\sigma^4}}. \tag{1.45}$$

The wave packet actually gets wider as it propagates with time, so the time variation is a shift of the centroid plus this broadening effect. The broadening of a Gaussian wave packet is familiar in the process of diffusion, and we recognize that the position wave packet actually undergoes a diffusive broadening as it propagates. This diffusive effect accounts for the increase in the uncertainty. The minimum uncertainty arises only at the initial time when the packet was formed. At later times, the various momentum components cause the wave packet position to become less certain since different spatial variations propagate at different effective frequencies. Thus, for any times after the initial one, it is not possible for us to know as much about the wave packet and there is more uncertainty in the actual position of the particle that is represented by the wave packet.

1.5 SUMMARY

Quantum mechanics furnishes a methodology for treating the wave–particle duality. The main importance of this treatment is for structures and times, both usually small, for which the *interference* of the waves can become important. The effect can be either the interference between two wave packets, or the interference of a wave packet with itself, such as in boundary value problems. In quantum mechanics, the boundary value problems deal with the equation that we will develop in the next chapter for the wave packet, the Schrödinger equation.

The result of dealing with the wave nature of particles is that dynamical variables have become operators which in turn operate upon the wave functions. As operators, these variables often no longer commute, and there is a basic uncertainty relation between non-commuting operators. The non-commuting nature arises from it being no longer possible to generate a wave function that yields eigenvalues for *both* of the operators, representing the fact that they cannot be simultaneously measured. It is this that introduces the uncertainty relationship.

Even if we generate a minimum-uncertainty wave packet in real space, it is correlated to a momentum space representation, which is the Fourier transform of the spatial variation. The time variation of this wave packet generates a diffusive broadening of the wave packet, which increases the uncertainty in the two operator relationships.

We can draw another set of conclusions from this behaviour that will be important for the differential equation that can be used to find the actual wave functions in different situations. The entire time variation has been found to derive from a single initial condition, which implies that the differential equation must be only first order in the time derivatives. Second, the motion has diffusive components, which suggests that the differential equation should bear a strong resemblance to a diffusion equation (which itself is only first order in the time derivative). These points will be expanded upon in the next chapter.

REFERENCES

Aharonov Y and Bohm D 1959 *Phys. Rev.* **115** 485

de Broglie L 1939 *Matter and Light, the New Physics* (New York: Dover) p 267 (this is a reprint of the original translation by W H Johnston of the 1937 original *Matière et Lumière)*

Einstein A 1905 *Ann. Phys., Lpz.* **17** 132

Washburn S and Webb R A 1986 *Adv. Phys.* **35** 375–422

Webb R A, Washburn S, Umbach C P and Laibowitz R B 1985 *Phys. Rev. Lett.* **54** 2696–99

PROBLEMS

1. Calculate the energy density for the plane electromagnetic wave described by the complex field strength

$$E_\mathrm{c} = E_0 \mathrm{e}^{\mathrm{i}(\omega t - \mathrm{i}kx)}$$

and show that its average over a temporal period T is $\omega = (\varepsilon/2)|E_\mathrm{c}|^2$.

2. What are the de Broglie frequencies and wavelengths of an electron and a proton accelerated to 100 eV? What are the corresponding group and phase velocities?

3. Show that the position operator x is represented by the differential operator

$$\mathrm{i}\hbar\frac{\partial}{\partial p}$$

in momentum space, when dealing with momentum wave functions. Demonstrate that (1.22) is still satisfied when momentum wave functions are used.

4. An electron represented by a Gaussian wave packet, with average energy 100 eV, is initially prepared with $\Delta p = 0.1\langle p\rangle$ and $\Delta x = \hbar/[2(\Delta p)]$. How much time elapses before the wave packet has spread to twice the original spatial extent?

5. Express the expectation value of the kinetic energy of a Gaussian wave packet in terms of the expectation value and the uncertainty of the momentum wave function.

6. A particle is represented by a wave packet propagating in a dispersive medium, described by

$$\omega = \frac{A}{\hbar}\left\{\sqrt{1 + \frac{\hbar^2 k^2}{mA}} - 1\right\}.$$

What is the group velocity as a function of k?

2

The Schrödinger equation

In the first chapter, it was explained that the introductory basics of quantum mechanics arise from the changes from classical mechanics that are brought to an observable level by the smallness of some parameter, such as the size scale. The most important effect is the appearance of operators for dynamical variables, and the non-commuting nature of these operators. We also found a wave function, either in the position or momentum representation, whose squared magnitude is related to the probability of finding the equivalent particle. The properties of the wave could be expressed as basically arising from a *linear* differential equation of a diffusive nature. In particular, because any subsequent form for the wave function evolved from a single initial state, the equation can only be of first order in the time derivative (and, hence, diffusive in nature).

In this chapter, we want now to specify such an equation—the Schrödinger equation, from which one version of quantum mechanics—wave mechanics—has evolved. In a later chapter, we shall turn to a second formulation of quantum mechanics based upon time evolution of the operators rather than the wave function, but here we want to gain insight into the quantization process, and the effects it causes in normal systems. In the following section, we will give a justification for the wave equation, but no formal derivation is really possible (as in the case of Maxwell's equations); rather, the equation is found to explain experimental results in a correct fashion, and its validity lies in that fact. In subsequent sections, we will then apply the Schrödinger equation to a variety of problems to gain the desired insight.

2.1 WAVES AND THE DIFFERENTIAL EQUATION

At this point, we want to begin to formulate an equation that will provide us with a methodology for determining the wave function in many different situations, but always in the position representation. We impose two requirements on the wave equation: (i) in the absence of any force, the wave packet must move in a free-particle manner, and (ii) when a force is present, the solution must reproduce Newton's law $F = ma$. As mentioned above, we cannot 'derive' this equation, because the equation itself is the basic postulate of wave mechanics, as formulated by Schrödinger (1926). We begin with (1.30) in the form

$$\Psi(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}(kx-\omega t)}\, \mathrm{d}k. \tag{2.1}$$

Because the wave function must evolve from a single initial condition, it must also be only first order in the time derivative. Thus, we take the partial derivative of (2.1) with respect to time, to yield

$$\frac{\partial \Psi}{\partial t} = -\frac{\mathrm{i}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \omega \mathrm{e}^{\mathrm{i}(kx-\omega t)}\, \mathrm{d}k \tag{2.2}$$

which can be rewritten as

$$\mathrm{i}\hbar \frac{\partial \Psi}{\partial t} = \frac{1}{\sqrt{2p}} \int_{-\infty}^{\infty} \phi(k) \mathcal{E} \mathrm{e}^{\mathrm{i}(kx-\omega t)}\, \mathrm{d}k. \tag{2.3}$$

In essence, the energy is the eigenvalue of the time derivative operator, although this is not a true operator, as time is not a dynamic variable. Thus, it may be thought that the energy represents a set of other operators that do represent dynamic variables. It is common to express the energy as a sum of kinetic and potential energy terms; for example

$$\mathcal{E} = T + V = \frac{p^2}{2m} + V(x,t). \tag{2.4}$$

The momentum does operate on the momentum representation functions, but by using our position space operator form (1.21), the energy term can be pulled out of the integral in (2.3), and we find

$$\mathrm{i}\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + V(x,t)\Psi(x,t). \tag{2.5}$$

This is the Schrödinger equation. We have written it with only one spatial dimension, that of the x-direction. However, the spatial second derivative is properly the Laplacian operator in three dimensions, and the results can readily be obtained for that case. For most of the work in this chapter, however, we will continue to use only the single spatial dimension.

Before proceeding, it is worthwhile to detour and consider to some extent how the classical limit is achieved from the Schrödinger equation. For this, let us define the wave function in terms of an amplitude and a phase, according to $\Psi(x,t) = a\mathrm{e}^{\mathrm{i}S/\hbar}$. The quantity S is known as the *action* in classical mechanics (but familiarity with this will not be required). Let us put this form for the wave function into (2.5), which gives (the exponential factor is omitted as it cancels equally from all terms)

$$-a\frac{\partial S}{\partial t}+i\hbar\frac{\partial a}{\partial t}=\frac{a}{2m}\left(\frac{\partial S}{\partial x}\right)^2-\frac{i\hbar a}{2m}\frac{\partial^2 S}{\partial x^2}-\frac{i\hbar}{m}\frac{\partial S}{\partial x}\frac{\partial a}{\partial x}-\frac{\hbar^2}{2m}\frac{\partial^2 a}{\partial x^2}+Va. \quad (2.6)$$

For this equation to be valid, it is necessary that the real parts and the imaginary parts balance separately, which leads to

$$\frac{\partial S}{\partial t}+\frac{1}{2m}\left(\frac{\partial S}{\partial x}\right)^2+V-\frac{\hbar^2}{2am}\frac{\partial^2 a}{\partial x^2}=0 \quad (2.7)$$

and

$$\frac{\partial a}{\partial t}+\frac{a}{2m}\frac{\partial^2 S}{\partial x^2}+\frac{1}{m}\frac{\partial S}{\partial x}\frac{\partial a}{\partial x}=0. \quad (2.8)$$

In (2.7), there is only one term that includes Planck's constant, and this term vanishes in the classical limit as $\hbar \to 0$. It is clear that the action relates to the phase of the wave function, and consideration of the wave function as a single-particle plane wave relates the gradient of the action to the momentum and the time derivative to the energy. Indeed, insertion of the wave function of (2.1) leads immediately to (2.4), which expresses the total energy. Obviously, here the variation that is quantum mechanical provides a correction to the energy, which comes in as the square of Planck's constant. This extra term, the last term on the left of (2.7), has been discussed by several authors, but today is usually referred to as the Bohm potential. Its interpretation is still under discussion, but this term clearly gives an additional effect in regions where the wave function amplitude varies rapidly with position. One view is that this term plays the role of a quantum pressure, but other views have been expressed. The second equation, (2.8), can be rearranged by multiplying by a, for which (in vector notation for simplicity of recognition)

$$\frac{\partial a^2}{\partial t}+\nabla\cdot\left(\frac{a^2}{m}\nabla S\right)=0. \quad (2.9)$$

The factor a^2 is obviously related to $|\Psi|^2$, the square of the magnitude of the wave function. If the gradient of the action is the momentum, as stated, then the second term is the divergence of the probability current, and the factor in the parentheses is the product of the probability function and its velocity. We explore this further in the next section.

2.2 DENSITY AND CURRENT

The Schrödinger equation is a complex diffusion equation. The wave function Ψ is a complex quantity. The potential energy $V(x,t)$, however, is usually a real quantity. Moreover, we discerned in chapter 1 that the probabilities were real quantities, as they relate to the chance of finding the particle at a particular position. Thus, the probability density is just

$$P(x,t) = \Psi^*(x,t)\Psi(x,t) = |\Psi(x,t)|^2. \tag{2.10}$$

This, of course, leads to the normalization of (1.9), which just expresses the fact that the sum of the probabilities must be unity. If (2.10) were multiplied by the electronic charge e, it would represent the charge density carried by the particle (described by the wave function).

One check of the extension of the Schrödinger equation to the classical limit lies in the continuity equation. That is, if we are to relate (2.10) to the local charge density, then there must be a corresponding current density $\boldsymbol{J}$, such that $(\rho = -eP)$

$$e\frac{\partial P}{\partial t} = \nabla \cdot \boldsymbol{J} \tag{2.11}$$

although we use only the x-component here. Now, the complex conjugate of (2.5) is just

$$-\mathrm{i}\hbar\frac{\partial \Psi^*}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2 \Psi^*}{\partial x^2} + V(x,t)\Psi^*(x,t). \tag{2.12}$$

We now use (2.10) in (2.11), with (2.5) and (2.12) inserted for the partial derivatives with respect to time, as (we neglect the charge, and will find the probability current)

$$\mathrm{i}\hbar\frac{\partial P}{\partial t} = -\frac{\hbar^2}{2m}[\Psi^* \nabla^2\Psi - (\nabla^2\Psi^*)\Psi] \tag{2.13}$$

where the terms involving the potential energy have cancelled. The terms in the brackets can be rewritten as the divergence of a probability current, if the latter is defined as

$$J_\Psi = \frac{\hbar}{2m\mathrm{i}}[\Psi^*(\nabla\Psi) - (\nabla\Psi^*)\Psi]. \tag{2.14}$$

If the wave function is to be a representation of a single electron, then this ‘current’ must be related to the velocity of that particle. On the other hand, if the wave function represents a large ensemble of particles, then the actual current (obtained by multiplying by e) represents some average velocity, with an average taken over that ensemble.

The probability current should be related to the momentum of the wave function, as discussed earlier. The gradient operator in (2.14) is, of course, related to the momentum operator, and the factors of the mass and Planck’s constant connect this to the velocity. In fact, we can rewrite (2.14) as

$$J_\Psi = \frac{1}{2m}(p + p^*)|\Psi|^2. \tag{2.15}$$

In general, when the momentum is a 'good' operator, which means that it is measurable, the eigenvalue is a real quantity. Then, the imaginary part vanishes, and (2.15) is simply the product of the velocity and the probability, which yields the probability current.

The result (2.15) differs from the earlier form that appears in (2.9). If the expectation of the momentum is real, then the two forms agree, as the gradient of the action just gives the momentum. On the other hand, if the expectation of the momentum is not real, then the two results differ. For example, if the average momentum were entirely imaginary, then (2.15) would yield zero identically, while (2.9) would give a non-zero result. However, (2.9) was obtained by separating the real and imaginary parts of (2.6), and the result in this latter equation assumed that S was entirely real. An imaginary momentum would require that S be other than purely real. Thus, (2.6) was obtained for a very special form of the wave function. On the other hand, (2.15) results from a quite general wave function, and while the specific result depended upon a plane wave, the approach was not this limited. If (2.1) is used for the general wave function, then (2.15) is evaluated using the expectation values of the momentum, and suggests that in fact these eigenvalues should be real, *if a real current is to be measured.*

By real eigenvalues, we simply recognize that if an operator A can be measured by a particular wave function, then this operator produces the eigenvalue $\underline{a}$, which is a real quantity (we may assert without proof that one can only measure real numbers in a measurement). This puts certain requirements upon the operator A, as we note that

$$\langle A \rangle = (\Psi, A\Psi) = (\Psi, \underline{a}\Psi) = \underline{a} \tag{2.16}$$

for a properly normalized wave function. Now,

$$\underline{a}^* = (\Psi, \underline{a}^*\Psi) = (\underline{a}\Psi, \Psi) = (A\Psi, \Psi) = (\Psi, A^+\Psi) \tag{2.17}$$

where the symbol $^+$ indicates the *adjoint* operator. If the eigenvalues are real, as required for a measurable quantity, the corresponding operator must be self-adjoint; for example, $\underline{a} = \underline{a}^* \Longrightarrow A = A^+$. Such operators are known as *Hermitian* operators. The most common example is just the total-energy operator, as the energy is most often measured in systems. Not all operators are Hermitian, however, and the definition of the probability current allows for consideration of those cases in which the momentum may not be a real quantity and may not be measurable, as well as those more normal cases in which the momentum is measurable.

2.3 SOME SIMPLE CASES

The Schrödinger equation is a partial differential equation both in position space and in time. Often, such equations are solvable by separation of variables, and this is also the case here. We proceed by making the *ansatz* that the wave function may be written in the general form $\Psi(x,t) \equiv \Psi(x)\chi(t)$. If we insert this into the Schrödinger equation (2.12), and then divide by this same wave function, we obtain

$$\frac{i\hbar}{\chi}\frac{\partial\chi}{\partial t} = -\frac{\hbar^2}{2m\Psi}\frac{\partial^2\Psi}{\partial x^2} + V(x). \tag{2.18}$$

We have acknowledged here that the potential energy term is almost always a static interaction, which is only a function of position. Then, the left-hand side is a function of time alone, while the right-hand side is a function of position alone. This can be achieved solely if the two sides are equal to a constant. The appropriate constant has earlier been identified as the energy $\mathcal{E}$. These lead to the general result for the energy function

$$\chi(t) = e^{-i\mathcal{E}t/\hbar} \tag{2.19}$$

and the *time-independent Schrödinger equation*

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + V(x)\Psi(x) = \mathcal{E}\Psi(x). \tag{2.20}$$

This last equation describes the quantum wave mechanics of the static system, where there is no time variation. Let us now turn to a few examples.

2.3.1 The free particle

We begin by first considering the situation in which the potential is zero. Then the time-independent equation becomes

$$\frac{\partial^2\Psi}{\partial x^2} + k^2\Psi(x) = 0 \tag{2.21}$$

where

$$\frac{\hbar^2k^2}{2m} = \mathcal{E} \qquad k = \sqrt{\frac{2m\mathcal{E}}{\hbar^2}}. \tag{2.22}$$

The solution to (2.21) is clearly of the form of sines and cosines, but here we will take the exponential terms, and

$$\Psi(x) = Ae^{ikx} + Be^{-ikx}. \tag{2.23}$$

These are just the plane-wave solutions with which we began our treatment of quantum mechanics. The plane-wave form becomes more obvious when the time variation (2.19) is re-inserted into the total wave function. Here, the amplitude is spatially homogeneous and requires the use of the box normalization conditions discussed in the previous chapter.

If we are in a system in which the potential is not zero, then the solutions become more complicated. We can redefine the wave vector $\boldsymbol{k}$ as

$$k = \sqrt{\frac{2m[\mathcal{E} - V(x)]}{\hbar^2}}. \tag{2.24}$$

If the potential is slowly varying with distance, then the phase of the wave function makes a great many oscillations in a distance over which the variation in potential is small. Then, we can still use the result (2.23) for the wave function. However, for this to be the case, we require that the spatial variation be small. One might try to meet this requirement with the Bohm potential, the last term on the left-hand side of (2.7), but this earlier result was obtained by assuming a very special form for the wave function. In the present case, it is desired that the variation of the momentum with position not lead to extra terms in the Schrödinger equation, and this requirement can be simply stated by requiring

$$\frac{\lambda}{V}\frac{\partial V}{\partial x} \ll 1 \tag{2.25}$$

which simply says that the variation over a wavelength should be small. For most cases, this can be handled by treating rapid variation in the potential through boundary conditions, but we shall return to a treatment of the spatially varying potential through an approximation technique (the WKB approximation) in chapter 3. This approximate treatment of the wave function in the spatially varying potential case uses the solutions of (2.23), with the exponential factors replaced by

$$\exp\left[\pm \int^{x} \sqrt{\frac{2m[\mathcal{E} - V(x')]}{\hbar^2}}\,\mathrm{d}x'\right]. \tag{2.26}$$

However, it is important to note that solutions such as (2.26) do not satisfy the Schrödinger equation, and rely upon a sufficiently slow variation in the potential with position. The problem is that when the potential varies with position, (2.20) changes from a simple second-order ordinary differential equation to one with varying coefficients. These usually generate quite unusual special functions as the solutions.

2.3.2 A potential step

To begin to understand the role that the potential plays, let us investigate a simple potential step, in which the potential is defined as

$$V = V_0\Theta(x) \qquad \text{with} \quad V_0 > 0 \tag{2.27}$$

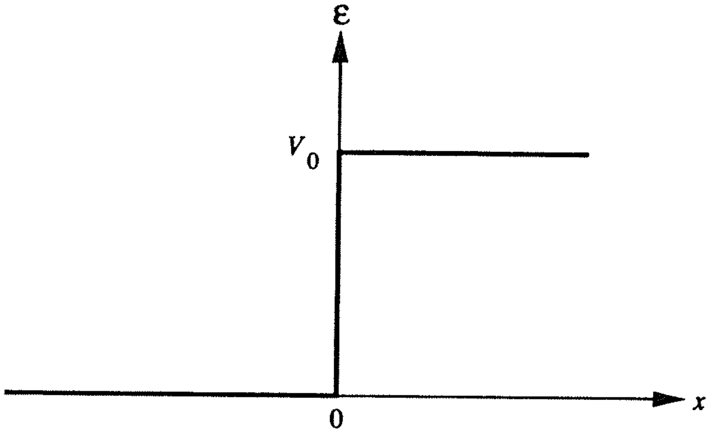


Figure 2.1 Schematic view of the potential of (2.27) which is non-zero (and constant) only in the positive half-space.

where $\Theta(x)$ is the Heaviside step function in which $\Theta = 1$ for $x \geqslant 0$, and $\Theta = 0$ for $x < 0$. This is shown in figure 2.1. Thus, the potential has a height of V_0 for positive x, and is zero for the negative-x region. This potential creates a barrier to the wave function, and a wave incident from the left (the negative region) will have part (or all) of its amplitude reflected from the barrier. The results that are obtained depend upon the relative energy of the wave. If the energy is less than V_0, the wave cannot propagate in the region of positive x. This is clearly seen from (2.24), as the wave vector is imaginary for $\mathcal{E} < V_0$. Only one exponent can be retained, as we require that the wave function remain finite (but zero) as $x \to \infty$.

Case I. $\mathcal{E} < V_0$

Let us first consider the low-energy case, where the wave is a non-propagating wave for $x > 0$. In the negative half-space, we consider the wave function to be of the form of (2.23), composed of an incident wave (the positive-exponent term) and a reflected wave (the negative-exponent term). In the positive half-space, the solution of the Schrödinger equation is simply

$$\Psi = C\mathrm{e}^{-\gamma x} \tag{2.28}$$

where

$$\gamma = \sqrt{\frac{2m[V_0 - \mathcal{E}]}{\hbar^2}}. \tag{2.29}$$

Here, we have defined a wave function in two separate regions, in which the potential is constant in each region. These two wave functions must be smoothly joined where the two regions meet.

While three constants are defined (A, B, C), one of these is defined by the resultant normalization of the wave function (we could e.g. let $A = 1$ without loss of generality). Two boundary conditions are required to evaluate the other two coefficients in terms of A. The boundary conditions can vary with the problem, but one must describe the continuity of the probability across the interface between the two regions. Thus, one boundary condition is that the wave function itself must be continuous at the interface, or

$$A + B = C. \tag{2.30}$$

To obtain a second boundary condition, we shall require that the derivative of the wave function is also continuous (that this is a proper boundary condition can be found by integrating (2.20) over a small increment from $x - \varepsilon$ to $x + \varepsilon$, which shows that the derivative of the wave function is continuous as long as this range of integration does not include an infinitely large potential or energy). In some situations, we cannot specify such a boundary condition, as there may not be a sufficient number of constants to evaluate (this will be the case in the next section). Equating the derivatives of the wave functions at the interface leads to

$$\mathrm{i}k(A - B) = -\gamma C. \tag{2.31}$$

This last equation can be rearranged by placing the momentum term in the denominator on the right-hand side. Then adding (2.30) and (2.31) leads to

$$\frac{C}{A} = \frac{2\mathrm{i}k}{\mathrm{i}k - \gamma}. \tag{2.32}$$

This result can now be used in (2.30) to find

$$\frac{B}{A} = \frac{\mathrm{i}k + \gamma}{\mathrm{i}k - \gamma}. \tag{2.33}$$

The amplitude of the reflected wave is unity, so there is no probability amplitude transmitted across the interface. In fact, the only effect of the interface is to phase shift the reflected wave; that is, the wave function is ($x < 0$)

$$\Psi(x) = A\left[\mathrm{e}^{ikx} + \mathrm{e}^{-\mathrm{i}(kx+\theta)}\right] \tag{2.34}$$

where

$$\theta = 2\tan^{-1}\left(\frac{\gamma}{k}\right). \tag{2.35}$$

The probability amplitude is given by

$$|\Psi|^2 = 2A[1 + \cos(2kx + \theta)] \qquad x < 0. \tag{2.36}$$

As may have been expected, this is a *standing-wave* pattern, with the probability oscillating from 0 to twice the value of A. The first peak occurs at a distance $x = -\theta/2k$, that is, the distance to the first peak is dependent upon the phase shift at the interface. If the potential amplitude is increased without limit, $V_0 \to \infty$, the damping coefficient $\gamma \to \infty$, and the phase shift approaches π. However, the first peak occurs at a value of $kx = \pi/2$, which also leads to the result that *the wave function becomes zero* at $x = 0$. We cannot examine the other limit ($V_0 \to 0$), as we do not have the proper transmitted wave, but this limit can be probed when the transmission mode is examined. It may also be noted that a calculation of the probability current for $x > 0$ leads immediately to zero as the wave function is real. Thus, no probability current flows into the right half-plane. It is a simple calculation to show that the net probability current in the left half-plane vanishes as well, as the reflected wave carries precisely the same current away from the interface as the incident wave carries toward the interface.

Case II. $\mathcal{E} > V_0$

We now turn to the case in which the wave can propagate on both sides of the interface. As above, the wave function in the left half-space is assumed to be of the form of (2.23), which includes both an incident wave and a reflected wave. Similarly, the transmitted wave will be assumed to be of the form

$$\Psi(x > 0) = C\mathrm{e}^{ik'x} \tag{2.37}$$

where k' is given by the right-hand side of (2.24). Again, we will match both the wave function and its derivative at $x = 0$. This leads to

$$\begin{aligned} A + B &= C \\ ik(A - B) &= ik'C. \end{aligned} \tag{2.38}$$

These equations can now be solved to obtain the constants C and B in terms of A. One difference here from the previous treatment is that these will be real numbers now, rather than complex numbers. Indeed, adding and subtracting the two equations of (2.38) leads to

$$\frac{C}{A} = \frac{2k}{k + k'} \qquad \frac{B}{A} = \frac{k - k'}{k + k'}. \tag{2.39}$$

Here, we see that if $V_0 \to 0$, $k' \to k$ and the amplitude of the reflected wave vanishes, and the amplitude of the transmitted wave is equal to the incident wave.

The probability current in the left-hand and right-hand spaces is found through the use of (2.14). For the incident and transmitted waves, these currents are simply

$$J_C = \frac{\hbar k'}{m}\left(\frac{2k}{k + k'}\right)^2 \qquad J_A = \frac{\hbar k}{m}. \tag{2.40}$$

The transmission coefficient is defined as the ratio of the transmitted current to the incident current, or

$$T = \frac{J_C}{J_A} = \frac{4kk'}{(k + k')^2} \tag{2.41}$$

which becomes unity when the potential goes to zero. By the same token, the reflection coefficient can be defined from the ratio of the reflected current to the incident current, or

$$R = -\frac{J_B}{J_A} = \left(\frac{k - k'}{k + k'}\right)^2 . \tag{2.42}$$

This leads to the result that

$$T + R = 1. \tag{2.43}$$

A critical point arises when $k' = 0$, that is, the energy is resonant with the top of the potential barrier. For this energy, the reflection coefficient from (2.42) is 1, so the transmission coefficient must vanish. The forms that have been used to solve for the wave function in the right-hand plane are not appropriate, as they are of exponential form. Here, however, the second derivative vanishes as the two terms with the potential energy and the energy cancel each other. This leads to a solution of the form $\Psi = C + Dx$, but D must vanish in order for the wave function to remain finite at large x. For the derivative of the wave function then to be continuous across the interface, (2.38) must become $B = A$. As a result of the first of equations (2.38), we then must have $C = 2A$. However, this constant wave function has no probability current associated with it, so the incident wave is fully reflected, consistent with $R = 1$. It is also reassuring that $C = 2A$ is consistent with (2.32) in the limit of $\gamma \to 0$, which also occurs at this limiting value of the energy.

For energies above the potential barrier height, the behaviour of the wave at the interface is quite similar in nature to what occurs with an optical wave at a dielectric discontinuity. This is to be expected as we are using the wave representation of the particle, and should expect to see optical analogues.

2.4 THE INFINITE POTENTIAL WELL

If we now put two barriers together, we have a choice of making a potential in which there is a barrier between two points in space, or a well between two points in space. The former will be treated in the next chapter. Here, we want to consider the latter case, as shown in figure 2.2. In this case the two barriers are located at $|x| = a$. In general, the wave function will penetrate into the barriers a distance given roughly by the decay constant γ. Before we consider this general case (treated in the next section), let us first consider the simpler case in which the amplitude of the potential increases without limit; that is, $V_0 \to \infty$.

From the results obtained in the last chapter, it is clear that the wave function decays infinitely rapidly under this infinite barrier. This leads to a boundary condition that requires the wave function to vanish at the barrier interfaces, that is $\Psi = 0$ at $|x| = a$. Within the central region, the potential vanishes, and the Schrödinger equation becomes just (2.21), with the wave vector defined by

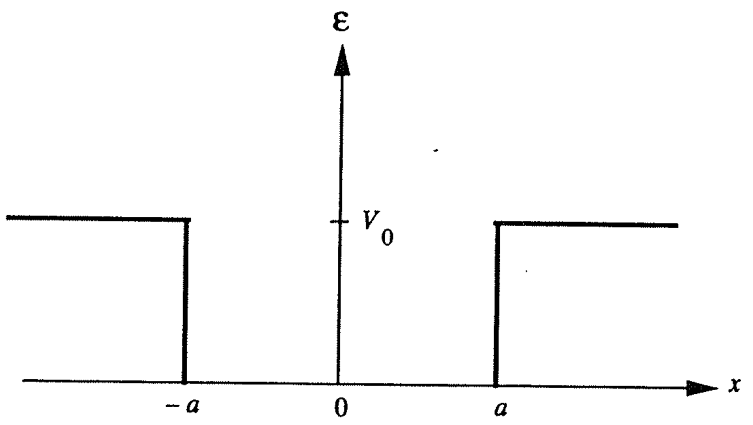


Figure 2.2 A potential well is formed by two barriers located at $|x| = a$.

(2.22). The solution is now given, just as in the free-particle case, by (2.23). At the right-hand boundary, this leads to the situation

$$Ae^{ika} + Be^{-ika} = 0 \tag{2.44}$$

and at the left-hand boundary,

$$Ae^{-ika} + Be^{ika} = 0. \tag{2.45}$$

Here, we have two equations with two unknowns, apparently. However, one of the constants must be determined by normalization, so only A or B can be treated as unknown constants. The apparent dilemma is resolved by recognizing that the wave vector k cannot take just any value, and the allowed values of k are recognized as the second unknown. Since the two equations cannot give two solutions, they must be *degenerate*, and the determinant of coefficients must vanish, that is

$$\begin{vmatrix} e^{ika} & e^{-ika} \\ e^{-ika} & e^{ika} \end{vmatrix} = 0. \tag{2.46}$$

This leads to the requirement that

$$\sin(2ka) = 0 \tag{2.47}$$

or

$$k = \frac{n\pi}{2a} \qquad \mathcal{E}_n = \frac{n^2\pi^2\hbar^2}{8ma^2} \qquad n = 1, 2, 3, \ldots . \tag{2.48}$$

Thus, there are an infinity of allowed energy values, with the spacing increasing quadratically with the index n.

In order to find the wave function corresponding to each of the energy levels, we put the value for k back into one of the equations above for the boundary conditions; we chose to use (2.44). This leads to

$$\frac{B}{A} = -e^{in\pi} = (-1)^{n+1}. \tag{2.49}$$

Thus, as we move up the hierarchy of energy levels, the wave functions alternate between cosines and sines. This can be summarized as

$$\Psi n(x) = \begin{cases} A\cos(n\pi x/(2a)) & n \text{ odd} \\ A\sin(n\pi x/(2a)) & n \text{ even} \end{cases} \tag{2.50}$$

These can be combined by offsetting the position, so that

$$\Psi_n(x) = A\sin\left[\frac{n\pi}{2a}(x+a)\right]. \tag{2.51}$$

This last solution fits both boundary conditions, and yields the two solutions of (2.50) when the multiple-angle expansion of the sine function is used. Of course, each indexed wave function of (2.51) corresponds to one of the Fourier expansion terms in the Fourier series that represents a square barrier. In fact, (2.21) is just one form of a general boundary value problem in which the Fourier series is a valid solution.

We still have to normalize the wave functions. To do this, we use (2.51), and the general inner product with the range of integration now defined from $-a$ to a. This leads to

$$(\Psi_n, \Psi_n) = A^2 \int_{-a}^{a} \sin^2\left[\frac{n\pi}{2a}(x+a)\right] dx = 1. \tag{2.52}$$

This readily leads to the normalization

$$A = \frac{1}{\sqrt{a}}. \tag{2.53}$$

If the particle resides exactly in a single energy level, we say that it is in a *pure* state. The more usual case is that it moves around between the levels and on the average many different levels contribute to the total wave function. Then the total wave function is a sum over the Fourier series, with coefficients related to the probability that each level is occupied. That is,

$$-\Psi(x) = \sum_n \frac{c_n}{\sqrt{a}} \sin\left[\frac{n\pi}{2a}(x+a)\right] \tag{2.54}$$

and the probability that the individual state n is occupied is given by $|c_n|^2$. This is subject to the limitation on total probability that

$$\sum_n |c_n|^2 = 1. \tag{2.55}$$

This summation over the available states for a particular system is quite universal and we will encounter it often in the coming sections and chapters.

It may be seen that the solutions to the Schrödinger equation in this situation were a set of odd wave functions and a set of even wave functions in (2.50), where by even and odd we refer to the symmetry when $x \to -x$. This is a general result when the potential is an even function; that is, $V(x) = V(-x)$. In the Schrödinger equation, the equation itself is unchanged when the substitution $x \to -x$ is made providing that the potential is an even function. Thus, for a bounded wave function, $\Psi(-x)$ can differ from $\Psi(x)$ by no more than a constant, say α. Repeated application of this variable replacement shows that $\alpha^2 = 1$, so α can only take on the values ± 1, which means that the wave function is either even or odd under the variable change. We note that this is only the case when the potential is even; no such symmetry exists when the potential is odd. Of course, if the wave function has an unbounded form, such as a plane-wave, it is not required that the wave function have this symmetry, although both symmetries are allowed for viable solutions.

2.5 THE FINITE POTENTIAL WELL

Now let us turn to the situation in which the potential is not infinite in amplitude and hence the wave function penetrates into the regions under the barriers. We continue to treat the potential as a symmetric potential centred about the point $x = 0$. However, it is clear that we will want to divide our treatment into two cases: one for energies that lie above the top of the barriers, and a second for energies that confine the particle into the potential well. For this, it is convenient (and we emphasize that it is only for convenience) to shift the energy scale so that the barrier heights are at $\mathcal{E} = 0$, and the potential-free region is shifted actually to fall inside a negative potential well of amplitude $-V_0$. This is shown in figure 2.3. Thus, for energies greater than zero, the particles are free to propagate, while for energies less than zero, the particle is confined in the potential well in a series of energy levels known as *bound states*.

Case I. $V_0 < \mathcal{E} < 0$

For energies below zero, the particle has freely propagating characteristics only for the range $|x| < a$, for which the Schrödinger equation becomes

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + k^2\Psi = 0 \qquad k^2 = \frac{2m}{\hbar^2}(V_0 + \mathcal{E}). \tag{2.56}$$

In (2.56), it must be remembered that V_0 is the magnitude of the negative potential well, and is a positive quantity, while $\mathcal{E}$ is a negative quantity. Similarly, in the range $|x| > a$, the Schrödinger equation becomes

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} - \gamma^2\Psi = 0 \qquad \gamma^2 = \frac{2m|\mathcal{E}|}{\hbar^2}. \tag{2.57}$$

We saw at the end of the last section that with the potential being a symmetric quantity, the solutions for the Schrödinger equation would have either even or

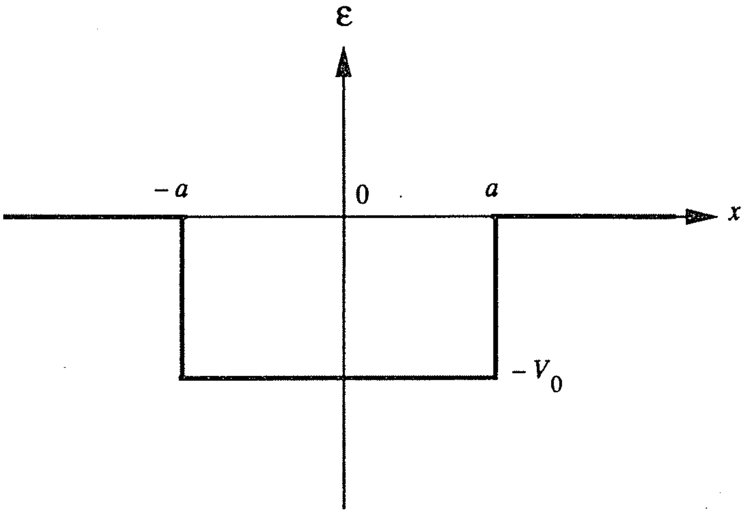


Figure 2.3 The finite potential well, in which the energy axis has been shifted for convenience.

odd symmetry. The basic properties of the last section will carry over to the present case, and we expect the solutions in the well region to be either sines or cosines. Of course, these solutions have the desired symmetry properties, and will allow us to solve for the allowed energy levels somewhat more simply.

Thus, we can treat the even and odd solutions separately. In either case, the solutions of (2.57) for the damped region will be of the form $Ce^{-\gamma|x|}$, $|x| > a$. We can match this to the proper sine or cosine function. However, in the normal case, both the wave function and its derivative are matched at each boundary. If we attempt to do the same here, this will provide four equations. However, there are only two unknowns—the amplitude of C relative to that of either the sine or cosine wave and the allowed values of the wave vector $\boldsymbol{k}$ (and hence γ, since it is not independent of $\boldsymbol{k}$) for the bound-state energy levels. We can get around this problem in one fashion, and that is to make the ratio of the derivative to the wave function itself continuous. That is, we make the logarithmic derivative Ψ'/Ψ continuous. (This is obviously called the logarithmic derivative since it is the derivative of the logarithm of Ψ.) Of course, if we choose the solutions to have even or odd symmetry, the boundary condition at $-a$ is redundant, as it is the same as that at a by these symmetry relations.

Let us consider the even-symmetry wave functions, for which the logarithmic derivative is

$$\frac{-k\sin(kx)}{\cos(kx)} = -k\tan(kx). \tag{2.58}$$

Similarly, the logarithmic derivative of the damped function is merely $-\gamma\,\mathrm{sgn}(x)$, where $\mathrm{sgn}(x)$ is the sign of x and arises because of the magnitude in the argument of the exponent. We note that we can match the boundary condition at either a or $-a$, and the result is the same, a fact that gives rise to the even function that we are using. Thus, the boundary condition is just

$$k\tan(ka) = \gamma. \tag{2.59}$$

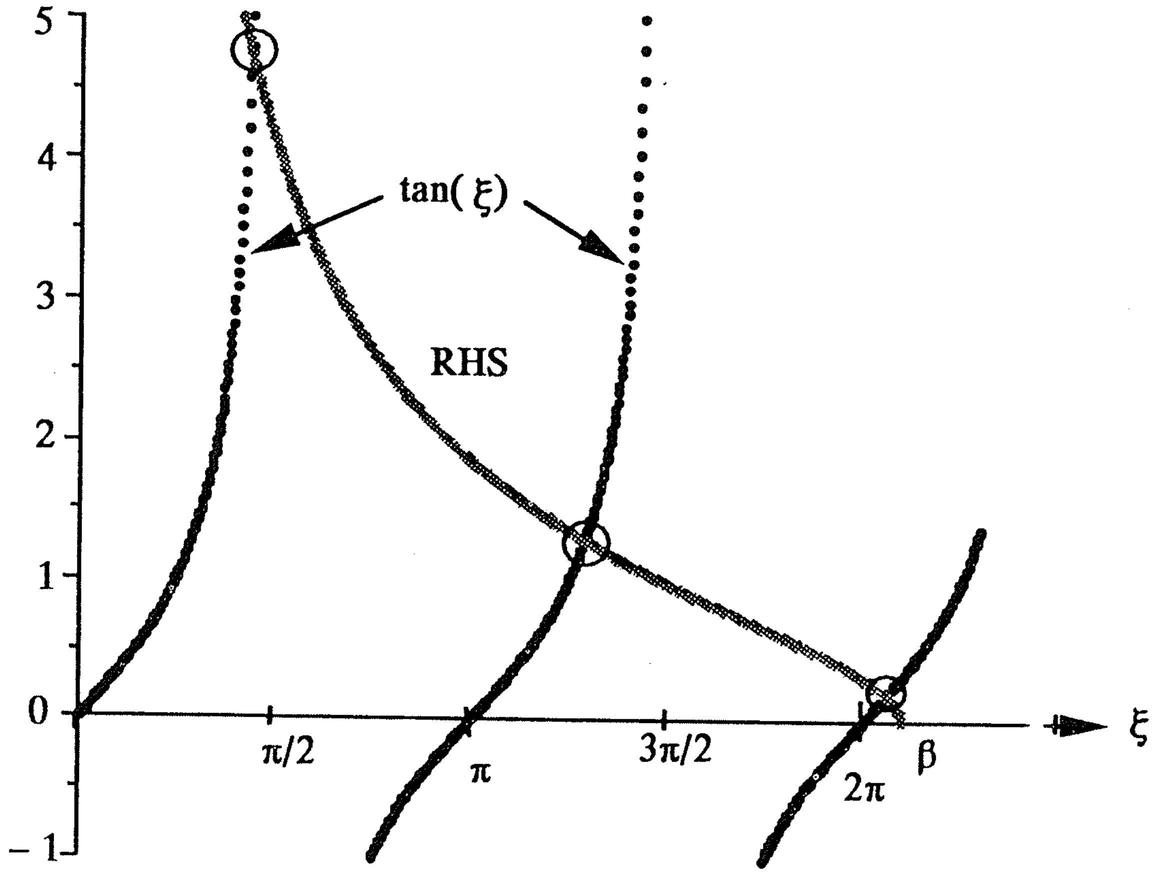


Figure 2.4 The graphical solutions of (2.6) are indicated by the circled crossings.

This transcendental equation ,now determines the allowed values of the energy for the bound states. If we define the new, reduced variable $\xi = ka$, then this equation becomes

$$\tan(\xi) = \frac{\gamma}{k} = \sqrt{\frac{\beta^2}{\xi^2} - 1} \qquad \beta^2 = \frac{2mV_0a^2}{\hbar^2}. \tag{2.60}$$

The right-hand side of the transcendental equation is a decreasing function, and it is only those values for which the energy lies in the range $(-V_0, 0)$ that constitute bound states. In general, the solution must be found graphically. This is shown in figure 2.4, in which we plot the left-hand side of (2.60) and the right-hand side separately. The crossings (circled) are allowed energy levels.

As the potential amplitude is made smaller, or as the well width is made smaller, the value of β is reduced, and there is a smaller range of ξ that can be accommodated before the argument of the square root becomes negative. Variations in the width affect both parameters, so we should prefer to think of variations in the amplitude, which affects only β. We note, however, that the right-hand side varies from infinity (for $\xi = 0$) to zero (for $\xi = \beta$), regardless of the value of the potential. A similar variation, in inverse range, occurs for the tangent function (that is, the tangent function goes to zero for $\xi = 0$ or $n\pi$, and the tangent diverges for ξ taking on odd values of $\pi/2$). Thus, there is always at least one crossing. However, there may only be the one. As the potential amplitude is reduced, the intercept β of the decreasing curve in figure 2.4 moves toward the origin. Thus, the solution point approaches $\xi = 0$, or $k = 0$. By expanding the tangent function for small ξ, it is found that the solution is approximately $\beta \simeq \xi$. However, this requires $\mathcal{E} \simeq 0$, which means that the energy level is just at the top of the well. Thus, there is at least one crossing of the curves for $\xi < \pi/2$. For larger values of the amplitude of the potential, the zero point (β) moves to the right and more allowed energy levels appear for the even functions. It is clear from the construction of figure 2.4 that at least one solution must occur, even if the width is the parameter made smaller, as the ξ-axis intersection cannot be reduced to a point where it does not cross the $\tan(\xi)$ axis at least once. The various allowed energy levels may be identified with the integers $1, 3, 5, \ldots$ just as is the case for the infinite well (it is a peculiarity that the even-symmetry wave functions have the odd integers) although the levels do not involve exact integers any more.

Let us now turn to the odd-symmetry wave functions in (2.50). Again, the logarithmic derivative of the propagating waves for $|x| < a$ may be found to be

$$\frac{k\cos(kx)}{\sin(kx)} = k\,\mathrm{cotan}(kx). \tag{2.61}$$

The logarithmic derivative for the decaying wave functions remains $-\gamma\,\mathrm{sgn}(x)$, and the equality will be the same regardless of which boundary is used for matching. This leads to

$$k\,\mathrm{cotan}(kx) = -\gamma \tag{2.62}$$

or

$$\mathrm{cotan}(\xi) = -\sqrt{\frac{\beta^2}{\xi^2} - 1}. \tag{2.63}$$

Again, a graphical solution is required. This is shown in figure 2.5. The difference between this case and that for the even wave functions is that the left-hand side of (2.63) starts on the opposite side of the ξ-axis from the right-hand side and we are not guaranteed to have even one solution point. On the other hand, it may be seen by comparing figures 2.4 and 2.5 that the solution points that do occur lie in between those that occur for the even-symmetry wave functions. Thus, these may be identified with the integers $2, 4, \ldots$ even though the solutions do not involve exact integers.

We can summarize these results by saying that for small amplitudes of the potential, or for small widths, there is at least one bound state lying just below the top of the well. As the potential, or width, increases, additional bound states become possible. The first (and, perhaps, only) bound state has an even-symmetry wave function. The next level that becomes bound will have odd symmetry. Then a second even-symmetry wave function will be allowed, then an odd-symmetry one, and so on. In the limit of an infinite potential well, there are an infinite number of bound states whose energies are given by (2.48).

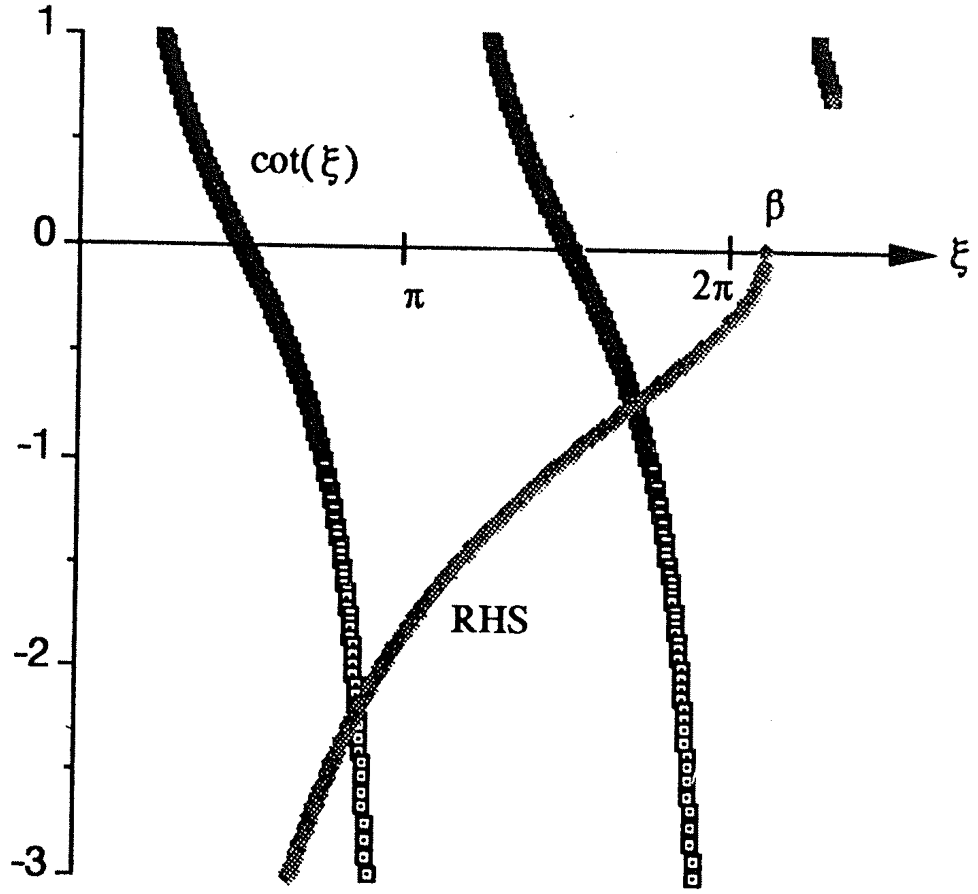


Figure 2.5 The graphical solution to (2.63).

Once the energy levels are determined for the finite potential well, the wave functions can be evaluated. We know the form of these functions, and the energy levels ensure the continuity of the logarithmic derivative, so we can generally easily match the partial wave functions in the well and in the barriers. One point that is obvious from the preceding discussion is that the energy levels lie below those of the infinite well. This is because the wave function penetrates into the barriers, which allows for example a sine function to *spread out* more, which means that the momentum wave vector k is slightly smaller, and hence corresponds to a lower energy level. Thus, the sinusoidal function does not vanish at the interface for the finite-barrier case, and in fact couples to the decaying exponential within the barrier. The typical sinusoid then adopts long exponential tails if the barrier is not infinite.

Some of the most interesting studies of these bound states have been directed at quantum wells in GaAs–AlGaAs heterojunctions. The alloy AlGaAs, in which there is about 20% AlAs alloyed into GaAs, has a band gap that is 0.25 eV larger than that of pure GaAs (about 1.75 eV versus 1.4 eV). A fraction of this band gap difference lies in the conduction band and the remainder in the valence band. Thus, if a GaAs layer is placed between two AlGaAs layers, a quantum well is formed both in the conduction band and in the valence band. Transitions between the hole bound states and the electron bound states can be probed optically, since these transitions will lie below the absorption band for the AlGaAs. Such an absorption spectrum is shown in figure 2.6. Transitions at the lowest heavy-

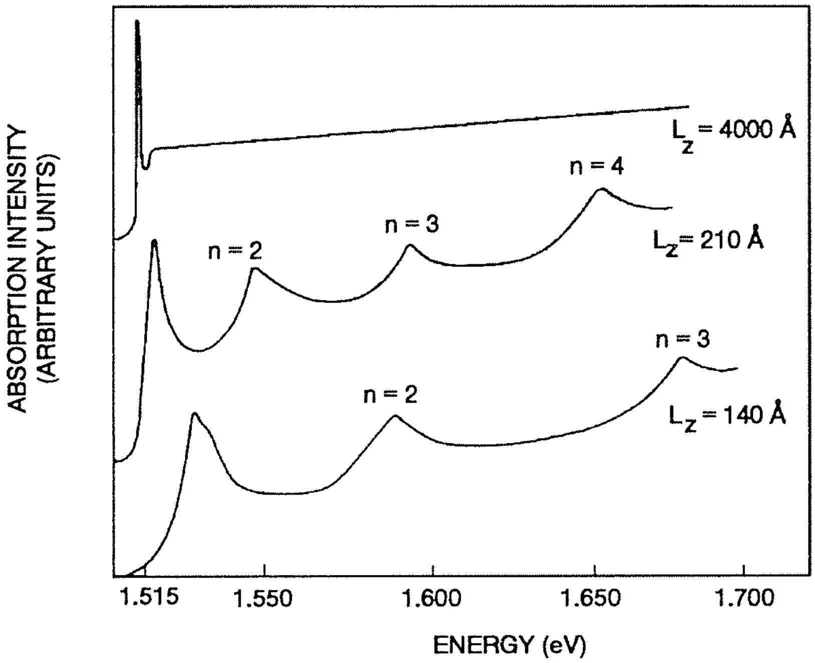


Figure 2.6 Absorption observed between bound states of holes and electrons in 21 nm and 14 nm quantum wells formed by placing a layer of GaAs between two layers of AlGaAs. For a well thickness of 400 nm, the absorption is uniform. (After Dingle *et al* (1974), by permission.)

hole to electron transition and the second heavy-hole to electron transition are seen (the spectrum is complicated by the fact that there are both heavy and light holes in the complicated valence band). The width of the absorption lines arises from thermal broadening of these states and broadening due to inhomogeneities in the width of the multiple wells used to see a sufficiently large absorption. Transitions such as these have been used actually to try to determine the band offset (the fraction of the band gap difference that lies in the valence band) through measurements for a variety of well widths. Such data are shown in figure 2.7, for this same system. While these data were used to try to infer that only 15% of the band gap difference lay in the valence band, these measurements are relatively insensitive to this parameter, and a series of more recent measurements gives this number as being more like 30%.

Case II. $\mathcal{E} > 0$

Let us now turn our attention to the completely propagating waves that exist for energies above the potential well. It might be thought that these waves will show no effect of the quantum well, but this is not the case. Each interface is equivalent to a dielectric interface in electromagnetics, and the thin layer is equivalent to a thin dielectric layer in which interference phenomena can occur. The same is expected to occur here. We will make calculations for these phenomena by calculating the transmission coefficient for waves propagating from the left (negative x) to the right (positive x).

Throughout the entire space, the Schrödinger equation is given by the form (2.56), with different values of k in the various regions. The value of k given

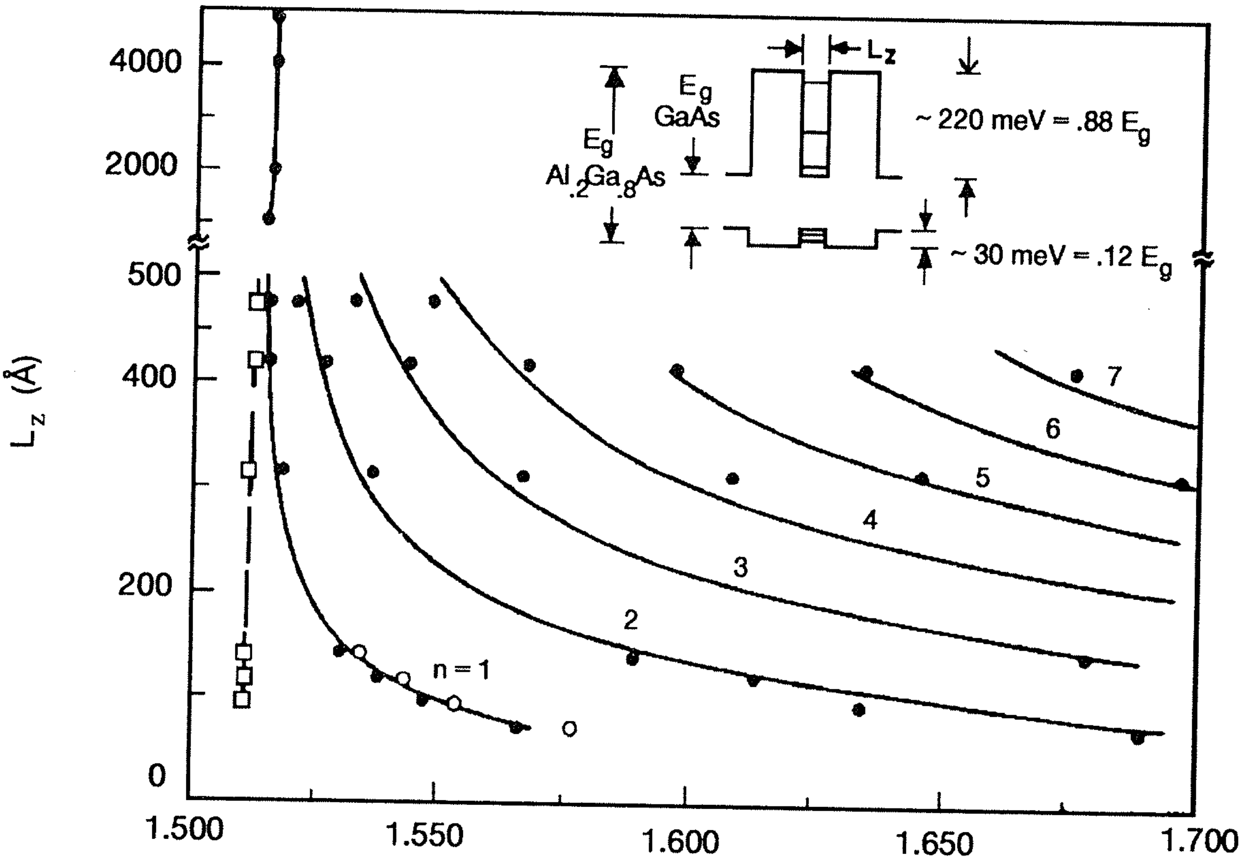


Figure 2.7 Variation of the absorption bands for transitions from heavy-hole (solid circles) and light-hole (open circles) levels to electron levels of the quantum wells as a function of well width. The solid curves are calculated positions. (After Dingle *et al* (1974), by permission.)

in (2.56) remains valid in the quantum well region, while for $|x| > a$,

$$k_0^2 = \frac{2m\mathcal{E}}{\hbar^2}. \tag{2.64}$$

For $x > a$, we assume that the wave function propagates only in the outgoing direction, and is given by

$$F\mathrm{e}^{\mathrm{i}k_0 x}. \tag{2.65}$$

In the quantum well region, we need to have waves going in both directions, so the wave function is assumed to be

$$C\mathrm{e}^{\mathrm{i}kx} + D\mathrm{e}^{-\mathrm{i}kx}. \tag{2.66}$$

Similarly, in the incident region on the left, we need to have a reflected wave, so the wave function is taken to be

$$\mathrm{e}^{\mathrm{i}k_0 x} + B\mathrm{e}^{-\mathrm{i}k_0 x} \tag{2.67}$$

where we have set $A = 1$ for convenience. We now develop four equations by using the continuity of both the wave function and its derivative at each of the two interfaces. This leads to the determinantal equation

$$\begin{bmatrix} 0 & \omega & \omega^{-1} & -\omega_0 \\ 0 & \omega & -\omega^{-1} & -(k_0/k)\omega_0 \\ -\omega_0 & \omega^{-1} & \omega & 0 \\ \omega_0 & (k/k_0)\omega^{-1} & -(k/k_0)\omega & 0 \end{bmatrix} \begin{bmatrix} B \\ C \\ D \\ F \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \omega^{-1} \\ \omega^{-1} \end{bmatrix}. \tag{2.68}$$

Here $\omega = \mathrm{e}^{\mathrm{i}ka}$, $\omega_0 = \mathrm{e}^{\mathrm{i}k_0 a}$. This can now be solved to find the coefficient of the outgoing wave:

$$F = \frac{\mathrm{e}^{-2\mathrm{i}k_0 a}}{\cos(2ka) - \mathrm{i}\left[(k^2 + k_0^2)/2kk_0\right]\sin(2ka)}. \tag{2.69}$$

Since the momentum wave vector is the same in the incoming region as in the outgoing region, the transmission coefficient can be found simply as the square of the magnitude of F in (2.69). This leads to

$$T = \frac{1}{1 + \left[(k^2 - k_0^2)/2kk_0\right]^2 \sin^2(2ka)}. \tag{2.70}$$

There are resonances, which occur when $2ka$ is equal to odd multiples of $\pi/2$, and for which the transmission is a minimum. The transmission rises to unity when $2ka$ is equal to even multiples of $\pi/2$, or just equal to $n\pi$. The reduction in transmission depends upon the amplitude of the potential well, and hence on the difference between k and k_0. We note that the transmission has minima that drop to small values only if the well is infinitely deep (and the energy of the wave is not infinite; i.e., $k_0 \gg k$). A deeper potential well causes a greater discontinuity in the wave vector, and this leads to a larger modulation of the transmission coefficient.

Such transmission modulation has been observed in studies of the transport of ballistic electrons across a GaAs quantum well base located between AlGaAs regions which served as the emitter and collector. The transport is shown in figure 2.8, and is a clear indication of the fact that quantum resonances, and quantum effects, can be found in real semiconductor devices in a manner that affects their characteristic behaviour. The device structure is shown in part (a) of the figure; electrons are injected (tunnel) through the barrier to the left (emitter side) of the internal GaAs quantum well at an energy determined by the Fermi energy in the emitter region (on the left of the figure). The injection coefficient, determined as the derivative of the injected current as a function of bias, reveals oscillatory behaviour due to resonances that arise from both the bound states and the so-called virtual states above the barrier. These are called virtual states as they are not true bound states but appear as variations in the transmission and reflection coefficients. Results are shown for devices with two different thicknesses of the quantum well, 29 and 51.5 nm. The injection coefficient is

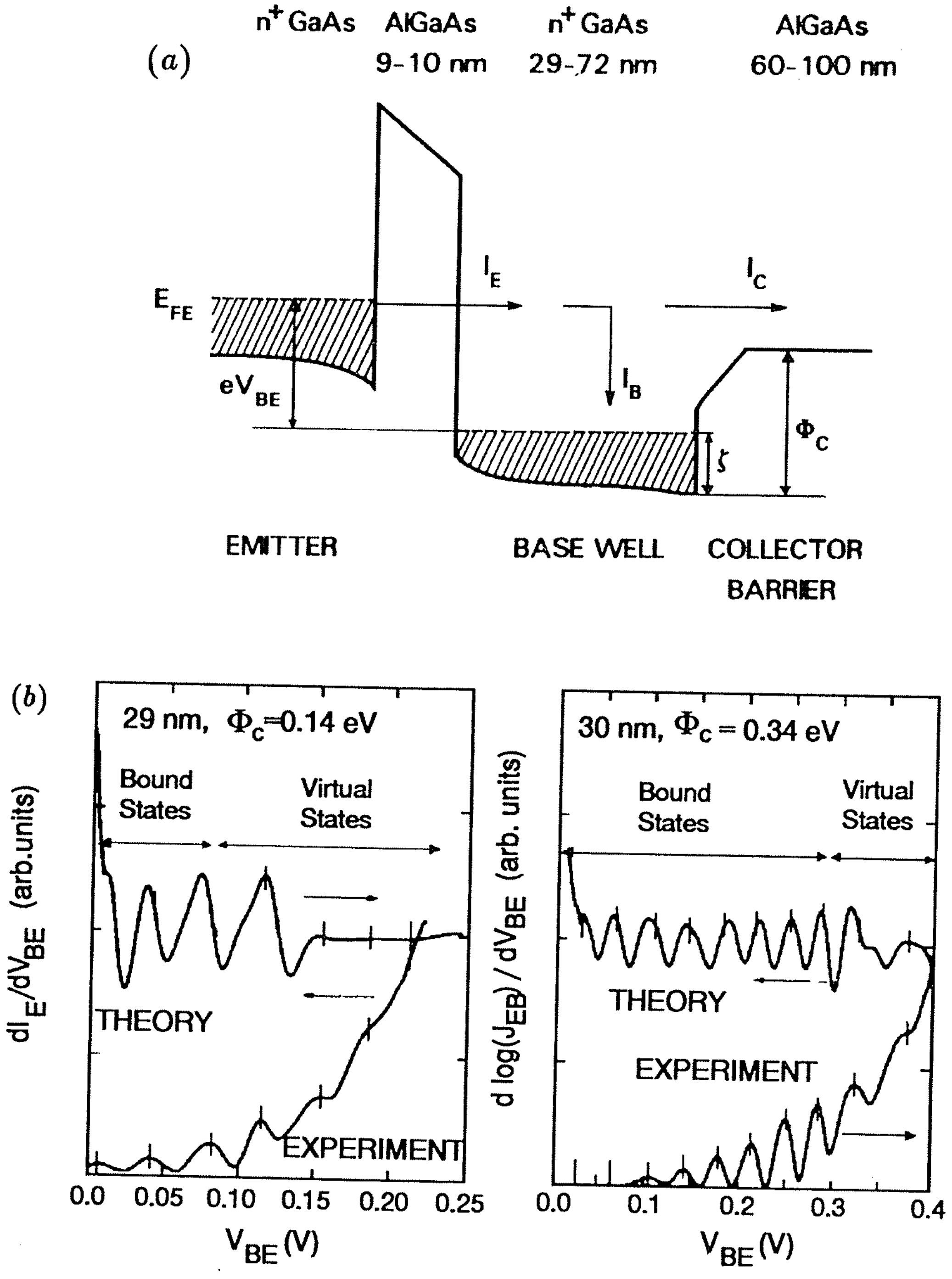


Figure 2.8 Transport of ballistic electrons through a double-barrier, ballistic transistor, whose potential profile is shown in (*a*). The quantum resonances of propagation over the well are evident in the density of collected electrons (*b*) for two different sizes. (After Heiblum *et al* (1987), by permission.)

shown rather than the transmission coefficient, as the former also illustrates the bound states.

It seems strange that a wave function that lies above the quantum well should not be perfectly transmitting. It is simple enough to explain this via the idea of 'dielectric discontinuity', but is this really telling the whole truth of the physics? Yes and no. It explains the physics with the mathematics, but it does not convey the understanding of what is happening. In fact, it is perhaps easier to think about the incident wave as a particle. When it impinges upon the region where the potential well exists, it cannot be trapped there, as its energy lies above the top of the well. However, the well potential can *scatter* the particle as it arrives. In the present case, the particle is scattered back into the direction from which it came with a probability given by the reflection coefficient. It proceeds in an unscattered state with a probability given by the transmission coefficient. In general, the scattering probability is non-zero, but at special values of the incident energy, the scattering vanishes. In this case, the particle is transmitted with unity probability. This type of potential scattering is quite special, because only the direction of the momentum (this is a one-dimensional problem) is changed, and the energy of the particle remains unchanged. This type of scattering is termed *elastic*, as in elastic scattering of a billiard ball from a 'cushion' in three dimensions. We will see other examples of this in the following chapters.

2.6 THE TRIANGULAR WELL

Another type of potential well is quite common in everyday semiconductor devices, such as the common metal–oxide–semiconductor (MOS) transistor (figure 2.9(*a*)). The latter is the workhorse in nearly all microprocessors and computers today, yet the presence of quantization has not really been highlighted in the operation of these devices. These devices depend upon capacitive control of the charge at the interface between the oxide and the semiconductor. If we consider a parallel-plate capacitor made of a metal plate, with an insulator made of silicon dioxide, and a second plate composed of the semiconductor silicon, we essentially have the MOS transistor. Voltage applied across the capacitor varies the amount of charge accumulated in the metal and in the semiconductor, in both cases at the interface with the insulator. On the semiconductor side, contacts (made of n-type regions embedded in a normally p-type material) allow one to pass current through the channel in which the charge resides in the semiconductor. Variation of the current, through variation of the charge via the capacitor voltage, is the heart of the transistor operation.

Consider the case in which the semiconductor is p-type, and hence the surface is in an 'inverted' condition (more electrons than holes) and mobile electrons can be drawn to the interface by a positive voltage on the metal plate (the channel region is isolated from the bulk of the semiconductor by the inversion process). The surface charge in the semiconductor is composed of two parts: (i) the surface electrons, and (ii) ionized acceptors from which the holes have been pushed into the interior of the semiconductor. In both cases the charge that results is negative and serves to balance the positive charge on the metal gate. The electron charge is localized right at the interface with the insulator, while the ionized acceptor charge is distributed over a large region. In fact, it is the localized electron charge that is mobile in the direction along the interface, and that is quantized in the resulting potential well. The field in the oxide is then given by the total surface charge through Gauss's law (we use the approximation of an infinite two-dimensional plane) as

$$E_s = \frac{e}{\varepsilon_{0x}}(N_a w + n_s) \tag{2.71}$$

where w is the thickness of the layer of ionized acceptors N_a (normal to the surface), the surface electron density n_s is assumed to be a two-dimensional sheet charge, and the permittivity is that of the oxide. On the semiconductor side of the interface, the normal component of $\boldsymbol{D}$ is continuous, which means that E in (2.71) is discontinuous by the dielectric constant ratio. Thus, just inside the interface, (2.71) represents the field if the oxide permittivity is replaced by that of the semiconductor. However, just a short distance further into the semiconductor, the field drops by the amount produced by the surface electron density. Thus, the average field in the semiconductor, in the region where the electrons are located, is approximately

$$E_s = \frac{e}{\varepsilon_s}\left(N_a w + \frac{n_s}{2}\right). \tag{2.72}$$

In this approximation, a constant electric field in this region gives rise to a linear potential in the Schrödinger equation (figure 2.9(*b*)). We want to solve for just the region inside the semiconductor, near to the oxide interface. Here, we can write the Schrödinger equation in the form

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + eE_s x\Psi = \mathcal{E}\Psi \qquad \text{for } x > 0. \tag{2.73}$$

We assume that the potential barrier at the interface is infinitely high, so no electrons can get into the oxide, which leads to the boundary condition that $\Psi(0) = 0$. The other boundary condition is merely that the wave function must remain finite, which means that it also tends to zero at large values of x.

To simplify the solution, we will make a change of variables in (2.73), which will put the equation into a standard form. For this, we redefine the position and energy variables as

$$z = \left(\frac{2meE_s}{\hbar^2}\right)^{1/3} x \qquad z_0 = \frac{2m\mathcal{E}}{\hbar^2}\left(\frac{\hbar^2}{2meE_s}\right)^{2/3}. \tag{2.74}$$

Then, using $\xi = z - z_0$, (2.73) becomes

$$\frac{\partial^2\Psi}{\partial\xi^2} - \xi\Psi = 0. \tag{2.75}$$

This is the Airy equation.

Airy functions are combinations of Bessel functions and modified Bessel functions. It is not important here to discuss their properties in excruciating detail. The important facts for us are that: (i) the Airy function $\mathrm{Ai}(-\xi)$ decays as an exponential for positive ξ, and (ii) $\mathrm{Ai}(\xi)$ behaves as a damped sinusoid with a period that also varies as ξ. For our purposes, this is all we need. The second solution of (2.75), the Airy functions $\mathrm{Bi}(\xi)$, diverge in each direction and must be discarded in order to keep the probability function finite. The problem is in meeting the desired boundary conditions. The requirement that the wave function decay for large x is easy. This converts readily into the requirement that the wave function decay for large ξ, which is the case for $\mathrm{Ai}(-\xi)$. However, the requirement that the wave function vanish at $x = 0$ is not arbitrarily satisfied for the Airy functions. On the other hand, the Airy functions are oscillatory. In the simple quantum well of the last two sections, we noted that the lowest bound state had a single peak in the wave function, while the second state had two, and so on. This suggests that we associate the vanishing of the wave function at $x = 0$ with the intrinsic zeros of the Airy function, which we will call a_s. Thus, choosing a wave function that put the first zero a_1 at the point $x = 0$ would fit all the boundary conditions for the lowest energy level (figure 2.9). Similarly, putting the second zero a_2 at $x = 0$ fits the boundary conditions for the next level, corresponding to $n = 2$. By this technique, we build the set of wave functions, and also the energy levels, for the bound states in the wells.

Here, we examine the lowest bound state as an example. For this, we require the first zero of the Airy function. Because the numerical evaluation of the Airy functions yields a complicated series, we cannot give exact values for the zeros. However, they are given approximately by the relation (Abramowitz and Stegun 1964)

$$a_s \simeq -\left(\frac{3\pi(4s-1)}{8}\right)^{2/3}. \tag{2.76}$$

Thus, the first zero appears at approximately $-(9\pi/8)^{2/3}$. Now, this may be related to the required boundary condition at $x = z = 0$ through

$$\xi = -\left(\frac{9\pi}{8}\right)^{2/3} = -z_0 = \frac{2m\mathcal{E}}{\hbar^2}\left(\frac{\hbar^2}{2meE_s}\right)^{2/3} \tag{2.77}$$

or

$$\mathcal{E}_1 = \frac{\hbar^2}{2m}\left(\frac{9\pi meE_s}{4\hbar^2}\right)^{2/3} \tag{2.78}$$

remembering, of course, that this is an approximate value since we have only an approximate value for the zero of the Airy function. In figure 2.10(a), the potential well, the first energy level and the wave function for this lowest bound state are shown. It can be seen from this that the wave function dies away exponentially in the region where the electron penetrates beneath the linear potential, just as for a normal step barrier.

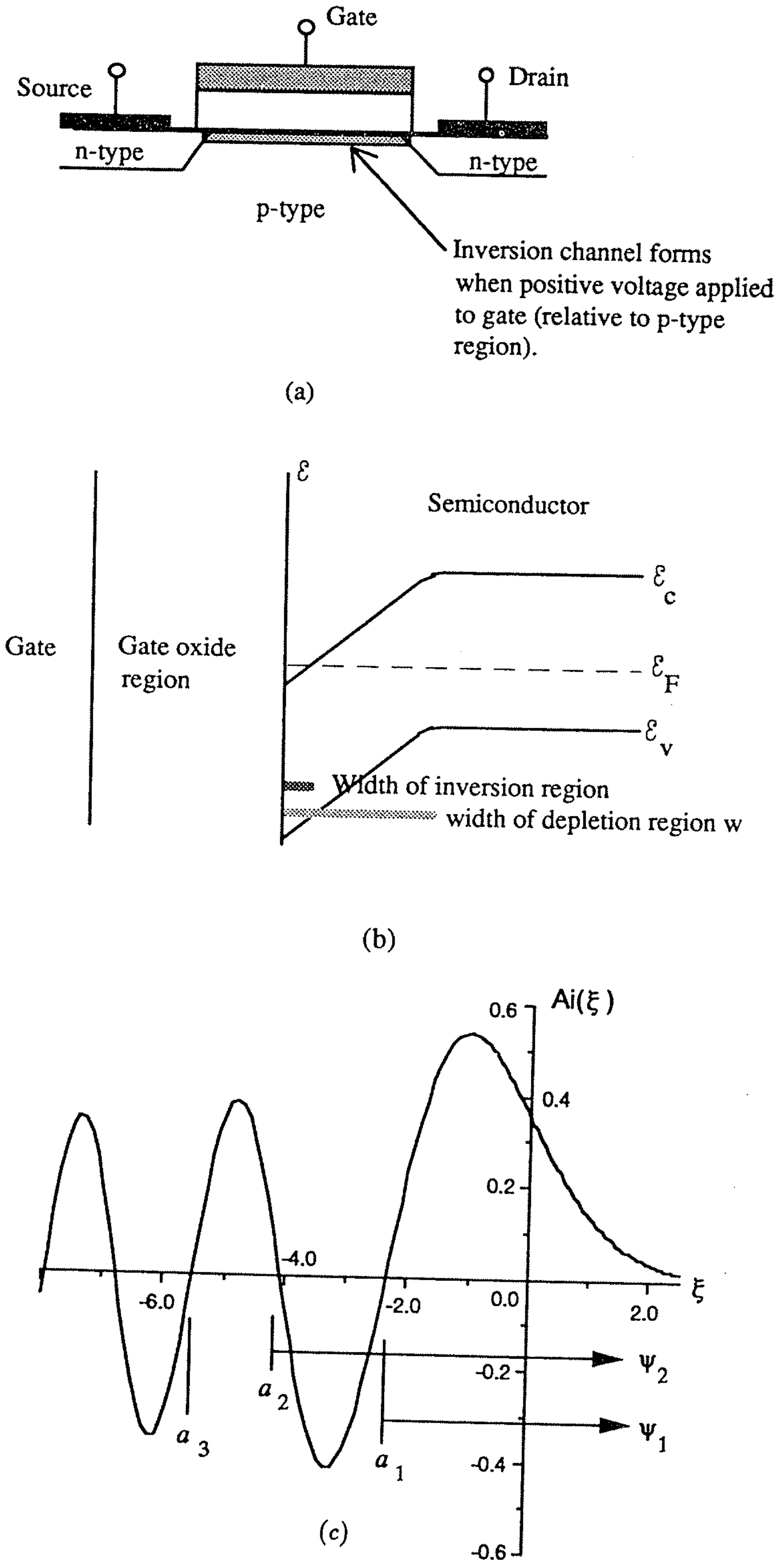


Figure 2.9 (*a*) A MOSFET, (*b*) the triangular potential, and (*c*) the Airy function and the use of the zeros to match the boundary conditions.

The quantization has the effect of moving the charge away from the surface. Classically, the free-electron charge density peaks right at the interface between the semiconductor and the oxide insulator, and then decays away into the semiconductor as

$$n(x) = n_s \exp\left[-\frac{eE_s x}{k_B T}\right]. \tag{2.79}$$

This decays to $1/e$ of the peak in a distance given by $k_B T/eE_s$. Typical values for the field may be of the order of 2 V across 20 nm of oxide, which leads to a field in the oxide of 10^6 V cm^{-1}, and this corresponds to a field in the semiconductor at the interface of (the oxide dielectric constant is about 3.8 while that for silicon is about 12) 3×10^5 V cm^{-1}. This leads to an effective thickness of the surface charge density of only about 0.9 nm, an incredibly thin layer. On the other hand, these values lead to a value for the lowest bound state of 50 meV, and an effective well width ($\mathcal{E}_1/eE$) of 1.7 nm. The wavelength corresponding to these electrons at room temperature is about 6 nm, so it is unlikely that these electrons can be confined in this small distance, and this is what leads to the quantization of these electrons. The quantized charge density in the lowest bound state is proportional to the square of the wave function, and the peak in this density occurs at the peak of the wave function, which is at the zero of the first derivative of the Airy function. These zeros are given by the approximate relation (Abramowitz and Stegun 1964)

$$a'_s = -\left(\frac{3\pi(4s-3)}{8}\right)^{2/3} \tag{2.80}$$

which for the lowest subband leads to $z_{\text{peak}} \simeq 2.1(3\pi)^{2/3}$. This leads to the peak occurring at a distance from the surface (e.g., from $-x_0$) of

$$x \simeq \left(\frac{\hbar^2}{2meE_s}\right)^{1/3}\left[\left(\frac{9\pi}{8}\right)^{2/3} - 2.1(3\pi)^{2/3}\right] \tag{2.81}$$

which for the above field gives a distance of 1.3 nm. The effective width of the quantum well, mentioned earlier, is larger than this, as this value is related to the 'half-width'. This value is smaller than the actual thermal de Broglie wavelength of the electron wave packet. The quantization arises from the confinement of the electron in this small region. In figure 2.10(*b*), the classical charge density and that resulting from the quantization is shown for comparison. It may be seen here that the quantization actually will decrease the total gate capacitance as it moves the surface charge away from the interface, producing an effective interface quantum capacitance contribution to the overall gate capacitance (in series with the normal gate capacitance to reduce the overall capacitance). In small transistors, this effect can be a significant modification to the gate capacitance, and hence to the transistor performance.

2.7 COUPLED POTENTIAL WELLS

What if there are two closely coupled potential wells? By closely coupled, it is meant that these two wells are separated by a barrier, as indicated in figure 2.11.

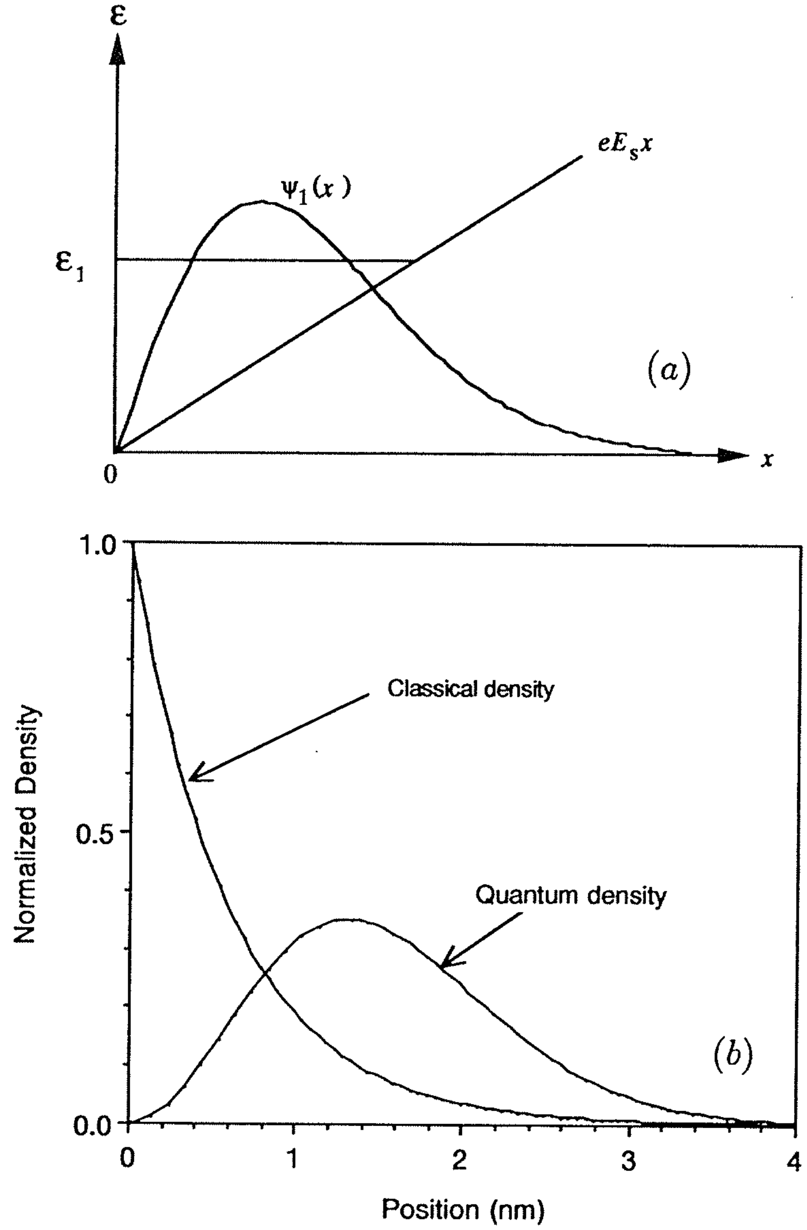


Figure 2.10 (*a*) The triangular potential well, the lowest energy level, and the Airy function wave function. (*b*) A comparison of the classical and quantum charge distributions.

However, the barrier is sufficiently thin that the decaying wave functions reach completely through the barrier into the next well. This will be quite important in the next chapter, but here we want to look at the interference that arises between the wave functions in the two wells. To simplify the problem, we will assume that the potential is infinite outside the two wells, zero in the wells, and a finite value between the wells; for example

$$V(x) = \begin{cases} \infty & |x| > b/2 + a \\ 0 & a + b/2 > |x| > b/2 \\ V_0 & |x| < b/2 . \end{cases} \tag{2.82}$$

(Note that the well width here is given by a, while it was $2a$ in the preceding sections on quantum wells.) Within the wells, the wave function is given by a sum of propagating waves, one moving to the right and one moving to the left,

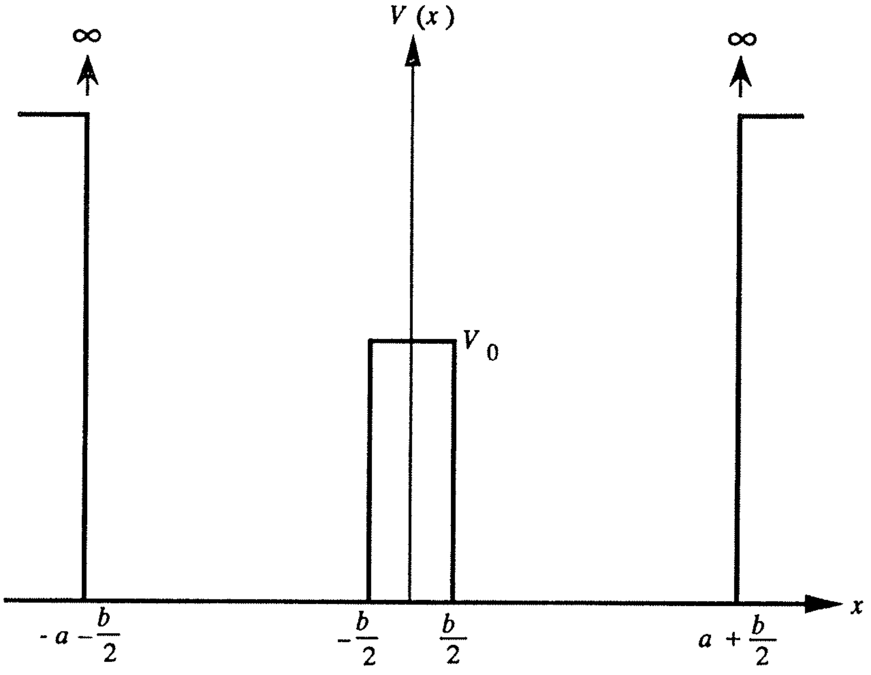


Figure 2.11 The double-well potential.

while within the barrier (where $\mathcal{E} < V_0$) the wave function is a set of decaying waves representing these same two motions. This leads to six coefficients, two of which are evaluated for $|x| = a + b/2$. The remaining four are evaluated by invoking the continuity of the wave function and its derivative at the two interfaces between the wells and the barrier, $|x| = b/2$.

We will treat only the case where the energy lies below the top of the barrier in this section. The above boundary conditions lead to a 4 × 4 matrix for the remaining coefficients. The determinant of this matrix gives the allowed energy levels. This determinantal equation is found to give real and imaginary parts

$$\tanh(\gamma b)[1 - \cos(2ka)] + \frac{k}{\gamma}\sin(2ka) = 0 \tag{2.83a}$$

and

$$\tanh(\gamma b)[1 + \cos(2ka)] + \frac{\gamma}{k}\sin(2ka) = 0 \tag{2.83b}$$

respectively. For a large potential barrier, the solution is found from the real equation (which also satisfies the imaginary one in the limit where γ goes to infinity) to be

$$\sin(ka) = 0 \quad \text{or} \quad ka = n\pi \tag{2.84}$$

which is the same result as for the infinite potential well found earlier. For a vanishing barrier, the result is the same with $a \to 2a$. Thus, the results from (2.83) satisfy two limiting cases that can be found from the infinite potential well. Our interest here is in finding the result for a weak interaction between the two wells. To solve for the general case, we will assume that the barrier is very large, and expand the hyperbolic tangent function around its value of unity for the very large limit. In addition, we expand $\cos(2ka)$ in (2.83a) about its relevant zero, where the latter is given by (2.84). This then leads to the approximate solutions

$$\sin(ka) = \pm|ka - n\pi|(1 - 2e^{-2\gamma b}). \qquad (2.85)$$

The pre-factor is very near zero, and the hyperbolic tangent function is very nearly unity, so there is a small shift of the energy level *both up and down from the bound state* of the single well. The lower level must be the symmetric combination of the wave functions of the two individual wells, which is the symmetric combination of wave functions that are each symmetric in their own wells. This lower level must be the symmetric combination since we have already ascertained that the lowest energy state is a symmetric wave function for a symmetric potential. The upper level must then be the anti-symmetric combination of the two symmetric wave functions. The actual levels from (2.85) can be found also by expanding the sine function around the zero point to give approximately

$$ka = n\pi \pm 2\sqrt{3}\,e^{-\gamma b}. \qquad (2.86)$$

While this result is for the approximation of a nearly infinite well, the general behaviour for finite wells is the same. The two bound states, one in each well that would normally lie at the same energy level, split due to the interaction of the wave functions penetrating the barrier. This leads to one level (in both wells) lying at a slightly lower energy due to the symmetric sum of the individual wave functions, and a second level lying at a slightly higher energy due to the anti-symmetric sum of the two wave functions. We will return to this in a later chapter, where we will develop formal approximation schemes to find the energy levels more exactly. In figure 2.12, experimental data on quantum wells in the GaAs/AlGaAs heterojunction system are shown to illustrate this splitting of the energy levels (Dingle *et al* 1975). Here, the coupling is quite strong, and the resulting splitting is rather large.

2.8 THE TIME VARIATION AGAIN

In each of the cases treated above, the wave function has been determined to be one of a number of possible *eigenfunctions*, each of which corresponds to a single energy level, determined by the eigenvalue. The general solution of the problem is composed of a sum over these eigenfunctions, with coefficients determined by the probability of the occupancy of each of the discrete states. This sum can be written as

$$\Psi(x) = \sum_n c_n \psi_n(x). \qquad (2.87)$$

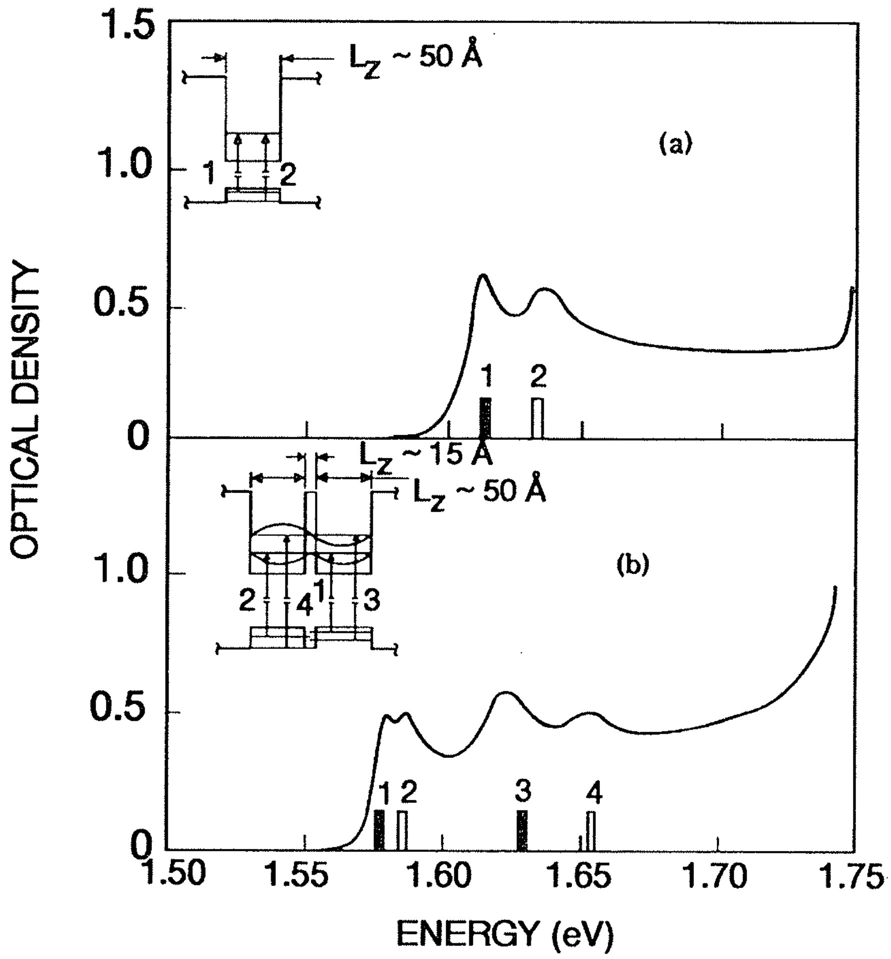


Figure 2.12 Optical absorption spectrum of a series of (*a*) eighty isolated GaAs quantum wells, of 5 nm thickness, separated by 18 nm of AlGaAs. In (*b*), the data are for sixty pairs of similar wells separated by a 1.5 nm barrier. The two bound states of the isolated wells each split into two combination levels in the double wells. (After Dingle *et al* (1975), by permission.)

In every sense, this series is strongly related to the Fourier series, where the expansion basis functions, our eigenfunctions, are determined by the geometry of the potential structure where the solution is sought. This still needs to be connected with the time-dependent solution. This is achieved by recalling that the separation coefficient that arose when the time variation was separated out from the total solution was the energy. Since the energy is different for each of the eigenfunctions, the particular energy of that function must be used, which means that the energy exponential goes inside the summation over the states. This gives

$$\Psi(x,t) = \sum_n c_n \psi_n(x) \exp\left[-\frac{\mathrm{i}\mathcal{E}_n t}{\hbar}\right]. \tag{2.88}$$

The exponential is, of course, just $\mathrm{e}^{-\mathrm{i}\omega_n t}$, the frequency variation of the particular 'mode' that is described by the corresponding eigenfunction. In many cases, the energy can be a continuous function, as in the transmission over the top of the potential well. In this case, the use of a discrete n is not appropriate. For the continuous-energy-spectrum case, it is more usual to use the energy itself as the 'index.' This is called the *energy representation.*

2.8.1 The Ehrenfest theorem

Let us now return to the concept of the expectation value. We recall that the expectation value of the position is found from

$$\langle x\rangle = (\Psi, x\Psi) = \int_{-\infty}^{\infty} \Psi^*(x,t)x\Psi(x,t)\,\mathrm{d}x. \tag{2.89}$$

What is the time variation of the position? Here, we do not refer specifically to the momentum or velocity *operator*, but to the time derivative of the *expectation value of the position.* These are two different things. In the first case, we are interested in the expectation value of the momentum operator. In the second, we are interested in whether the expectation value of the position may be changing. As stated, the problem is to determine whether the time derivative of the position is indeed the expectation value of the momentum.

Consider, for example, the situations discussed above for the various bound states for the potential wells, say the infinite well or the triangular well, where all states are bound. The time derivative of (2.86) is given by (it is assumed that the position operator is one of a set of conjugate variables and does not have an intrinsic time variation)

$$\begin{aligned}\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\int x\rho\,\mathrm{d}x = \int x\frac{\partial\rho}{\partial t}\,\mathrm{d}x = -\int x\frac{\partial J}{\partial x}\,\mathrm{d}x\\ &= -\int\frac{\partial xJ}{\partial x}\,\mathrm{d}x + \int J\frac{\partial x}{\partial x}\,\mathrm{d}x = \int J\,\mathrm{d}x.\end{aligned} \tag{2.90}$$

The continuity equation has been used to get the last term on the first line from the previous one. For the states in the wells considered, the first term in the second line vanishes since the wave function itself vanishes exponentially at the large-x limits. Since these states are not current-carrying states, the current J also vanishes and the time derivative of the position expectation vanishes. By not being a current-carrying state, we mean that the bound states are real, and so the current (2.14) is identically zero. This is not the case for the propagating wave solutions that exist above the potential barriers, for example in the finite potential well.

If the current does not vanish, as in the propagating waves, then the last term in (2.90) is identically the expectation value of the momentum. If (2.14) is used in (2.90), we find (in vector notation and with volume integrations)

$$\begin{aligned}\frac{\mathrm{d}\langle \boldsymbol{x}\rangle}{\mathrm{d}t} &= \int \boldsymbol{J}\,\mathrm{d}\boldsymbol{x} = \frac{\hbar}{2mi}\int[\Psi^*(\nabla\Psi) - (\nabla\Psi^*)\Psi]\,\mathrm{d}\boldsymbol{x}\\ &= \frac{\hbar}{mi}\int\Psi^*(\nabla\Psi)\,\mathrm{d}\boldsymbol{x}\end{aligned} \tag{2.91}$$

where we have used $(\nabla\Psi^*)\Psi = \nabla(\Psi^*\Psi) - \Psi^*(\nabla\Psi)$. The last term expresses the desired result that the time derivative of the expectation value of the position is given by the expectation value of the momentum. The important point here is that we are working with expectation values and not with the operators themselves. The connection between position and momentum in classical mechanics carries over to a connection between their expectation values in quantum mechanics.

How does this carry over to Newton's law on acceleration? In the beginning, this was one of the points that we wanted to establish—that the classical equations of motion carried over to equivalent ones in quantum mechanics. To express this result, let us seek the time derivative of the expectation value of the momentum:

$$\begin{aligned}\frac{\mathrm{d}\langle p\rangle}{\mathrm{d}t} &= -\mathrm{i}\hbar\frac{\partial}{\partial t}\int \Psi^*(\nabla\Psi)\,\mathrm{d}x \\ &= \int\left(-\mathrm{i}\hbar\frac{\partial\Psi^*}{\partial t}\right)\nabla\Psi\,\mathrm{d}x - \int \Psi^*\nabla\left(\mathrm{i}\hbar\frac{\partial\Psi^*}{\partial t}\right)\mathrm{d}x \\ &= \int\left[-\frac{\hbar^2}{2m}\nabla^2\Psi^*\right]\nabla\Psi\,\mathrm{d}x - \int \Psi^*\nabla\left[-\frac{\hbar^2}{2m}\nabla^2\Psi^*\right]\mathrm{d}x \\ &\quad + \int[(V\Psi^*)\nabla\Psi - \Psi^*\nabla(V\Psi)]\,\mathrm{d}x. \qquad (2.92)\end{aligned}$$

The first two terms in the last line can be combined and converted to a surface integral which vanishes. This follows since the momentum operator has real eigenvalues and is a Hermitian operator, and thus is self-adjoint. These two terms may be expressed as

$$\begin{aligned}(p^2\Psi, p\Psi) - (\Psi, p^3\Psi) &= (\Psi, (p^+)^2 p\Psi) - (\Psi, p^3\Psi) \\ &= (\Psi, p^3\Psi) - (\Psi, p^3\Psi) = 0. \qquad (2.93)\end{aligned}$$

The last term just becomes the gradient of the potential, and

$$\frac{\mathrm{d}\langle p\rangle}{\mathrm{d}t} = -\langle\nabla V(x)\rangle. \qquad (2.94)$$

Thus, the time derivative of the momentum is given by the expectation value of the gradient of the potential. This is a very interesting result, since it says that rapid variations in the potential will be smoothed out by the wave function itself, and it is only those variations that are of longer range and survive the averaging process that give rise to the acceleration evident in the expectation value of the momentum. This result is known as Ehrenfest's theorem.

2.8.2 Propagators and Green's functions

Equation (2.88), which we developed earlier, clearly indicates that the wave function can easily be obtained by an expansion in the basis functions appropriate to the problem at hand. It goes further, however, and even allows us to determine fully the time variation of any given initial wave function. This follows from the Schrödinger equation being a linear differential equation, with the time evolution deriving from a single initial state. To see formally how this occurs, consider a case where we know the wave function at $t = 0$ to be $\Psi(x, 0)$. This can be used with (2.88) to determine the coefficients in the generalized Fourier series, which this latter equation represents as

$$c_n = \int \psi_n^*(x)\Psi(x, 0)\,\mathrm{d}x. \tag{2.95}$$

This can be re-inserted into (2.88) to give the general solution

$$\begin{aligned}\Psi(x, t) &= \sum_n \int \psi_n^*(x')\psi_n(x)\Psi(x', 0)\exp\left[-\frac{\mathrm{i}\mathcal{E}_n t}{\hbar}\right]\mathrm{d}x' \\ &= \int K(x, x'; t, 0)\Psi(x', 0)\,\mathrm{d}x'\end{aligned} \tag{2.96}$$

where the *propagator kernel* is

$$K(x, x'; t, 0) = \sum_n \psi_n^*(x')\psi_n(x)\exp\left[-\frac{\mathrm{i}\mathcal{E}_n t}{\hbar}\right]. \tag{2.97}$$

The kernel (2.97) describes the general propagation of any initial wave function to any time $t > 0$. In fact, however, this is not required, and we could set the problem up with any initial state at any time t_0. For example, say that we know that the wave function is given by $\Psi(x, t_0)$ at time t_0. Then, the Fourier coefficients are found to be

$$c_n = \int \psi_n^*(x)\Psi(x, t_0)\exp\left[\frac{\mathrm{i}\mathcal{E}_n t_0}{\hbar}\right]\mathrm{d}x. \tag{2.98}$$

Following the same procedure—that is, re-introducing this into (2.85)—the general solution at arbitrary time for the wave function is then

$$\Psi(x, t) = \int K(x, x'; t, t_0)\Psi(x, t_0)\,\mathrm{d}x' \tag{2.99}$$

where

$$K(x, x'; t, t_0) = \sum_n \psi_n^*(x')y_n(x)\exp\left[-\frac{\mathrm{i}\mathcal{E}_n(t - t_0)}{\hbar}\right]. \tag{2.100}$$

We note that the solution is a function of $t - t_0$, and not a function of these two times separately. This is a general property of the linear solutions, and is always expected (unless for some reason the basis set is changing with time).

The interesting fact about (2.99) is that we can find the solutions either for $t > t_0$, or for $t < t_0$. This means that we can propagate forward in time to find the future solution, or we can propagate backward in time to find the earlier state that produced the wave function at t_0.

In general, it is preferable to separate the propagation in forward and reverse times to obtain different functions for retarded behaviour (forward in time) and for advanced behaviour (backward in time). We can do this by introducing the retarded Green's function as

$$G_{\mathrm{r}}(x, x'; t, t_0) = -\mathrm{i}\Theta(t - t_0)K(x, x'; t, t_0) \tag{2.101}$$

where Θ is the Heaviside function. Hence, the retarded Green's function vanishes by construction for $t < t_0$. Similarly, the advanced Green's function can be defined as

$$G_{\mathrm{a}}(x, x'; t, t_0) = \mathrm{i}\Theta(t_0 - t)K(x, x'; t, t_0) \tag{2.102}$$

which vanishes by construction for $t > t_0$. These can be put together to give

$$K(x, x'; t, t_0) = \mathrm{i}[G_{\mathrm{r}}(x, x'; t, t_0) - G_{\mathrm{a}}(x, x'; t, t_0)]. \tag{2.103}$$

We can compute the kernel from the general Schrödinger equation itself. To see this, note that when $t = t_0$, equation (2.100) becomes just a sum over a complete set of basis states, and a property of these orthonormal functions is that

$$K(x, x'; t_0, t_0) = \delta(x - x') \tag{2.104}$$

which is expected just by investigating (2.97). This suggests that we can develop a differential equation for the kernel, which has the unique initial condition (2.104). This is done by beginning with the time derivative, as (for a free wave propagation, $V = 0$)

$$\begin{aligned}\frac{\partial K}{\partial t} &= -\sum_n \psi_n^*(x')\psi_n(x)\left[\frac{\mathrm{i}\mathcal{E}_n}{\hbar}\right]\exp\left[-\frac{\mathrm{i}\mathcal{E}_n(t - t_0)}{\hbar}\right] \\ &= -\frac{\mathrm{i}}{\hbar}\sum_n \psi_n^*(x')[H\psi_n(x)]\exp\left[-\frac{\mathrm{i}\mathcal{E}_n(t - t_0)}{\hbar}\right] \\ &= -\frac{\mathrm{i}}{\hbar}HK \end{aligned} \tag{2.105}$$

or

$$\mathrm{i}\hbar\frac{\partial K}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2 K}{\partial x^2}. \tag{2.106}$$

The easiest method for solving this equation is to Laplace transform in time, and then solve the resulting second-order differential equation in space with the initial boundary condition (2.104) and vanishing of K at large distances. This leads to (we take t_0 as zero for convenience)

$$K(x, x', t) = \sqrt{\frac{m}{2\pi\hbar t}} \exp\left[-\frac{m(x-x')^2}{2\mathrm{i}\hbar t}\right]. \tag{2.107}$$

It may readily be ascertained that this satisfies the condition (2.104) at $t = 0$.

The definition of the kernel (2.100) is, in a sense, an inverse Fourier transform from a frequency space, with the frequency defined by the discrete (or continuous) energy levels. In this regard the product of the two basis functions, at different positions, gives the amplitude of each Fourier component (in time remember, as we are also dealing with generalized Fourier series in space). Another way of thinking about this is that the kernel represents a summation over the *spectral* components, and is often called the *spectral density.* In fact, if we Fourier transform (2.97) in time, the kernel is just

$$K(x, x', \omega) = \sum_n \frac{\psi_n^*(x')\psi_n(x)}{\mathrm{i}(\omega - \omega_n)} \tag{2.108}$$

where $\omega_n = \mathcal{E}_n/\hbar$. It is clear that the numerical factors included in the definition of the Green's functions convert the denominator to energy and cancel the factor i. The difference between the retarded and advanced Green's functions lies in the way in which the contour of the inverse transform is closed, and it is typical to add a convergence factor η as in

$$G(x, x', \omega) = \sum_n \frac{\psi_n^*(x')\psi_n(x)}{(\omega_n - \omega \pm \mathrm{i}\eta)} \tag{2.109}$$

where the upper sign is used for the retarded function and the lower sign is used for the advanced function.

REFERENCES

Abramowitz M and Stegun I A 1964 *Handbook of Mathematical Functions* (Washington, DC: US National Bureau of Standards) p 450

Dingle R, Gossard A C and Wiegman W 1975 *Phys. Rev. Lett.* **34** 1327–30

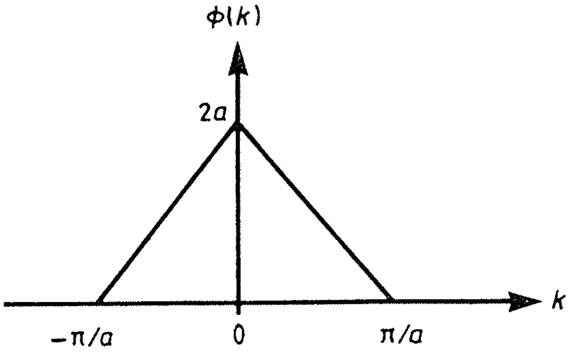
Dingle R, Wiegman W and Henry C H 1974 *Phys. Rev. Lett.* **33** 827–30

Heiblum M, Fischetti M V, Dumke W P, Frank D J, Anderson I M, Knoedler C M and Osterling L 1987 *Phys. Rev. Lett.* **58** 816–9

Schrödinger E 1926 *Ann. Phys., Lpz.* **79** 361, **79** 489, **81** 109

PROBLEMS

1. For the wave packet defined by $\phi(k)$, shown below, find $\Psi(x)$. What are Δx and Δk?



2. If a Gaussian wave packet approaches a potential step ($V > 0$ for $x > 0$, $k_0 > 0$), it is found that it becomes broader for the region $x > 0$. Why?

3. Assume that $\psi_n(x)$ are the eigenfunctions in an infinite square well ($V \to \infty$ for $|x| > d/2$). Calculate the overlap integrals

$$\int_{-d/2}^{d/2} \psi_n(x)\psi_m(x)\,\mathrm{d}x.$$

4. Suppose that electrons are confined in an infinite potential well of width 0.5 nm. What spectral frequencies will result from transitions between the lowest four energy levels? Use the free-electron mass in your computations.

5. A particle confined to an infinite potential well has an uncertainty that is of the order of the well width, $\Delta x \simeq a$. The momentum can be estimated as its uncertainty value as well. Using these simple assumptions, estimate the energy of the lowest level. Compare with the actual value.

6. In terms of the momentum operator $\boldsymbol{p} = -\mathrm{i}\hbar\boldsymbol{\nabla}$, and

$$H = \frac{p^2}{2m} + \frac{m\omega^2}{2}x^2$$

and using the fact that $\langle p\rangle = \langle x\rangle = 0$ in a bound state, with

$$\langle p^2\rangle = (\Delta p)^2 + \langle p\rangle^2 = (\Delta p)^2$$
$$\langle x^2\rangle = (\Delta x)^2 + \langle x\rangle^2 = (\Delta x)^2$$

use the uncertainty principle to estimate the lowest bound-state energy. (Hint: recall the classical relation between the average kinetic and potential energies.)

7. Consider a potential well with $V = -0.3$ eV for $|x| < a/2$, and $V = 0$ for $|x| > a/2$, with $a = 7.5$ nm. Write a computer program that computes the energy levels for $\mathcal{E} < 0$ (use a mass appropriate for GaAs, $m \simeq 6.0 \times 10^{-32}$ kg). How many levels are bound in the well, and what are their energy eigenvalues? Using a simple wave-function-matching technique, plot the wave functions for each bound state. Plot the transmission coefficient for $\mathcal{E} > 0$.

8. For the situation in which a linear potential is imposed on a system, compute the momentum wave functions. Show that these wave functions form a normalized set.

9. Using the continuity of the wave function and its derivative at each interior interface, verify (2.83).

10. Consider an infinite potential well that is 10 nm wide. At time zero, a Gaussian wave packet, with half-width of 1 nm, is placed 2 nm from the centre of the well. Plot the evolving wave functions for several times up to the stable steady state. How does the steady state differ from the initial state, and why does this occur?

11. Verify that (2.107) is the proper solution for the kernel function.

3

Tunnelling

When we dealt in the last chapter (section 2.7) with the double potential well, coupled through a thin barrier, it was observed that the wave function penetrated through the barrier and interacted with the wave function in the opposite well. This process does not occur in classical mechanics, since a particle will in all cases bounce off the barrier. However, when we treat the particle as a wave, then the wave nature of barrier penetration can occur. This is familiar in electromagnetic waves, where the decaying wave (as opposed to a propagating wave) is termed an *evanescent* wave. For energies below the top of the barrier, the wave is attenuated, and it decays exponentially. Yet, it takes a significant distance for this decay to eliminate the wave completely. If the barrier is thinner than this critical distance, the evanescent wave can excite a propagating wave in the region beyond the barrier. Thus, the wave can penetrate the barrier, and continue to propagate, with an attenuated amplitude, in the trans-barrier region. This process is termed *tunnelling*, with analogy to the miners who burrow through a mountain in order to get to the other side! This process is quite important in modern semiconductor devices, and Leo Esaki received the Nobel prize for first recognizing that tunnelling was important in degenerately doped p–n junction diodes. In this chapter, we will address this tunnelling process. First we will treat those few cases in which the tunnelling probability can be obtained exactly. Then we will discuss its use in solid-state electronics. Following this, we will move to approximate treatments suitable for those cases in which the solution is not readily obtainable in an exact manner. Finally, we turn to periodic tunnelling structures, which give rise for example to the band structure discussed in semiconductors.

3.1 THE TUNNEL BARRIER

The general problem is that posed in figure 3.1. Here, we have a barrier, whose height is taken to be V_0, that exists in the region $|x| < a$. To the left and to the right of this barrier, the particle can exist as a freely propagating wave, but, in the region $|x| < a$, and for energies $\mathcal{E} < V_0$, the wave is heavily

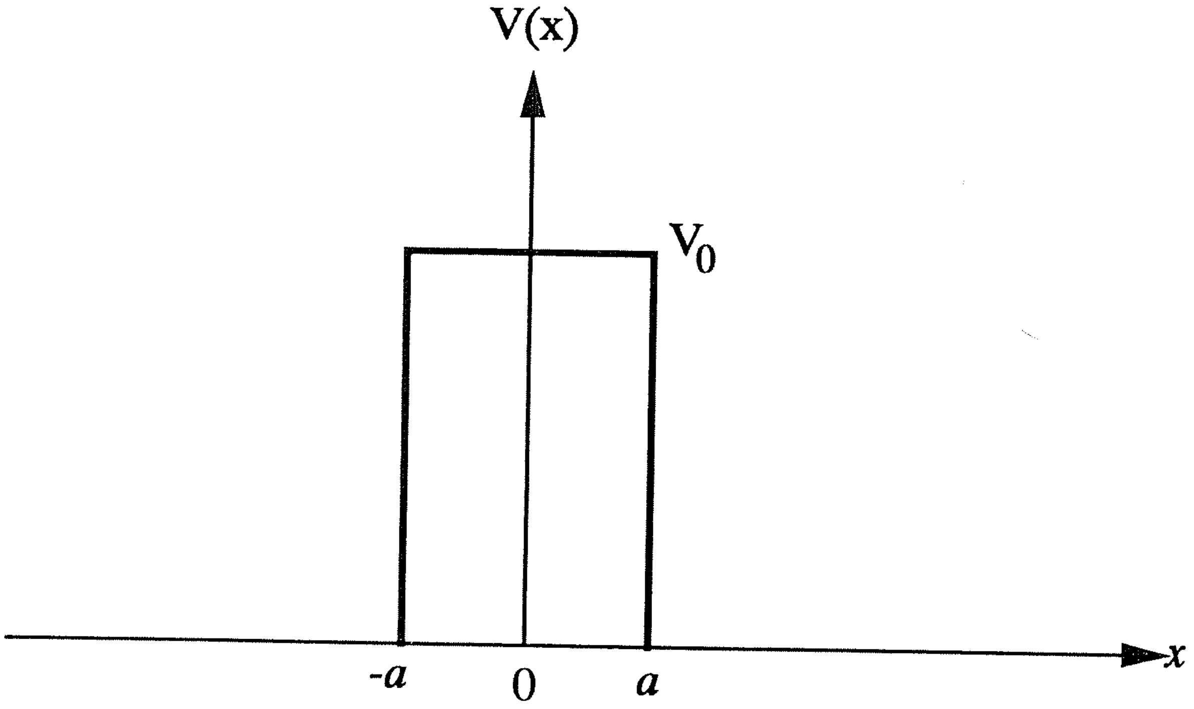


Figure 3.1 The simple rectangular tunnelling barrier.

attenuated and is characterized by a decaying exponential 'wave'. Our interest is in determining just what the transmission probability through the barrier is for an incident particle. We are also interested in the transmission behaviour for energies above the top of the barrier. To solve for these factors, we proceed in precisely the same fashion as we did for the examples of the last chapter. That is, we assume waves with the appropriate propagation characteristics in each of the regions of interest, but with unknown coefficients. We then apply boundary conditions, in this case the continuity of the wave function and its derivative at each interface, in order to evaluate the unknown coefficients. We consider first a simple barrier.

3.1.1 The simple rectangular barrier

The simple barrier is shown in figure 3.1. Here the potential is defined to exist only between $-a$ and a, and the zero of potential for the propagating waves on either side is the same. We can therefore define the wave vector k in the region $|x| > a$, and the decaying wave vector γ in the region $|x| < a$, by the equations ($\mathcal{E} < V_0$)

$$k = \sqrt{\frac{2m}{\hbar^2}\mathcal{E}} \qquad \gamma = \sqrt{\frac{2m}{\hbar^2}(V_0 - \mathcal{E})} \tag{3.1}$$

respectively. To the right and left of the barrier, the wave is described by propagating waves, while in the barrier region, the wave is attenuated. Thus, we can write the wave function quite generally as

$$\Psi(x) = \begin{cases} A\mathrm{e}^{\mathrm{i}kx} + B\mathrm{e}^{-\mathrm{i}kx} & x < -a \\ C\mathrm{e}^{\gamma x} + D\mathrm{e}^{-\gamma x} & |x| < a \\ E\mathrm{e}^{\mathrm{i}kx} + F\mathrm{e}^{-\mathrm{i}kx} & x > a\,. \end{cases} \tag{3.2}$$

We now have six unknown coefficients to evaluate. However, we can get only four equations from the two boundary conditions, and a fifth from normalizing the incoming wave from one side or the other. If we keep the most general set of six coefficients, we will have incoming waves from both sides, both of which must be normalized in some fashion. For our purposes, however, we will throw away the incoming wave from the right, and assume that our interest is in determining the transmission of a wave incident from the left. In a later section, though, we will need to keep both solutions, as we will have multiple barriers with multiple reflections. Here, however, while we keep all the coefficients, we will eventually set $F = 0$. We can count on eventually using the principle of superposition, as the Schrödinger equation is linear; thus, our approach is perfectly general.

The boundary conditions are applied by asserting continuity of the wave function and its derivative at each interface. Thus, at the interface $x = -a$, continuity of these two quantities gives rise to

$$A\mathrm{e}^{-\mathrm{i}ka} + B\mathrm{e}^{\mathrm{i}ka} = C\mathrm{e}^{-\gamma a} + D\mathrm{e}^{\gamma a} \tag{3.3a}$$

$$\mathrm{i}k\left[A\mathrm{e}^{-\mathrm{i}ka} - B\mathrm{e}^{\mathrm{i}ka}\right] = \gamma\left[C\mathrm{e}^{-\gamma a} - D\mathrm{e}^{\gamma a}\right]. \tag{3.3b}$$

As in the last chapter, we can now solve for two of these coefficients in terms of the other two coefficients. For the moment, we seek A and B in terms of C and D. This leads to the matrix equation

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \left(\dfrac{\mathrm{i}k+\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{(\mathrm{i}k-\gamma)a} & \left(\dfrac{\mathrm{i}k-\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{(\mathrm{i}k+\gamma)a} \\ \left(\dfrac{\mathrm{i}k-\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{-(\mathrm{i}k+\gamma)a} & \left(\dfrac{\mathrm{i}k+\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{-(\mathrm{i}k-\gamma)a} \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix}. \tag{3.4}$$

Now, we turn to the other boundary interface. The continuity of the wave function and its derivative at $x = a$ leads to

$$E\mathrm{e}^{\mathrm{i}ka} + F\mathrm{e}^{-\mathrm{i}ka} = C\mathrm{e}^{\gamma a} + D\mathrm{e}^{-\gamma a} \tag{3.5a}$$

$$\mathrm{i}k[E\mathrm{e}^{\mathrm{i}ka} - F\mathrm{e}^{-\mathrm{i}ka}] = \gamma[C\mathrm{e}^{\gamma a} - D\mathrm{e}^{-\gamma a}]. \tag{3.5b}$$

Again, we can solve for two of these coefficients in terms of the other two. Here, we seek to find C and D in terms of E and F (we will eliminate the former two through the use of (3.4)). This leads to the matrix equation

$$\begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} \left(\dfrac{\mathrm{i}k+\gamma}{2\gamma}\right)\mathrm{e}^{(\mathrm{i}k-\gamma)a} & -\left(\dfrac{\mathrm{i}k-\gamma}{2\gamma}\right)\mathrm{e}^{-(\mathrm{i}k+\gamma)a} \\ -\left(\dfrac{\mathrm{i}k-\gamma}{2\gamma}\right)\mathrm{e}^{(\mathrm{i}k+\gamma)a} & \left(\dfrac{\mathrm{i}k+\gamma}{2\gamma}\right)\mathrm{e}^{-(\mathrm{i}k-\gamma)a} \end{bmatrix} \begin{bmatrix} E \\ F \end{bmatrix}. \tag{3.6}$$

From the pair of equations (3.4) and (3.6), the two propagating coefficients on the left of the barrier, A and B, can be related directly to those on the right of the barrier, E and F, with the two under the barrier dropping out of consideration. This leads to the matrix equation

$$\left.\begin{matrix}A\\B\end{matrix}\right]\begin{bmatrix}M_{11} & M_{12}\\M_{21} & M_{22}\end{bmatrix}=\left.\begin{matrix}E\\F\end{matrix}\right]. \tag{3.7}$$

Here, the elements are defined by the relations

$$\begin{aligned}M_{11} &= \left(\frac{\mathrm{i}k+\gamma}{2\mathrm{i}k}\right)\left(\frac{\mathrm{i}k+\gamma}{2\gamma}\right)\mathrm{e}^{2(\mathrm{i}k-\gamma)a}-\left(\frac{\mathrm{i}k-\gamma}{2\gamma}\right)\left(\frac{\mathrm{i}k-\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{2(\mathrm{i}k+\gamma)a}\\ &= \left[\cosh(2\gamma a)-\frac{\mathrm{i}}{2}\left(\frac{k^2-\gamma^2}{k\gamma}\right)\sinh(2\gamma a)\right]\mathrm{e}^{2\mathrm{i}ka}\end{aligned} \tag{3.8}$$

$$\begin{aligned}M_{21} &= \left(\frac{\mathrm{i}k+\gamma}{2\gamma}\right)\left(\frac{\mathrm{i}k-\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{-2\gamma a}-\left(\frac{\mathrm{i}k+\gamma}{2\mathrm{i}k}\right)\left(\frac{\mathrm{i}k-\gamma}{2\gamma}\right)\mathrm{e}^{2\gamma a}\\ &= -\frac{\mathrm{i}}{2}\left(\frac{k^2+\gamma^2}{k\gamma}\right)\sinh(2\gamma a)\end{aligned} \tag{3.9}$$

$$M_{22}=M_{11}^{*} \qquad M_{12}=M_{21}^{*}. \tag{3.10}$$

It is a simple algebraic exercise to show that, for the present case, the determinant of the matrix **M** is unity, so this matrix has quite interesting properties. It is *not* a unitary matrix, because the diagonal elements are complex. In the simple case, where we will take $F=0$, the transmission coefficient is simply given by the reciprocal of $|M_{11}|^2$, since the momentum is the same on either side of the barrier and hence the current does not involve any momentum adjustments on the two sides.

3.1.2 The tunnelling probability

In the formulation that leads to (3.7), A and F are incoming waves, while B and E are outgoing waves. Since we are interested in the tunnelling of a particle from one side to the other, we treat an incoming wave from only one of the two sides, so that we will set $F=0$ for this purpose. Then, we find that $A=M_{11}E$. The transmission probability is the ratio of the currents on the two sides of the barrier, directed in the same direction of course, so

$$T=\frac{1}{|M_{11}|^2}. \tag{3.11}$$

Inserting the value for this from (3.8), we find

$$\begin{aligned}T &= \left[\cosh^2 2\gamma a)+\left(\frac{k^2-\gamma^2}{2k\gamma}\right)^2\sinh^2(2\gamma a)\right]^{-1}\\ &= \frac{1}{1+\left(\frac{k^2+\gamma^2}{2k\gamma}\right)^2\sinh^2(2\gamma a)}.\end{aligned} \tag{3.12}$$

There are a number of limiting cases that are of interest. First, for a very weak barrier, in which $2\gamma a \ll 1$, the transmission coefficient becomes

$$T \to \frac{1}{1+(ka)^2}. \tag{3.13}$$

On the other hand, when the potential is very strong, where $2\gamma a \gg 1$, the transmission coefficient falls off exponentially as

$$T \to \left(\frac{4k\gamma}{k^2+\gamma^2}\right)^2 \mathrm{e}^{-4\gamma a}. \tag{3.14}$$

It is important to note that the result (3.13) is valid only for a weak potential for which the energy is actually *below the top of the barrier*. If we consider an incident energy above the barrier, we expect the barrier region to act as a thin dielectric and cause interference fringes. We can see this by making the simple substitution suggested by (3.1) through $\gamma \to -ik'$. This changes (3.12) into

$$T(\mathcal{E} > V_0) = \frac{1}{1+\left(\dfrac{k^2-k'^2}{2kk'}\right)^2 \sin^2(2k'a)} \tag{3.15}$$

which is precisely the result (2.70) obtained in the last chapter (with a suitable change in the definition of the wave function in the barrier region). Thus, above the barrier, the transmission has oscillatory behaviour as a function of energy, with resonances that occur for $2k'a = n\pi$. The overall behaviour of the tunnelling coefficient is shown in figure 3.2.

3.2 A MORE COMPLEX BARRIER

In the previous section, the calculations were quite simple as the wave momentum was the same on either side of the barrier. Now, we want to consider a somewhat more realistic barrier in which the momentum differs on the two sides of the barrier. Consider the barrier shown in figure 3.3. The interface at $x = -a$ is the same as treated previously, and the results of (3.4) are directly used in the present problem. However, the propagating wave on the right-hand side of the barrier ($x > a$) is characterized by a different wave vector through

$$k_1 = \sqrt{\frac{2m}{\hbar^2}(\mathcal{E}+V_1)}. \tag{3.16}$$

Matching the wave function and its derivative at $x = a$ leads to

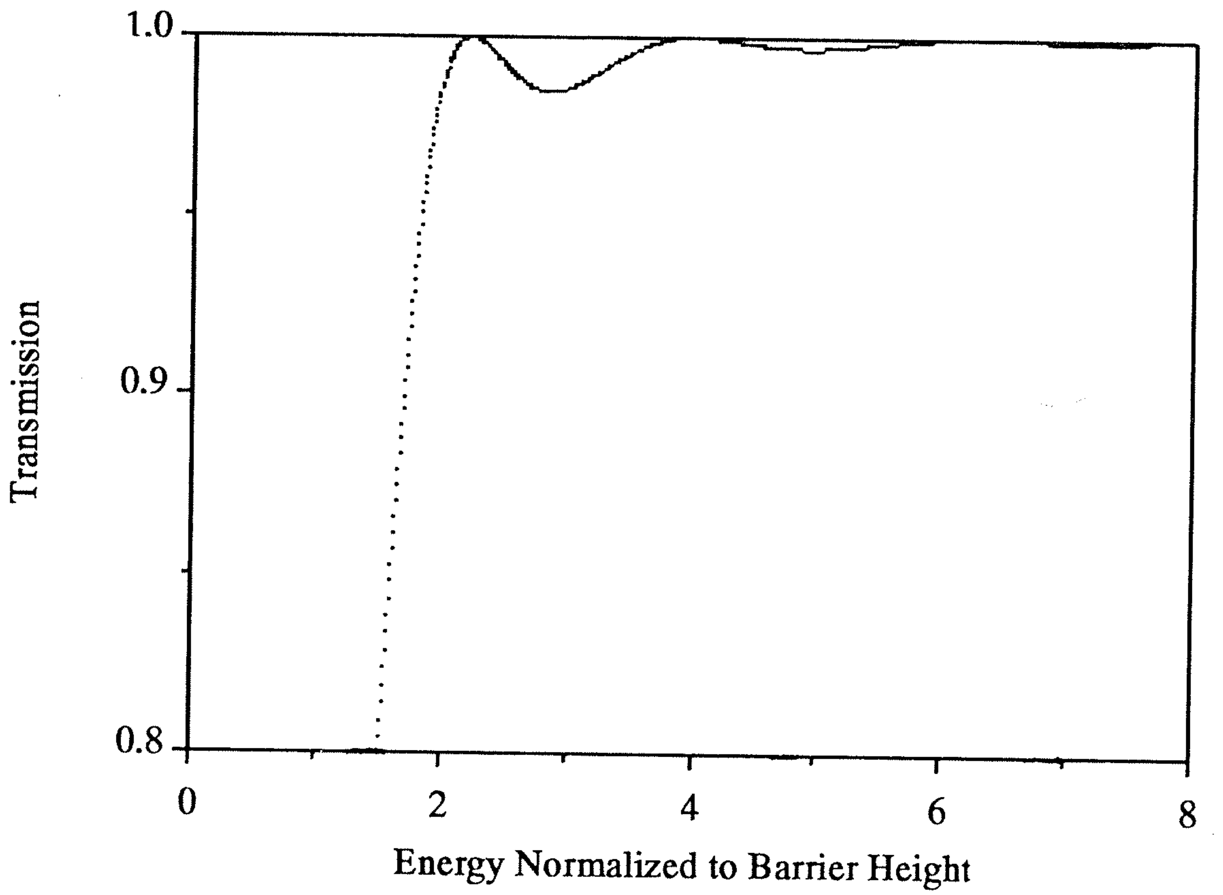


Figure 3.2 Tunnelling (transmission) probability for a simple barrier (for generic values).

$$Ee^{ik_1a} + Fe^{-ik_1a} = Ce^{\gamma a} + De^{-\gamma a} \tag{3.17a}$$

$$ik_1\left[Ee^{ik_1a} - Fe^{-ik_1a}\right] = \gamma\left[Ce^{\gamma a} - De^{-\gamma a}\right]. \tag{3.17b}$$

This result is an obvious modification of (3.5). This will also occur for the matrix equation (3.6), and the result is

$$\begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} \left(\dfrac{ik_1+\gamma}{2\gamma}\right)e^{(ik_1-\gamma)a} & -\left(\dfrac{ik_1-\gamma}{2\gamma}\right)e^{-(ik_1+\gamma)a} \\ -\left(\dfrac{ik_1-\gamma}{2\gamma}\right)e^{(ik_1+\gamma)a} & \left(\dfrac{ik_1+\gamma}{2\gamma}\right)e^{-(ik_1-\gamma)a} \end{bmatrix}\begin{bmatrix} E \\ F \end{bmatrix}. \tag{3.18}$$

We can now eliminate the coefficients C and D by combining (3.6) and (3.18). The result is again (3.7), but now the coefficients are defined by

$$M_{11} = \left(\frac{ik+\gamma}{2ik}\right)\left(\frac{ik_1+\gamma}{2\gamma}\right)e^{(ik+ik_1-2\gamma)a}$$
$$-\left(\frac{ik_1-\gamma}{2\gamma}\right)\left(\frac{ik-\gamma}{2ik}\right)e^{(ik+ik_1+2\gamma)a}$$

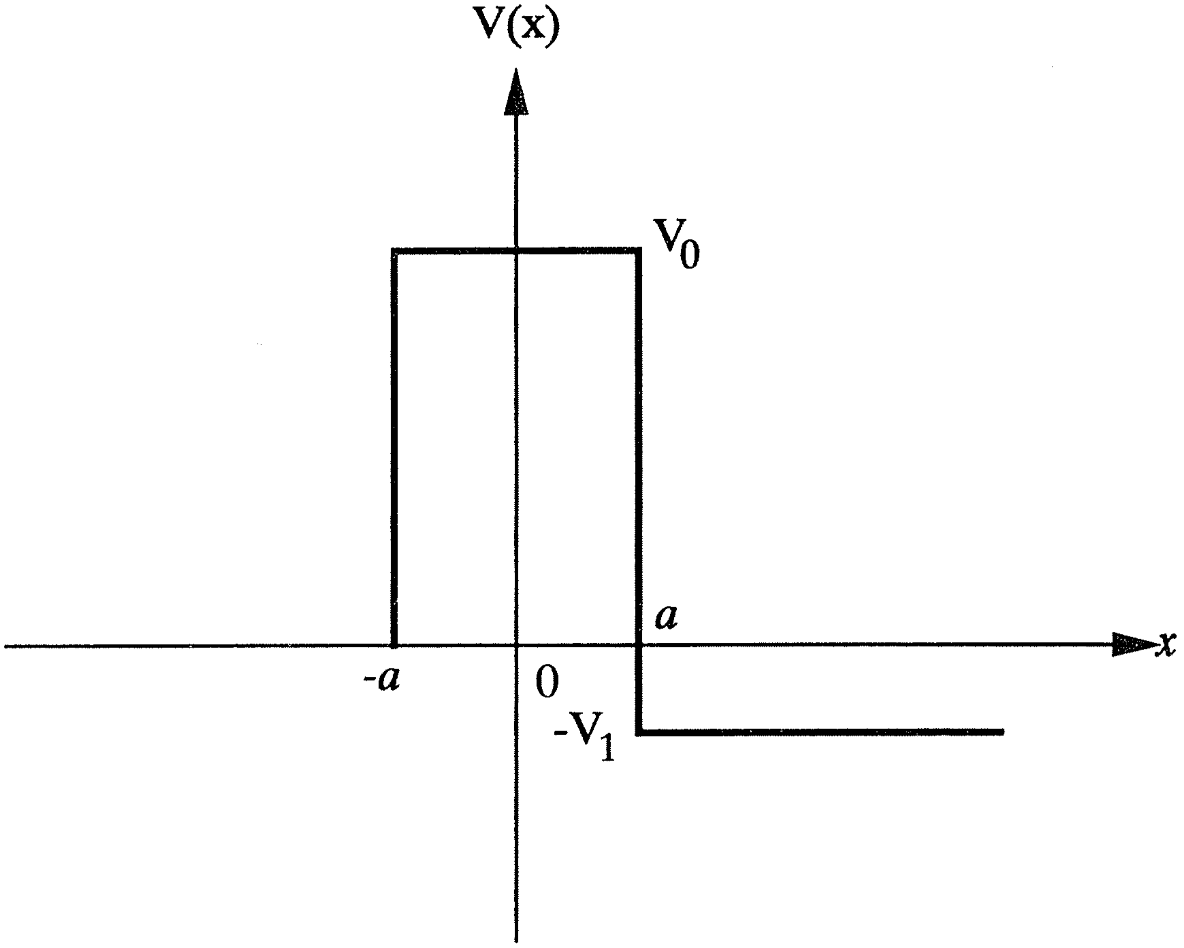


Figure 3.3 A more complex tunnelling barrier.

$$= \left[\frac{1}{2}\left(1+\frac{k_1}{k}\right)\cosh(2\gamma a) - \frac{\mathrm{i}}{2}\left(\frac{kk_1-\gamma^2}{k\gamma}\right)\sinh(2\gamma a)\right]\mathrm{e}^{\mathrm{i}(k+k_1)a} \tag{3.19}$$

$$\begin{aligned} M_{21} &= \left(\frac{\mathrm{i}k_1+\gamma}{2\gamma}\right)\left(\frac{\mathrm{i}k-\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{-2\gamma a-\mathrm{i}(k-k_1)a} \\ &\quad - \left(\frac{\mathrm{i}k+\gamma}{2\mathrm{i}k}\right)\left(\frac{\mathrm{i}k_1-\gamma}{2\gamma}\right)\mathrm{e}^{2\gamma a-\mathrm{i}(k-k_1)a} \\ &= -\left[\frac{\mathrm{i}}{2}\left(\frac{kk_1+\gamma^2}{k\gamma}\right)\sinh(2\gamma a) + \frac{1}{2}\left(\frac{k_1}{k}-1\right)\cosh(2\gamma a)\right]\mathrm{e}^{-\mathrm{i}(k-k_1)a} \end{aligned} \tag{3.20}$$

and the complex conjugate symmetry still holds for the remaining terms.

The determinant of the matrix **M** is also no longer unity, but is given by the ratio k_1/k. This determinant also reminds us that we must be careful in calculating the transmission coefficient as well, due to the differences in the momenta, at a given energy, on the two sides of the barrier. We proceed as in the previous section, and take $F = 0$ in order to compute the transmission coefficient. The actual transmission coefficient relates the currents as in (2.39)–(2.41), and we find that

$$T = \frac{k_1}{k}\frac{1}{|M_{11}|^2} = \frac{4k_1k/(k_1+k)^2}{1+\dfrac{(\gamma^2+k^2)(\gamma^2+k_1^2)}{\gamma^2(k_1+k)^2}\sinh^2(2\gamma a)}. \tag{3.21}$$

In (3.21), there are two factors. The first factor is the one in the numerator, which describes the discontinuity between the propagation constants in the two regions to the left and to the right of the barrier. The second factor is the denominator, which is the actual tunnelling coefficient describing the *transparency* of the barrier. It is these two factors together that describe the total transmission of waves from one side to the other. It should be noted that if we take the limiting case of $k_1 = k$, we recover the previous result (3.12).

There is an important relationship that is apparent in (3.21). The result represented by (3.21) is reciprocal in the two wave vectors. They appear symmetrical in the transmission coefficient T. This is a natural and important result of the symmetry. Even though the barrier and energy structure of figure 3.3 does not appear symmetrical, the barrier is a linear structure that is passive (there is no active gain in the system). Therefore, the electrical properties should satisfy the principal of reciprocity, and the transmission should be the same regardless of from which direction one approaches the barrier. This is evident in the form of the transmission coefficient (3.20) that is obtained from these calculations.

3.3 THE DOUBLE BARRIER

We now want to put together two tunnel barriers separated by a quantum well. The quantum well (that is, the region between the two barriers) will have discrete energy levels because of the confinement quantization, just as in section 2.5. We will find that, when the incident wave energy corresponds to one of these resonant energy states of the quantum well, the transmission through the double barrier will rise to a value that is unity (for equal barriers). This resonant tunnelling, in which the transmission is unity, is quite useful as an energy filter.

There are two approaches to solving for the composite tunnelling transmission coefficient. In one, we resolve the entire problem from first principles, matching the wave function and its derivative at four different interfaces (two for each of the two barriers). The second approach, which we will pursue here, uses the results of the previous sections, and we merely seek knowledge as to how to put together the transmission matrices that we already have found. The reason we can pursue this latter approach effectively is that the actual transmission matrices found in the previous sections depend only upon the wave vectors (the ks and γ), and the thickness of the barrier, $2a$. They do *not* depend upon the position of the barrier, so the barrier may be placed at an arbitrary point in space without modifying the transmission properties. Thus, we consider the generic problem of figure 3.4, where we have indicated the coefficients in the same manner as that in which they were defined in the earlier sections. To differentiate between the

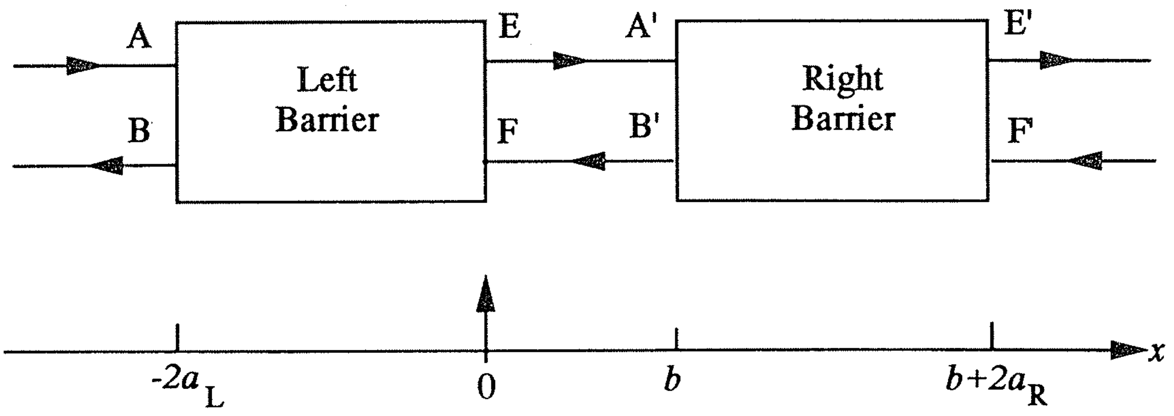


Figure 3.4 Two generic barriers are put together to form a double-barrier structure.

two barriers, we have used primes on the coefficients of the right-hand barrier. Our task is to now relate the coefficients of the left-hand barrier to those of the right-hand barrier.

We note that both E and A' describe a wave propagating to the right. Denoting the definition of the thickness of the well region as b, we can simply relate these two coefficients via

$$A' = E\mathrm{e}^{ikb} \tag{3.22}$$

where k is the propagation constant *in the well region*. Similarly, F and B' relate the same wave propagating in the opposite direction. These two can thus be related by

$$B' = F\mathrm{e}^{-ikb}. \tag{3.23}$$

These definitions now allow us to write the connection as a matrix in the following manner:

$$\begin{bmatrix} E \\ F \end{bmatrix} = \begin{bmatrix} \mathrm{e}^{ikb} & 0 \\ 0 & \mathrm{e}^{-ikb} \end{bmatrix} \begin{bmatrix} A' \\ B' \end{bmatrix}. \tag{3.24}$$

Equation (3.24) now defines a matrix $\mathbf{M}_{\mathrm{W}}$, where the subscript indicates the well region. This means that we can now take the matrices defined in sections 3.1 and 3.2 for the left-hand and right-hand regions and write the overall tunnelling matrix as

$$\begin{bmatrix} A \\ B \end{bmatrix} = [\mathbf{M}_{\mathrm{L}}]\,[\mathbf{M}_{\mathrm{W}}]\,[\mathbf{M}_{\mathrm{R}}] \begin{bmatrix} E' \\ F' \end{bmatrix}. \tag{3.25}$$

From this, it is easy to now write the composite M_{11} as

$$M_{\mathrm{T}11} = M_{\mathrm{L}11}M_{\mathrm{R}11}\mathrm{e}^{-ikb} + M_{\mathrm{L}12}M_{\mathrm{R}21}\mathrm{e}^{ikb} \tag{3.26}$$

and it is apparent that the resonance behaviour arises from the inclusion of the off-diagonal elements of each transmission matrix, weighted by the propagation factors. At this point, we need to be more specific about the individual matrix elements.

3.3.1 Simple, equal barriers

For the first case, we use the results of section 3.1, where a simple rectangular barrier was considered. Here, we assume that the two barriers are exactly equal, so the same propagation wave vector k exists in the well and in the regions to the left and right of the composite structure. By the same token, each of the two barriers has the same potential height and therefore the same γ. We note that this leads to a magnitude-squared factor in the second term of (3.26), *but not in the first term* with one notable exception. The factor of e^{i2ka} does cancel since we are to the left of the right-hand barrier ($-a$-direction) but to the right of the left-hand barrier ($+a$-direction). Thus, the right-hand barrier contributes a factor of e^{-i2ka}, and the left-hand barrier contributes a factor of e^{i2ka}, so the two cancel each other. In order to simplify the mathematical details, we write the remainder of (3.8) as

$$M_{11} = m_{11}e^{-i\theta} \tag{3.27}$$

where

$$m_{11} = \sqrt{\cosh^2(2\gamma a) + \left(\frac{k^2-\gamma^2}{2k\gamma}\right)^2 \sinh^2(2\gamma a)} \tag{3.28}$$

is the magnitude and

$$\theta = \tan^{-1}\left[\left(\frac{k^2-\gamma^2}{2k\gamma}\right)\tanh(2\gamma a)\right] \tag{3.29}$$

is the phase of M_{11}. We can then use this to write

$$\begin{aligned} |M_{T11}|^2 &= |M_{11}|^4 + |M_{21}|^4 + 2|M_{11}|^2|M_{21}|^2\cos[2(kb+\theta)] \\ &= (|M_{11}|^2 - |M_{21}|^2) + 4|M_{11}|^2|M_{21}|^2\cos^2(kb+\theta). \end{aligned} \tag{3.30}$$

The first term, the combination within the parentheses, is just the determinant of the individual barrier matrix, and is unity for the simple rectangular barrier. Thus, the overall transmission is now

$$|M_{T11}|^2 = 1 + 4|m_{11}|^2|M_{21}|^2\cos^2(kb+\theta). \tag{3.31}$$

In general, the cosine function is non-zero, and the composite term of (3.31) is actually larger than that for the single barrier T_1, with

$$T_{\text{total}} \sim \frac{T_1}{4|M_{21}|^2} \quad \text{off resonance.} \tag{3.32}$$

However, for particular values of the wave vector, the cosine term vanishes, and

$$T_{\text{total}} = 1 \qquad kb + \theta = (2n+1)\frac{\pi}{2}. \tag{3.33}$$

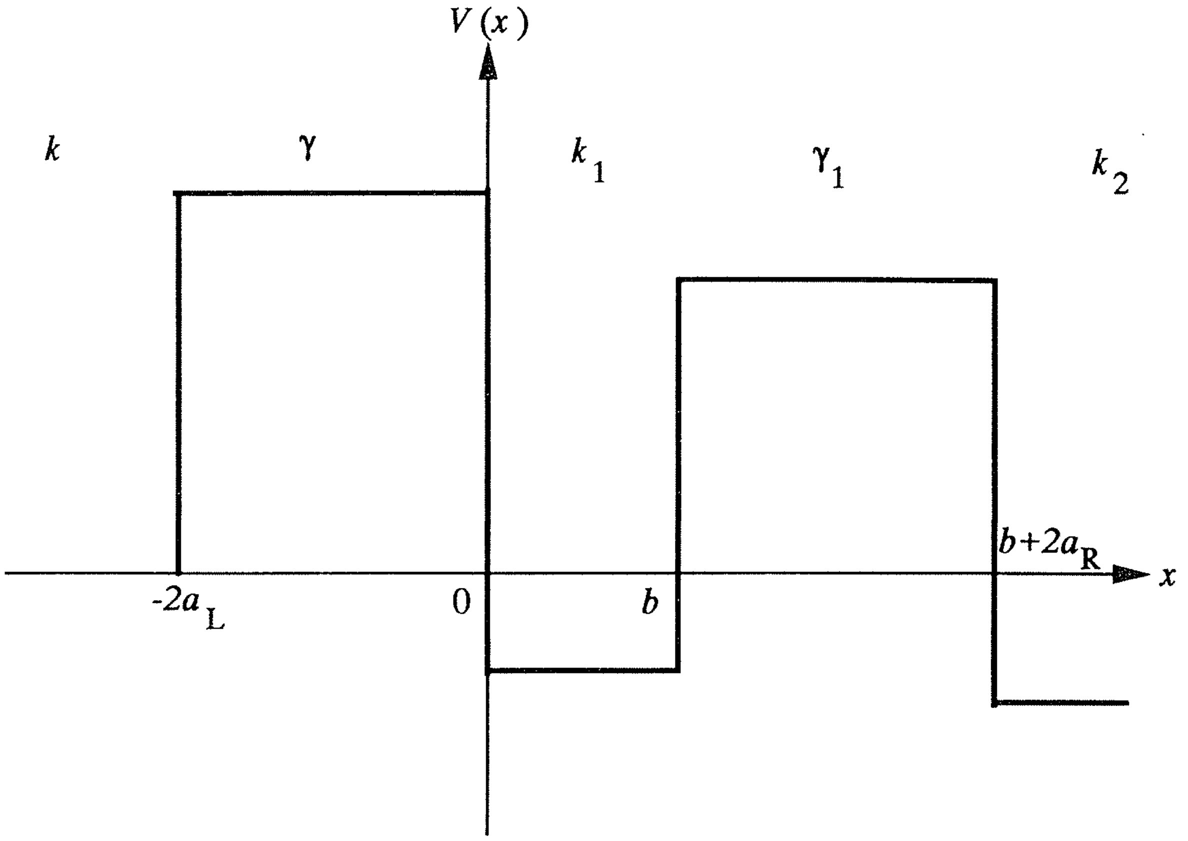


Figure 3.5 The potential structure for a general double barrier. The definition of the various constants is given for each region.

These values of the wave vector correspond to the resonant levels of a finite-depth quantum well (the finite-well values are shifted in phase from the infinite-well values by θ, which takes the value $-\pi/2$ in the latter case). Hence, as we supposed, the transmission rises to unity at values of the incident wave vector that correspond to resonant levels of the quantum well. In essence, the two barriers act like mirrors, and a resonant structure is created just as in electromagnetics. The incoming wave excites the occupancy of the resonance level until an equilibrium is reached in which the incoming wave is balanced by an outgoing wave and the overall transmission is unity. This perfect match is broken up if the two barriers differ, as we see below.

3.3.2 The unequal-barrier case

In the case where the two barriers differ, the results are more complicated, and greater care is necessary in the mathematics. The case we consider is indicated in figure 3.5. Here, we have individual wave vectors for the regions to the left and right of the composite barrier, as well as in the well. In addition, the decay constants of the two barriers differ, so the thicknesses and heights of the two barriers may also be different. Nevertheless, the result (3.26) still holds and will be our guide for obtaining the solution.

The definitions of the various functions are now taken from (3.19) and (3.20). We define the important quantities as

$$M_\mathrm{i} = m_\mathrm{i} e^{i\theta_\mathrm{i}} \tag{3.34}$$

where i = L11, L12, R11, R21. This leads to

$$m_{\mathrm{L}11} = \sqrt{\frac{1}{4}\left(1+\frac{k_1}{k}\right)^2 \cosh^2(2\gamma a_\mathrm{L}) + \frac{1}{4}\left(\frac{kk_1-\gamma^2}{k\gamma}\right)^2 \sinh^2(2\gamma a_\mathrm{L})} \tag{3.35}$$

$$m_{\mathrm{L}12} = \sqrt{\frac{1}{4}\left(\frac{kk_1+\gamma^2}{k\gamma}\right)^2 \sinh^2(2\gamma a_\mathrm{L}) + \frac{1}{4}\left(\frac{k_1}{k}-1\right)^2 \cosh^2(2\gamma a_\mathrm{L})} \tag{3.36}$$

$$m_{\mathrm{R}11} = \sqrt{\frac{1}{4}\left(1+\frac{k_2}{k_1}\right)^2 \cosh^2(2\gamma_1 a_\mathrm{R}) + \frac{1}{4}\left(\frac{k_1k_2-\gamma_1^2}{k_1\gamma_1}\right)^2 \sinh^2(2\gamma_1 a_\mathrm{R})} \tag{3.37}$$

$$m_{\mathrm{R}21} = \sqrt{\frac{1}{4}\left(\frac{k_1k_2+\gamma_1^2}{k_1\gamma_1}\right)^2 \sinh^2(2\gamma_1 a_\mathrm{R}) + \frac{1}{4}\left(\frac{k_2}{k_1}-1\right)^2 \cosh^2(2\gamma_1 a_\mathrm{R})} \tag{3.38}$$

$$\theta_{\mathrm{L}11} = -\tan^{-1}\left[\frac{kk_1-\gamma^2}{(k+k_1)\gamma}\tanh(2\gamma a_\mathrm{L})\right] + (k+k_1)a_\mathrm{L} \tag{3.39}$$

$$\theta_{\mathrm{L}12} = -\tan^{-1}\left[\frac{kk_1+\gamma^2}{(k_1-k)\gamma}\tanh(2\gamma a_\mathrm{L})\right] + \pi + (k-k_1)a_\mathrm{L} \tag{3.40}$$

$$\theta_{\mathrm{R}11} = -\tan^{-1}\left[\frac{k_1k_2-\gamma_1^2}{(k_1+k_2)\gamma_1}\tanh(2\gamma_1 a_\mathrm{R})\right] - (k_1+k_2)a_\mathrm{R} \tag{3.41}$$

$$\theta_{\mathrm{R}21} = \tan^{-1}\left[\frac{k_1k_2+\gamma_1^2}{(k_2-k_1)\gamma_1}\tanh(2\gamma_1 a_\mathrm{R})\right] + \pi + (k_1-k_2)a_\mathrm{R}. \tag{3.42}$$

These results for the phases and magnitudes of the individual terms of the transmission matrices can now be used in (3.26) to yield the net transmission matrix element, following the same procedure as above:

$$\begin{aligned}|M_{\mathrm{T}11}|^2 &= (m_{\mathrm{L}11}m_{\mathrm{R}11} - m_{\mathrm{L}12}m_{\mathrm{R}21})^2 \\ &\quad + 4m_{\mathrm{L}11}m_{\mathrm{R}11}m_{\mathrm{L}12}m_{\mathrm{R}21}\cos^2\left(kb + \frac{\theta_{\mathrm{L}12}+\theta_{\mathrm{R}21}-\theta_{\mathrm{L}11}-\theta_{\mathrm{R}11}}{2}\right).\end{aligned} \tag{3.43}$$

Now, as opposed to what was the case in the last sub-section, the first term (in the parentheses) does become unity. There is still a resonance, which occurs when the argument of the cosine function is an odd multiple of $\pi/2$. This is not a simple resonance, as it is in the previous case. Rather, the resonance involves phase shifts at each of the two interfaces. It is, however, convenient that the products of the wave vectors and the barrier thicknesses (the last terms in the four equations above for the phases) all reduce to a single term, that is

$k_1(a_L - a_R)/2$. This contribution to the phase shifts arises from the fact that the resonant energy level sinks below the infinite-quantum-well value due to the penetration of the wave function into the barrier region. This penetration is affected by the fact that the barrier has a finite thickness, and this term is a correction for the difference in thickness of the two sides. Obviously, if the thicknesses of the two barriers were made equal this term would drop out, and we would be left with the simple phase shifts at each boundary to consider.

At resonance, the overall transmission does not rise to unity because of the mismatch between the two barriers, which causes the first term to differ from unity. To find the actual value, we manipulate (3.43) somewhat to find the appropriate value. For this, we will assume that the attenuation of the barriers is rather high, so that the transmission of either would be $\ll 1$. Then, on resonance, we can write (3.43) as

$$\begin{aligned} |M_{T11}|^2 &= (m_{L11}m_{R11} - m_{L12}m_{R21})^2 \\ &= m_{L11}^2 m_{R11}^2 \left(1 - \frac{m_{L12}m_{R21}}{m_{L11}m_{R11}}\right)^2. \end{aligned} \tag{3.44}$$

Now, let us use (3.35) and (3.36) to write

$$\begin{aligned} \left(\frac{m_{L12}}{m_{L11}}\right)^2 &= \frac{1 + \dfrac{\gamma^2(k_1 - k)^2}{(\gamma^2 + k_2)(\gamma^2 + k_1^2)\sinh^2(2\gamma a_L)}}{1 + \dfrac{\gamma^2(k_1 + k)^2}{(\gamma^2 + k_2)(\gamma^2 + k_1^2)\sinh^2(2\gamma a_L)}} \\ &\simeq 1 - \frac{4kk_1\gamma^2}{(\gamma^2 + k^2)(\gamma^2 + k_1^2)\sinh^2(2\gamma a_L)} \simeq 1 - T_L \end{aligned} \tag{3.45}$$

and

$$\frac{m_{L12}}{m_{L11}} \simeq 1 - \frac{T_L}{2} \tag{3.46}$$

where we have used (3.21) in the limit $2\gamma a_L \gg 1$. We can do a similar evaluation for the other factor in (3.44), and finally can write (incorporating the ratio k_2/k to get the currents)

$$T = \frac{T_L T_R}{\left[1 - (1 - T_L/2)(1 - T_R/2)\right]^2} \simeq \frac{4T_L T_R}{(T_L + T_R)^2}. \tag{3.47}$$

Equation (3.47) is significant in that if the two transmissions are equal, a value of unity is obtained for the net transmission. On the other hand, if the two are not equal, and one is significantly different from the other, the result is that

$$T \simeq 4\frac{T_{\min}}{T_{\max}}. \tag{3.47a}$$

This implies that the transmission on resonance is given by the ratio of the minimum of the two individual barrier transmissions to the maximum of these two.

The opposite extreme is reached when we are away from one of the resonant levels. In this case, the maximum attenuation is achieved when the cosine function has the value of unity, and the resulting minimum in overall transmission is given by the value

$$T = T_L T_R/2 \tag{3.48}$$

in the limit of low transmission. It is clear from these discussions that if we are to maximize the transmission on resonance in any device application, it is important to have the transmission of the two barriers equal under any bias conditions that cause the resonance to appear.

3.4 APPROXIMATION METHODS—THE WKB METHOD

So far, the barriers that we have been treating are simple barriers in the sense that the potential $V(x)$ has always been piecewise constant. The reason for this lies in the fact that if the barrier height is a function of position, then the Schrödinger equation is a complicated equation that has solutions that are special functions. The example we treated in the last chapter merely had a linear variation of the potential—a constant electric field—and the result was solutions that were identified as Airy functions which already are quite complicated. What are we to do with more complicated potential variations? In some cases, the solutions can be achieved as well known special functions—we treat Hermite polynomials in the next chapter—but in general these solutions are quite complicated. On the other hand, nearly all of the solution techniques that we have used involve propagating waves or decaying waves, and the rest of the problem lay in matching boundary conditions. This latter, quite simple, observation suggests an approximation technique to find solutions, the Wentzel–Kramers–Brillouin approach (Wentzel 1926, Kramers 1926, Brillouin 1926).

Consider figure 3.6, in which we illustrate a general spatially varying potential. At a particular energy level, there is a position (shown as a) at which the wave changes from propagating to decaying. This position is known as a *turning point*. The description arises from the simple fact that the wave (particle) would be reflected from this point in a classical system. In fact, we can generally extend the earlier arguments and definitions of this chapter to say that

$$k(x) = \sqrt{\frac{2m}{\hbar^2}[E - V(x)]} \qquad \mathcal{E} > V(x) \tag{3.49}$$

and

$$\gamma(x) = \sqrt{\frac{2m}{\hbar^2}[V(x) - E]} \qquad \mathcal{E} < V(x). \tag{3.50}$$

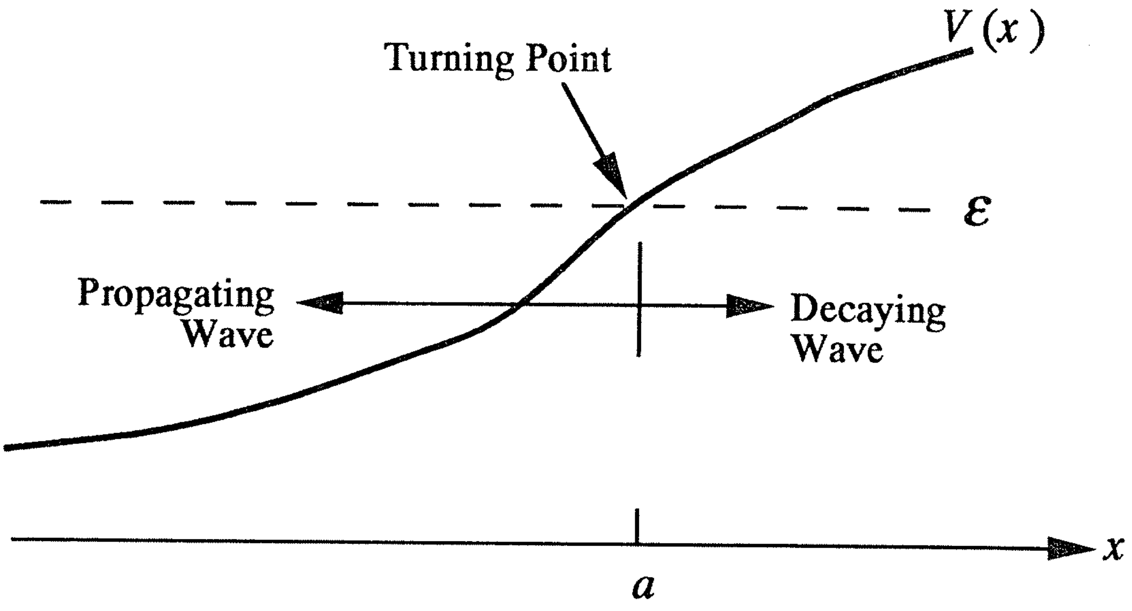


Figure 3.6 A simple variation of potential and the corresponding energy surface.

These solutions suggest that, at least to zero order, the solutions can be taken as simple exponentials that correspond either to propagating waves or to decaying waves.

The above ideas suggest that we consider a wave function that is basically a wave-type function, either decaying or propagating. We then adopt the results (3.49) and (3.50) as the lowest approximation, but seek higher approximations. To proceed, we assume that the wave function is generically definable as

$$\Psi(x) \sim \mathrm{e}^{\mathrm{i}u(x)} \tag{3.51}$$

and we now need to determine just what form $u(x)$ takes. This, of course, is closely related to the formulation adopted in section 2.1, and the differential equation for $u(x)$ is just (2.6) when the variation of the pre-factor of the exponent is ignored. This gives

$$\mathrm{i}\frac{\partial^2 u}{\partial x^2} - \left(\frac{\partial u}{\partial x}\right)^2 + k^2(x) = 0 \tag{3.52}$$

and equivalently for the decaying solution (we treat only the propagating one, and the decaying one will follow easily via a sign change). If we had a true free particle, the last two terms would cancel ($u = kx$) and we would be left with

$$\mathrm{i}\frac{\partial^2 u}{\partial x^2} = 0. \tag{3.53}$$

This suggests that we approximate $u(x)$ by making this latter equality an initial assumption for the lowest-order approximation to $u(x)$. To carry this further, we can then write the ith iteration of the solution as the solution of

$$\left(\frac{\partial u_i}{\partial x}\right)^2 = k^2(x) + \mathrm{i}\frac{\partial^2 u_{i-1}}{\partial x^2}. \tag{3.54}$$

We will only concern ourselves here with the first-order correction and approximation. The insertion of the zero-order approximation (which neglects the last term in (3.54)) into the equation for the first-order approximation leads to

$$\frac{\partial u_1}{\partial x} = \sqrt{k_2(x) + \mathrm{i}\frac{\partial k}{\partial x}} \simeq \pm k(x) + \mathrm{i}\frac{1}{k(x)}\frac{\partial k}{\partial x}. \tag{3.55}$$

In arriving at this last expression, we have assumed, in keeping with the approximations discussed, that the second term on the right-hand side in (3.55) is much smaller than the first term on the right. This implies that, in keeping with the discussion of section 2.1, the potential is slowly varying on the scale of the wavelength of the wave packet.

The result (3.55) can now be integrated over the position, with an arbitrary initial position as the reference point. This gives

$$u_1 \simeq \pm \int^x k(x')\,\mathrm{d}x' + \frac{\mathrm{i}}{2}\ln k(x) + \ln C_1 \tag{3.56}$$

which leads to

$$\Psi(x) \sim \frac{C_1}{\sqrt{k(x)}} \exp\left[\pm\mathrm{i}\int^x k(x')\,\mathrm{d}x'\right]. \tag{3.57}$$

The equivalent solution for the decaying wave function is

$$\Psi(x) \sim \frac{C_1}{\sqrt{\gamma(x)}} \exp\left[\pm\int^x \gamma(x')\,\mathrm{d}x'\right]. \tag{3.58}$$

It may be noted that these results automatically are equivalent to the requirement of making the current continuous at the turning point, which is achieved via the square-root pre-factors.

The remaining problem lies in connecting the waves of one type with those of the other at the turning point. The way this is done is through a method called the method of stationary phase. The details are beyond the present treatment, but are actually quite intuitive. In general, the connection formulas are written in terms of sines and cosines, rather than as propagating exponentials, and this will insert a factor of two, but only in the even functions of the propagating waves. In addition, the cosine waves always couple to the decaying solution, and a factor of $\pi/4$ is always subtracted from the phase of the propagating waves (this is a result of the details of the stationary-phase relationship and arises from the need to include a factor that is the square root of i). In figure 3.6, the turning point is to the right of the classical region (where $\mathcal{E} > V$). For this case, the connection formulas are given by

$$\frac{2}{\sqrt{k}}\cos\left(\int_x^a k\,\mathrm{d}x' - \frac{\pi}{4}\right) \leftrightarrow \frac{1}{\sqrt{\gamma}}\exp\left(-\int_a^x \gamma\,\mathrm{d}x'\right) \tag{3.59}$$

$$\frac{1}{\sqrt{k}}\sin\left(\int_x^a k\,\mathrm{d}x' - \frac{\pi}{4}\right) \leftrightarrow \frac{1}{\sqrt{\gamma}}\exp\left(\int_a^x \gamma\,\mathrm{d}x'\right) \tag{3.60}$$

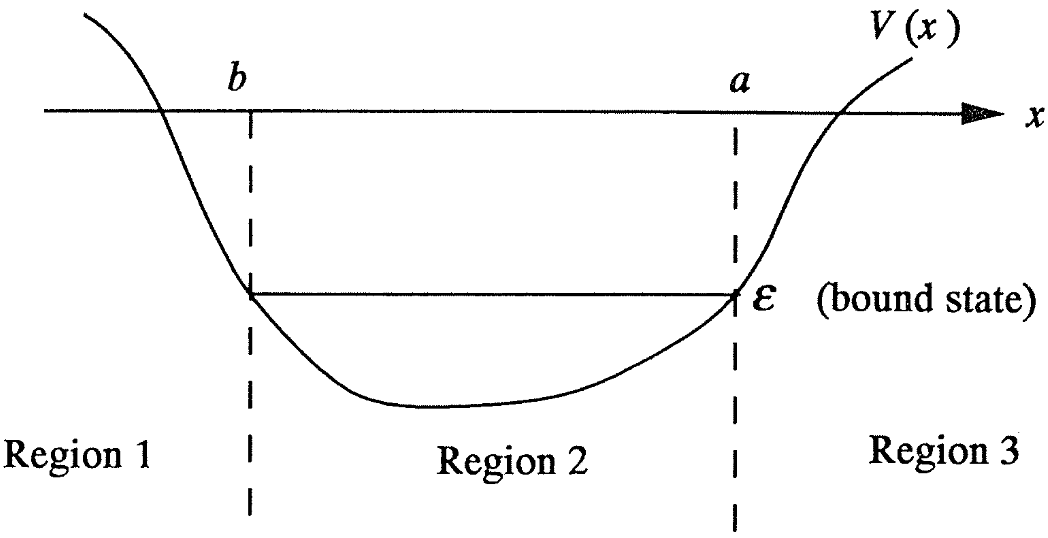


Figure 3.7 An arbitrary potential well in which to apply the WKB method.

The alternative case is for the mirror image of figure 3.6, in which the turning point is to the left of the classical region (in which the potential would be a decreasing function of x rather than an increasing function). For this case, the matching formulas are given as (the turning point is taken as $x = b$ in this case)

$$\frac{1}{\sqrt{\gamma}} \exp\left(-\int_x^b \gamma \, dx'\right) \leftrightarrow \frac{2}{\sqrt{k}} \cos\left(\int_b^x k \, dx' - \frac{\pi}{4}\right) \tag{3.61}$$

$$-\frac{1}{\sqrt{\gamma}} \exp\left(\int_x^b \gamma \, dx'\right) \leftrightarrow \frac{1}{\sqrt{k}} \sin\left(\int_b^x \gamma \, dx' - \frac{\pi}{4}\right). \tag{3.62}$$

To illustrate the application of these matching formulas, we consider some simple examples.

3.4.1 Bound states of a general potential

As a first example of the WKB technique, and the matching formulas, let us consider the general potential shown in figure 3.7. Our aim is to find the bound states, or the energy levels to be more exact. It is assumed that the energy level of interest is such that the turning points are as indicated; that is, the points $x = a$ and $x = b$ correspond to the turning points. Now, in region 1, to the left of $x = b$, we know that the solution has to be a decaying exponential as we move away from b. This means that we require that

$$\Psi_1(x) \simeq \frac{1}{\sqrt{\gamma}} \exp\left(-\int_x^b \gamma \, dx'\right) \qquad x < b. \tag{3.63}$$

At $x = b$, this must match to the cosine wave if we use (3.61). Thus, we know that in region 2, the wave function is given by

$$\Psi_2(x) \simeq \frac{2}{\sqrt{k}} \cos\left(\int_b^x k \, dx' - \frac{\pi}{4}\right) \qquad b < x < a. \tag{3.64}$$

We now want to work our way across to $x = a$, and this is done quite simply with simple manipulations of (3.64), as

$$\begin{aligned}\Psi_2(x) &\simeq \frac{2}{\sqrt{k}} \cos\left(\int_b^x k\,\mathrm{d}x' + \frac{\pi}{4} - \frac{\pi}{2}\right) = \frac{2}{\sqrt{k}} \sin\left(\int_b^x k\,\mathrm{d}x' + \frac{\pi}{4}\right) \\ &= \frac{2}{\sqrt{k}} \sin\left(\int_b^a k\,\mathrm{d}x' - \int_x^a k\,\mathrm{d}x' + \frac{\pi}{4}\right) \\ &= -\frac{2}{\sqrt{k}} \cos\left(\int_b^a k\,\mathrm{d}x'\right) \sin\left(\int_x^a k\,\mathrm{d}x' - \frac{\pi}{4}\right) \\ &\quad + \frac{2}{\sqrt{k}} \sin\left(\int_b^a k\,\mathrm{d}x'\right) \cos\left(\int_x^a k\,\mathrm{d}x' - \frac{\pi}{4}\right). \end{aligned} \tag{3.65}$$

We also know that the solution for the matching at the interface $x = a$ must satisfy (3.59), as the wave function in region 3 must be a decaying wave function. This means that at this interface, $\Psi_2(a)$ must be given *only* by the second term of (3.65). This can *only* be achieved by requiring that

$$\cos\left(\int_b^a k\,\mathrm{d}x'\right) = 0 \tag{3.66}$$

or

$$\int_b^a k\,\mathrm{d}x' = (2n+1)\frac{\pi}{2} \qquad n = 0, 1, 2, \ldots . \tag{3.67}$$

This equation now determines the energy eigenvalues of the potential well, at least within the WKB approximation.

If we compare (3.67) with the result for a sharp potential as the infinite quantum well of (2.48), with $b = -a$, we see that there is an additional phase shift of $\pi/2$ on the left-hand side. While one might think that this is an error inherent in the WKB approach, we note that the sharp potentials of the last chapter violate the assumptions of the WKB approach (slowly varying potentials). The extra factor of $\pi/2$ arises from the soft variation of the potentials. Without exactly solving the true potential case, one cannot say whether or not this extra factor is an error, but this factor is a general result of the WKB approach.

3.4.2 Tunnelling

It is not necessary to work out the complete tunnelling problem here, since we are interested only in the decay of the wave function from one side of the barrier to the other (recall that the input wave was always normalized to unity). It suffices to say that the spirit of the WKB approximation lies in the propagation (or decaying) wave vector, and the computation of the argument of the exponential decay function. The result (3.67) is that it is only the combination of forward and reverse waves that matter. For a barrier in which the attenuation is relatively large, only the decaying forward wave is important, and the tunnelling probability is approximately

$$T \sim \exp\left(-2\int_b^a \gamma \, \mathrm{d}x\right) \tag{3.68}$$

which implies that it is only the numerical coefficients (which involve the propagating and decaying wave vectors) that are lost in the WKB method. This tells us that we can use the limiting form of (3.14) ($b = -a$), or the equivalent limit of (3.21), with the argument of the exponential replaced with that of (3.68).

3.5 TUNNELLING DEVICES

One of the attractions of tunnelling devices is that it is possible to apply textbook quantum mechanics to gain an understanding of their operation, and still achieve a reasonable degree of success in actually getting quantitative agreement with experimental results. The concept of the tunnel 'diode' goes back several decades, and is usually implemented in heavily doped p–n junctions. In this case, the tunnelling is through the forbidden energy gap, as we will see below. Here, the tunnelling electrons make a transition from the valence band, on one side of the junction, to the conduction band on the other side. More recently, effort has centred on resonant tunnelling devices which can occur in a material with a single carrier type. Each of these will be discussed below, but first we need to formulate a current representation for the general tunnelling device.

3.5.1 A current formulation

In the treatment of the tunnelling problem that we have encountered in the preceding sections, the tunnelling process is that of a single plane-wave energy state from one side of the barrier to the other. The tunnelling process, in this view, is an energy-conserving process, since the energy at the output side is the same as that at the input side. In many real devices, the tunnelling process can be more complex, but we will follow this simple approach and treat a general tunnelling structure, such as that shown in figure 3.8. In the 'real' device, the tunnelling electrons are those within a narrow energy range near the Fermi energy, where the range is defined by the applied voltage as indicated in the figure. For this simple view, the device is treated in the linear-response regime, even though the resulting current is a non-linear function of the applied voltage. The general barrier can be a simple square barrier, or a multitude of individual barriers, just so long as the total tunnelling probability through the entire structure is *coherent*. By coherent here, we mean that the tunnelling through the entire barrier is an energy- and momentum-conserving process, so

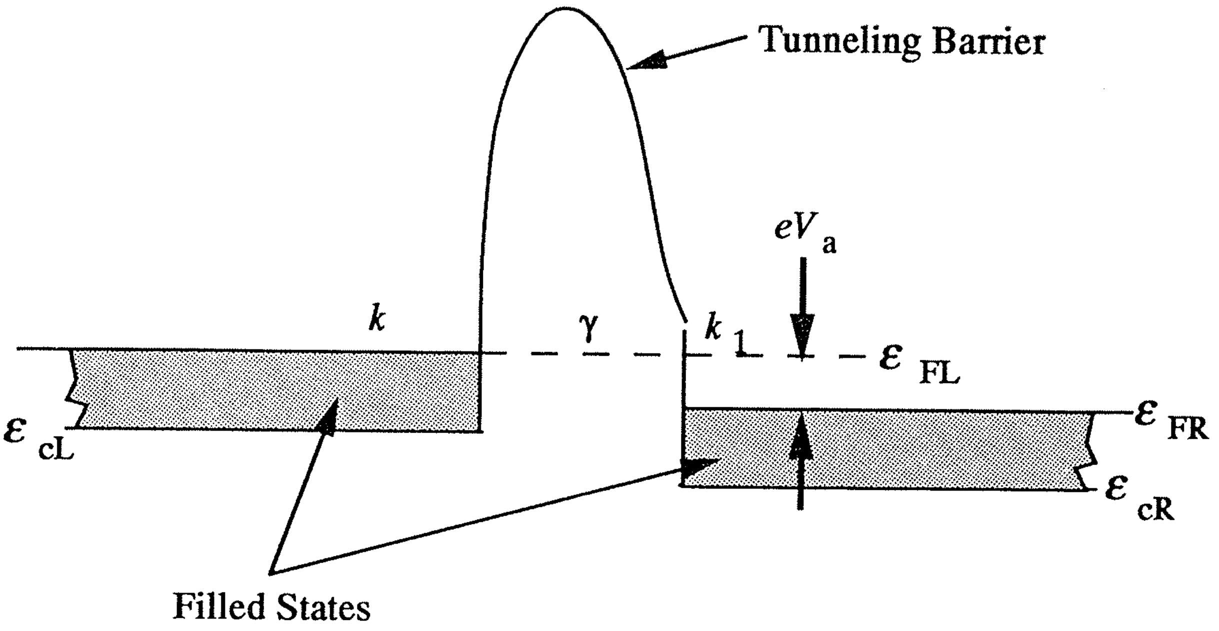


Figure 3.8 Tunnelling occurs from filled states on one side of the barrier to the empty states on the opposite side. The current is the net flow of particles from one side to the other.

no further complications are necessary. Hence, the properties of the barrier are completely described by the quantity $T(k)$.

In equilibrium, where there is no applied bias, the left-going and right-going waves are equivalent and there is no net current. By requiring that the energy be conserved during the process, we can write the z-component of energy as (we take the z-direction as that of the tunnelling current)

$$\mathcal{E} = \frac{\hbar^2 k_z^2}{2m} = \frac{\hbar^2 k_{1z}^2}{2m} + \text{constant} \tag{3.69}$$

where the constant accounts for the bias and is negative for a positive potential applied to the right of the barrier. The two wave vectors are easily related to one another by this equation, and we note that the derivative allows us to relate the velocities on the two sides. In particular, we note that

$$v_z(k_z)\,\mathrm{d}k_z = v_z(k_{1z})\,\mathrm{d}k_{1z}. \tag{3.70}$$

The current flow through the barrier is related to the tunnelling probability and to the total number of electrons that are available for tunnelling. Thus, the flow from the left to the right is given by

$$J_{\mathrm{LR}} = 2e \int \frac{\mathrm{d}^3 k}{(2\pi)^3}\, v_z(k_z) T(k_z) f(\mathcal{E}_{\mathrm{L}}) \tag{3.71}$$

where the factor of 2 is for spin degeneracy of the electron states, the $(2\pi)^3$ is the normalization on the number of $\boldsymbol{k}$ states (related to the density of states in $\boldsymbol{k}$

space), and $f(\mathcal{E}_{\mathrm{L}})$ is the electron distribution function *at the barrier*. Similarly, the current flow from the right to the left is given by

$$J_{\mathrm{RL}} = 2e \int \frac{\mathrm{d}^3 k_1}{(2\pi)^3} v_z(k_{1z}) T(k_{1z}) f(\mathcal{E}_{\mathrm{R}}). \tag{3.72}$$

Now, we know that the tunnelling probability is equal at the same energy, regardless of the direction of approach, and these two equations can be combined as

$$J = 2e \int \frac{\mathrm{d}^3 k}{(2\pi)^3} v_z(k_z) T(k_z) [f(\mathcal{E}_{\mathrm{L}}) - f(\mathcal{E}_{\mathrm{L}} + eV_{\mathrm{a}})] \tag{3.73}$$

where we have related the energy on the left to that on the right through the bias, as shown in figure 3.8, and expressed in (3.69). In the following, we will drop the subscript 'L' on the energy, but care must be used to ensure that it is evaluated on the left of the barrier.

Before proceeding, we want to simplify some of the relationships in (3.73). First, we note that the energy is a scalar quantity and can therefore be decomposed into its z-component and its transverse component, as

$$\mathcal{E} = \mathcal{E}_z + \mathcal{E}_\perp \tag{3.74}$$

and

$$\mathrm{d}^3 k = \mathrm{d}^2 k_\perp \, \mathrm{d}k_z. \tag{3.75}$$

We would like to change the last differential to one over the z-component of energy, and

$$\mathrm{d}k_z = \left(\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}k_z} \right)^{-1} \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}\mathcal{E}_z} \mathrm{d}\mathcal{E}_z. \tag{3.76}$$

The second term on the right-hand side is unity, so it drops out. The first term may be evaluated from (3.69) as

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}k_z} = \frac{\hbar^2 k_z}{m} = \hbar v_z. \tag{3.77}$$

The velocity term here will cancel that in (3.73), and we can write the final representation of the current as

$$J = \frac{e}{\pi\hbar} \int \frac{\mathrm{d}^2 k_\perp}{(2\pi)^2} \int \mathrm{d}\mathcal{E}_z \, T(\mathcal{E}_z) [f(\mathcal{E}_z + \mathcal{E}_\perp) - f(\mathcal{E}_z + \mathcal{E}_\perp + eV_{\mathrm{a}})]. \tag{3.78}$$

At this point in the theory, we really do not know the form of the distributions themselves, other than some form of simplifying assumption such as saying that they are Fermi–Dirac distributions. In fact, in metals, the distribution functions are well approximated by Fermi–Dirac distributions.

In semiconductors, however, the electric field and the current flow work to perturb the distributions significantly from their equilibrium forms, and this will introduce some additional complications. Additionally, the amount of charge in semiconductors is much smaller and charge fluctuations near the barriers can occur. This is shown in figure 3.9 as an example, where the density is plotted as a function of position along one axis and as a function of z-momentum along the other axis. There is a deviation of the distribution from its normal form as one approaches the barrier. This is quite simply understood. Electrons see a barrier in which the tunnelling is rather small. Thus, the wave function tries to have a value near zero at the interface with the barrier. The wave function then peaks at a distance of approximately $\lambda/2$ from the barrier. But this leads to a charge depletion right at the barrier, and the self-consistent potential will try to pull more charge toward the barrier. Electrons with a higher momentum will have their peaks closer to the barrier, so this charging effect leads to a distribution function with more high-energy electrons close to the barrier. In essence, this is a result of the Bohm potential of (2.7), as quantum mechanics does not really like to have a strongly varying density. In metals, where the number of electrons is quite high, this effect is easily screened out, but in semiconductors it can be significant. Whether or not it affects the total current is questionable, depending upon the size of the tunnelling coefficient. Nevertheless, we need to account for the distribution function being somewhat different from the normal Fermi–Dirac function.

We can avoid the approximations, at least in the linear-response regime, by deriving a relationship between the distribution functions on the two sides that will determine the deviations from equilibrium. For example, the electron population at the level k_z is obviously related to that on the right of the barrier by (Landauer 1957, 1970)

$$f_{\mathrm{L}}(-k_z) = Rf_{\mathrm{L}}(k_z) + Tf_{\mathrm{R}}(-k_{1z}) \tag{3.79}$$

where $R = 1 - T$ is the reflection coefficient. This means that electrons that are in the state $-k_z$ must arise either by tunnelling from the right-hand side or by reflection from the barrier. Using the relation between the reflection and tunnelling coefficients, we can rewrite this as

$$f_{\mathrm{L}}(k_z) - f_{\mathrm{L}}(-k_z) = Tf_{\mathrm{L}}(k_z) - Tf_{\mathrm{R}}(-k_{1z}). \tag{3.80}$$

Similarly, we can arrive at the equivalent expression for the distribution on the right-hand side of the barrier:

$$f_{\mathrm{R}}(k_{1z}) - f_{\mathrm{R}}(-k_{1z}) = Tf_{\mathrm{L}}(k_z) - Tf_{\mathrm{R}}(-k_{1z}). \tag{3.81}$$

The first thing that is noted from (3.80) and (3.81) is that the two left-hand sides must be equal, since the two right-hand sides are equal. Secondly, the terms on the right are exactly the terms necessary for the current equation (3.78).

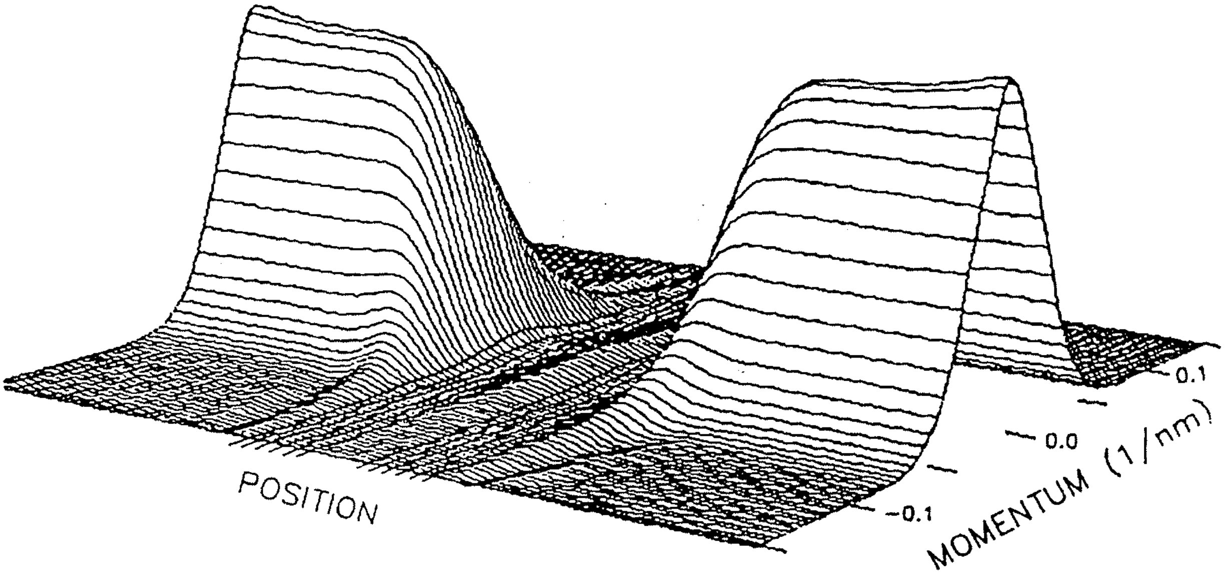


Figure 3.9 A quantum charge distribution, with a single tunnelling barrier located in the centre. The charge is plotted as a function of position along one axis and as a function of the z-component of momentum along the other. The double-barrier structure is indicated by the heavier lines parallel to the momentum axis that are drawn at the centre of the density.

To proceed further, we want to dissect the distribution functions in a manner suggested by (3.80) and (3.81). Here, we break each of the two functions into its symmetric and anti-symmetric parts, as

$$f(k_z) = f^{\mathrm{s}}(k_z) + f^{\mathrm{a}}(k_z) \tag{3.82}$$

and where we assume that each is still multiplied by the appropriate term for the transverse directions. Thus, we may write the two parts as

$$f^{\mathrm{s}}(k_z) = \tfrac{1}{2}[f(k_z) + f(-k_z)] \tag{3.83a}$$

$$f^{\mathrm{a}}(k_z) = \tfrac{1}{2}[f(k_z) - f(-k_z)]. \tag{3.83b}$$

Equations (3.80) and (3.81) now require that the two anti-symmetric parts of the distribution functions must be equal (the two left-hand sides, which are equal, are just the anti-symmetric parts), or

$$f_{\mathrm{L}}^{\mathrm{a}}(k_z) = f_{\mathrm{R}}^{\mathrm{a}}(k_{1z}) = f^{\mathrm{a}}(k_z). \tag{3.84}$$

This can now be used to find a value for the anti-symmetric term from the values of the symmetric terms, as

$$2f^{\mathrm{a}}(k_z) = T[f_{\mathrm{L}}^{\mathrm{s}}(k_z) - f_{\mathrm{R}}^{\mathrm{s}}(k_{1z})] + 2Tf^{\mathrm{a}}(k_z) \tag{3.85}$$

and

$$f^{\mathrm{a}}(k_z) = \frac{1}{2}\frac{T}{1-T}[f_{\mathrm{L}}^{\mathrm{s}}(k_z) - f_{\mathrm{R}}^{\mathrm{s}}(k_{1z})]. \tag{3.86}$$

It is this quantity, the anti-symmetric part of the distribution function, that is responsible for the tunnelling current (or for any current). The normalization of the symmetric part is the same as the equilibrium distribution function. That is, each of these normalizes to give the proper total density on either side of the barrier. For this reason, many authors linearize the treatment by replacing the symmetric part of the total distribution function with the Fermi–Dirac distribution function, and this is perfectly acceptable in the linear-response regime. The charge deviation that we saw in figure 3.9 is symmetric, but its effect is reflected in the ratio of the transmission to the reflection coefficients that appears in (3.86). Technically, the distortion shown in this latter value differs from the calculation that has been carried out here to find the anti-symmetric part of the overall distribution. However, both of these corrections are small (we are in linear response), and the effect of the factor $T/(1-T)$ introduces corrections that can account for both effects. When T is near unity, the latter factor can be much larger than unity. In principle, such corrections must include the extra high-energy carriers near the barrier, but this is an after-the-fact assertion. In the next section, we will see how the corrections of figure 3.9 should properly be included. The final equation for the current is then

$$J = \frac{e}{\pi\hbar}\int\frac{\mathrm{d}^2k_\perp}{(2\pi)^2}\int \mathrm{d}\mathcal{E}_z\frac{T(\mathcal{E}_z)}{1-T(\mathcal{E}_z)}[f^{\mathrm{s}}(\mathcal{E}_z+\mathcal{E}_\perp)-f^{\mathrm{s}}(\mathcal{E}_z+\mathcal{E}_\perp+eV_{\mathrm{a}})]. \tag{3.87}$$

(The factor of 2 in (3.86) cancels when the two distributions in (3.78) are put together, using the fact that the distribution on the right of the barrier is for a negative momentum, which flips its sign in the latter equation.)

3.5.2 The p–n junction diode

The tunnel diode is essentially merely a very heavily doped p–n junction, so the built-in potential of the junction is larger than the band gap. This is shown in figure 3.10(a). When a small bias is applied, as shown in figure 3.10(b), the filled states on one side of the junction overlap empty, allowed states on the other side, which allows current to flow. So far, this is no different from a normal junction diode, other than the fact that the carriers tunnel across the forbidden gap at the junction rather than being injected. However, it may be noted from figure 3.10(b) that continuing to increase the forward bias (the polarity shown) causes the filled states to begin to overlap states in the band gap, which are forbidden. Thus, the forward current returns to zero with increasing forward bias, and a negative differential conductance is observed. When combined with the normal p–n junction injection currents, an N-shaped conductance curve is obtained, which leads to the possibility of the use of the device for many novel electronic applications. In the reverse bias direction, the overlap of filled and

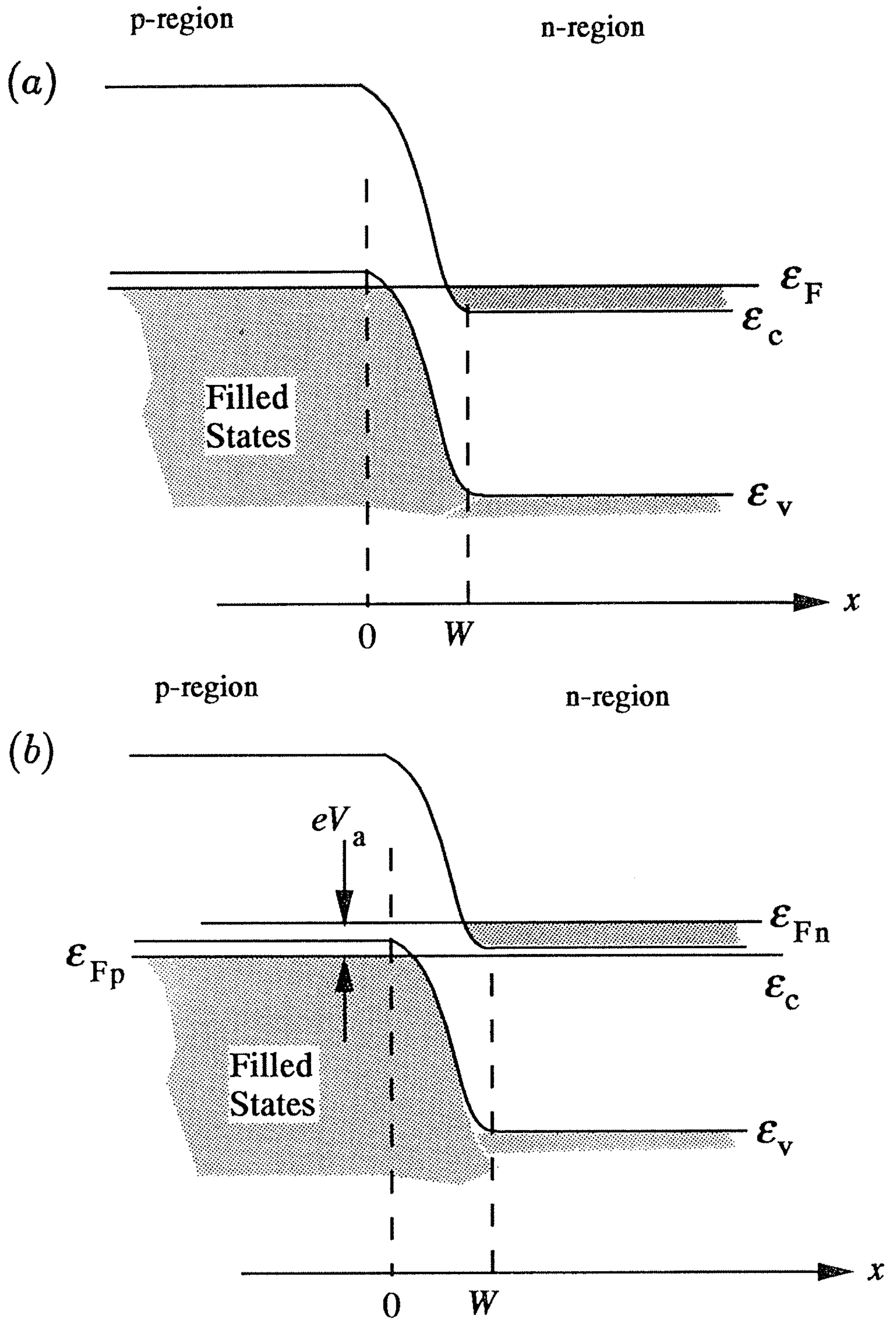


Figure 3.10 The band line-up for degenerately doped p–n junctions (*a*), and the possible tunnelling transitions for small forward bias (*b*).

empty (allowed) states continues to increase with all bias levels, so no negative conductance is observed in this direction of the current.

When the electric field in the barrier region is sufficiently large, the probability of tunnelling through the gap region is non-zero; for example, tunnelling can occur when the depletion width W is sufficiently small. One view of the tunnelling barrier is that it is a triangular potential, whose height is approximately equal to the band gap, and whose width at the tunnelling energy is the depletion width W. In section 2.6, we found that a triangular-potential region gave rise to wave functions that were Airy functions. The complications of these functions provide a strong argument for the use of the WKB approximation. Here, we can take the decay coefficient as

$$\gamma(x) \simeq \begin{cases} \sqrt{\dfrac{2m\mathcal{E}_G}{\hbar^2}\left(1 - \dfrac{x}{W} + \dfrac{\mathcal{E}_\perp}{\mathcal{E}_G}\right)} & 0 < x < W \\ 0 & \text{elsewhere} \end{cases} \tag{3.88}$$

where we have factored the energy gap out of the potential term and evaluated the electric field as $\mathcal{E}_G/eW$. The last term in the square root accounts for the transverse energy, since the tunnelling coefficient depends upon only the z-component of momentum (the z-component of energy must be reduced below the total energy by the transverse energy). This expression must now be integrated according to (3.68) over the tunnelling region, which produces

$$\begin{aligned} T &\simeq \exp\left[-2\int_0^W \sqrt{\frac{2m\mathcal{E}_G}{\hbar^2}\left(1 - \frac{x}{W} + \frac{\mathcal{E}_\perp}{\mathcal{E}_G}\right)}\,\mathrm{d}x\right] \\ &\simeq \exp\left[-\frac{4W}{3}\sqrt{\frac{2m\mathcal{E}_G}{\hbar^2}}\left(1 + \frac{3\mathcal{E}_\perp}{2\mathcal{E}_G}\right)\right] \end{aligned} \tag{3.89}$$

where we have expanded the radical to lowest order, and retained only the leading term in the transverse energy since it is considerably smaller than the band gap. It turns out that the result (3.89) is not sensitive to the actual details of the potential, since it is actually measuring the area under the V–$\mathcal{E}$ curve. Different shapes give the same result if the areas are equal. Recognizing this assures us that the approximation (3.89) is probably as good as any other. We can rewrite (3.88) as

$$T \simeq T_0 \exp\left[-\frac{\mathcal{E}_\perp}{\mathcal{E}_0}\right] \tag{3.90}$$

where

$$\mathcal{E}_0 = \frac{eE}{2}\sqrt{\frac{\hbar^2}{2m\mathcal{E}_G}}. \tag{3.91}$$

This can now be used in (3.86) to find the current.

We first will tackle the transverse energy integral. To lowest order, we note that the term involving the Fermi–Dirac functions is mainly a function of the longitudinal z-component of the energy, which we will show below, so the transverse terms are given by

$$\int_0^{\mathcal{E}_F - eV_a} \frac{\mathrm{d}^2k_\perp}{(2\pi)^2}\exp(-\mathcal{E}_\perp/\mathcal{E}_0) = \frac{m\mathcal{E}_0}{2\pi\hbar^2}[1 - \exp(-(\mathcal{E}_{Ft} - eV_a)/\mathcal{E}_0)]. \tag{3.92}$$

The limits on the previous integral are set by the fact that the transverse energy can only increase up to the sum of the Fermi energies on the two sides of the junction (measured from the band edges) reduced by the longitudinal energy.

The longitudinal contribution may be found by evaluating the energies in the Fermi–Dirac integrals, through shifting the energy on one side by the applied voltage eV_{a}. This leads to the result, in the linear-response limit, that

$$\begin{aligned}[f^{\mathrm{s}}(\mathcal{E}_z + \mathcal{E}_\perp) - f^{\mathrm{s}}(\mathcal{E}_z + \mathcal{E}_\perp + eV_{\mathrm{a}})] &\simeq \frac{eV_{\mathrm{a}}}{k_{\mathrm{B}}T} f^{\mathrm{s}}(1 - f^{\mathrm{s}}) \\ &\simeq -eV_{\mathrm{a}} \frac{\partial f^{\mathrm{s}}}{\partial \mathcal{E}_z} \simeq eV_{\mathrm{a}}\, \delta(\mathcal{E}_z - \mathcal{E}_{\mathrm{F}}).\end{aligned} \tag{3.93}$$

The last approximation is for strongly degenerate material (or equivalently, very low temperature). Then the integration over $\mathcal{E}_z$ gives just eV_{a} times the tunnelling probability T_0. We can now put (3.92) and (3.93) in the general equation (3.86) to obtain the total current density

$$J_z = \frac{eT_0}{\pi\hbar} \frac{m\mathcal{E}_{\mathrm{F}t}}{2\pi\hbar^2} \left(1 - \frac{eV_{\mathrm{a}}}{\mathcal{E}_{\mathrm{F}t}}\right) eV_{\mathrm{a}}. \tag{3.94}$$

As we discussed at the beginning of this section, the current rises linearly with applied bias, but then decreases as the electron states on the right-hand side begin to overlap the forbidden states in the energy gap, which cuts off the current. We show the tunnelling current in figure 3.11, along with the normal p–n junction current due to injection and diffusion.

3.5.3 The resonant tunnelling diode

The resonant tunnelling diode is one in which a double barrier is inserted into, say, a conduction band, and the current through the structure is metered via the resonant level. The latter corresponds to the energy at which the transmission rises to a value near unity. The structure of such a system, in the GaAs/AlGaAs/GaAs/AlGaAs/GaAs system with the AlGaAs forming the barriers, is shown in figure 3.12. Typically, the barriers are 3–5 nm thick and about 0.3 eV high, and the well is also 3–5 nm thick.

To proceed, we will use the same approximations as used for the p–n junction diode, at least for the distribution function. The difference beween the Fermi–Dirac distributions on the left-hand and right-hand sides, in the limit of very low temperature ($T \to 0$ K) gives

$$[f^{\mathrm{s}}(\mathcal{E}_z + \mathcal{E}_\perp) - f^{\mathrm{s}}(\mathcal{E}_z + \mathcal{E}_\perp + eV_{\mathrm{a}})] \simeq eV_{\mathrm{a}}\, \delta(\mathcal{E}_z + \mathcal{E}_\perp - \mathcal{E}_{\mathrm{F}}). \tag{3.95}$$

We retain the transverse energy in this treatment, since we must be slightly more careful in the integrations in this model. The tunnelling probability can also be

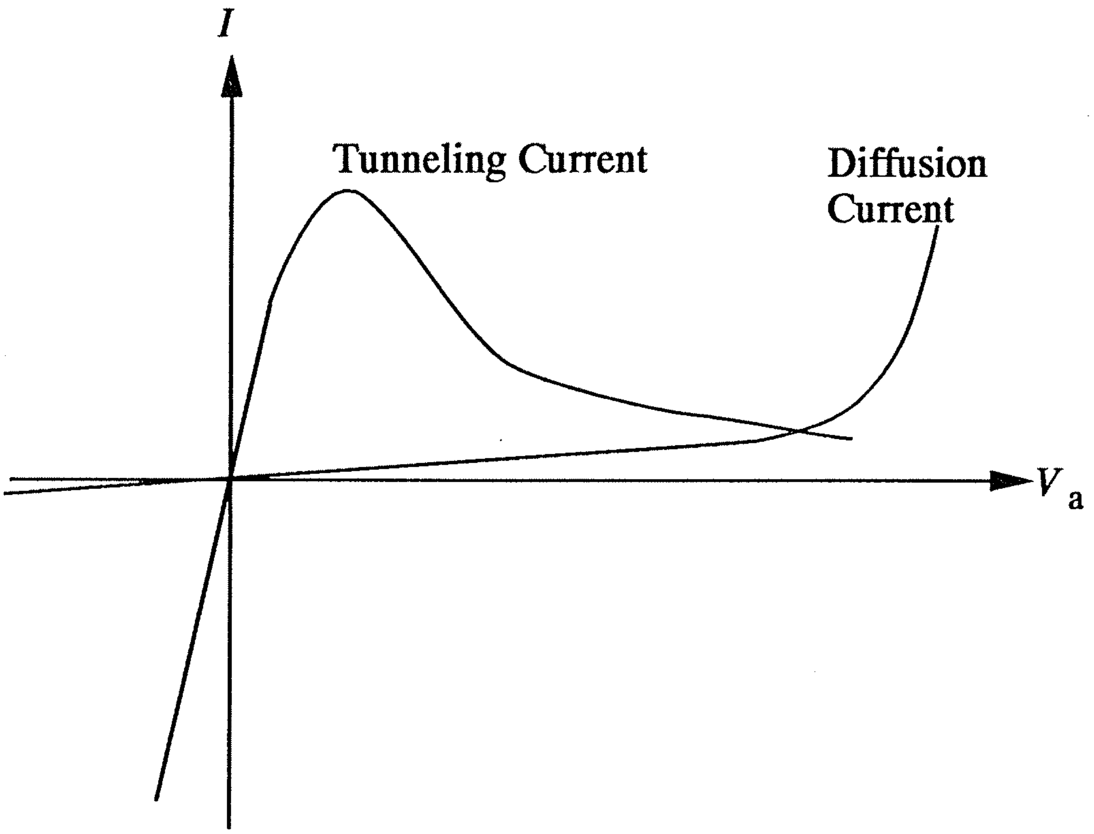


Figure 3.11 The contribution of the tunnelling current to the overall current of a tunnel diode.

taken as approximately a delta function, but with a finite width describing the nature of the actual lineshape (an alternative is to use something like a Lorentzian line, but this does not change the physics). Thus, we write (we note that the transmission will be less than unity and ignore the T-term in the denominator)

$$T(\mathcal{E}) \simeq \mathcal{E}_{\mathrm{W}}\, \delta(\mathcal{E}_z + eV_{\mathrm{a}}/2 - \mathcal{E}_0) \tag{3.96}$$

where we have assumed that the width of the transmission is $\mathcal{E}_{\mathrm{W}}$, and that the resonant level is shifted downward by an amount equal to half the bias voltage (everything is with reference to the Fermi energy on the left-hand side of the barrier, as indicated in the figure). Thus, the current can be written from (3.86) as

$$\begin{aligned} J_z &= \frac{e^2 V_{\mathrm{a}}}{\pi\hbar}\frac{m\mathcal{E}_{\mathrm{W}}}{2\pi\hbar^2}\int \mathrm{d}\mathcal{E}_z \int_0^{\mathcal{E}_{\mathrm{F}}} \delta(\mathcal{E}_z + \mathcal{E}_\perp - \mathcal{E}_{\mathrm{F}})\,\delta(\mathcal{E}_z + eV_{\mathrm{a}}/2 - \mathcal{E}_0)\,\mathrm{d}\mathcal{E}_\perp \\ &= \frac{e^2 V_{\mathrm{a}}}{\pi\hbar}\frac{m\mathcal{E}_{\mathrm{W}}}{2\pi\hbar^2}\int_0^{\mathcal{E}_{\mathrm{F}}} \delta(\mathcal{E}_{\mathrm{F}} - \mathcal{E}_\perp + eV_{\mathrm{a}}/2 - \mathcal{E}_0)\,\mathrm{d}\mathcal{E}_\perp \\ &= \frac{e^2 V_{\mathrm{a}}}{\pi\hbar}\frac{m\mathcal{E}_{\mathrm{W}}}{2\pi\hbar^2} \qquad 2(\mathcal{E}_0 - \mathcal{E}_{\mathrm{F}}) < eV_{\mathrm{a}} < 2\mathcal{E}_0. \end{aligned} \tag{3.97}$$

Outside of the indicated range of applied bias, the current is zero. At finite temperature (or if a Lorentzian lineshape for T is used), the current rises more

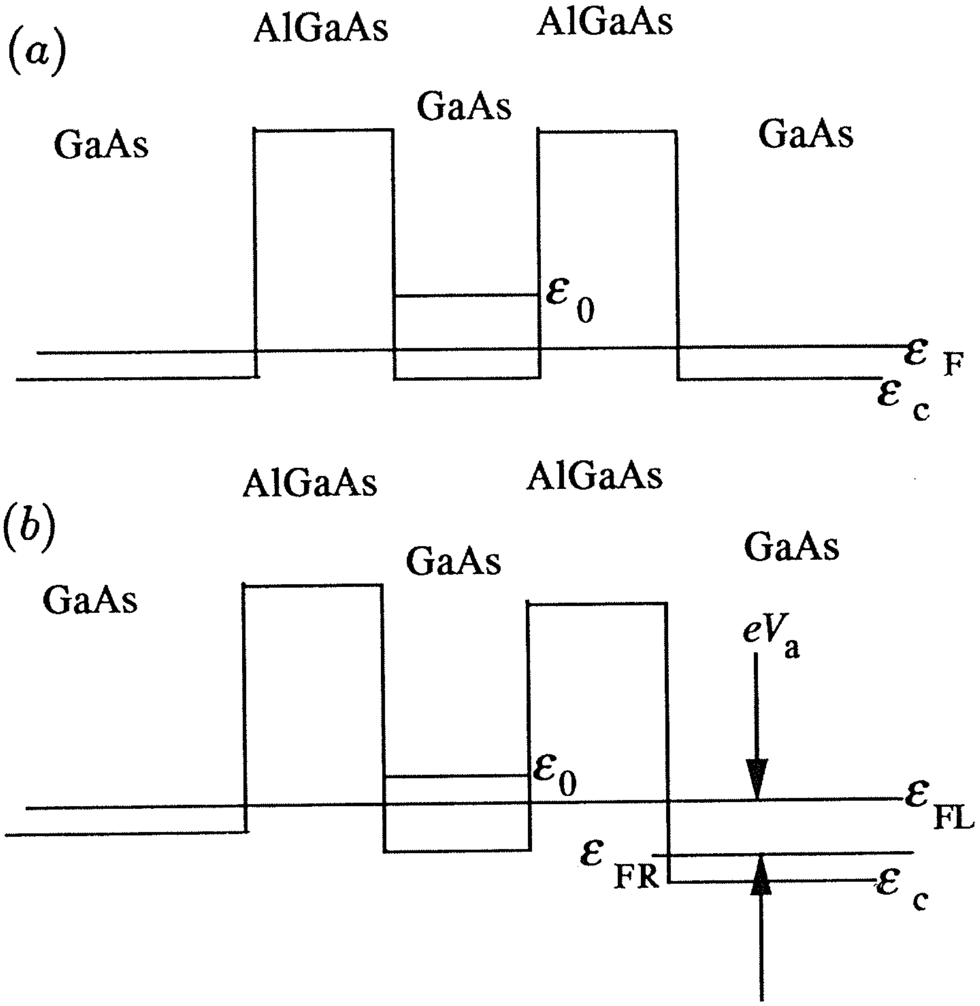


Figure 3.12 A typical double-barrier resonant tunnelling diode potential system, grown by heteroepitaxy in the GaAs–AlGaAs system. In (*a*), the basic structure is shown for an n-type GaAs well and cladding. In (*b*), the shape under bias is shown.

smoothly and drops more smoothly. Essentially, the current begins to flow as soon as the resonant level $\mathcal{E}_0$ is pulled down to the Fermi energy on the left-hand side (positive bias is to the right), and current ceases to flow when the resonant level passes the bottom of the conduction band. This is shown in figure 3.13, while experimentally observed curves are shown in figure 3.14.

3.6 THE LANDAUER FORMULA

The general approach that was used to evaluate the current equation (3.87) was to expand the difference between the distribution functions and use the resulting 'delta functions' to define a range of energies over which the tunnelling probability is summed. These energies correspond to those states that are full on one side of the barrier and empty on the other side (and, of course, allowed). Through the entire process, the current is 'metered' by the tunnelling probability. By this, we mean that the current is limited by this process. One question that has been raised is quite obvious: we have a current passing through a region defined by the tunnelling barriers and their cladding layers; we have a voltage

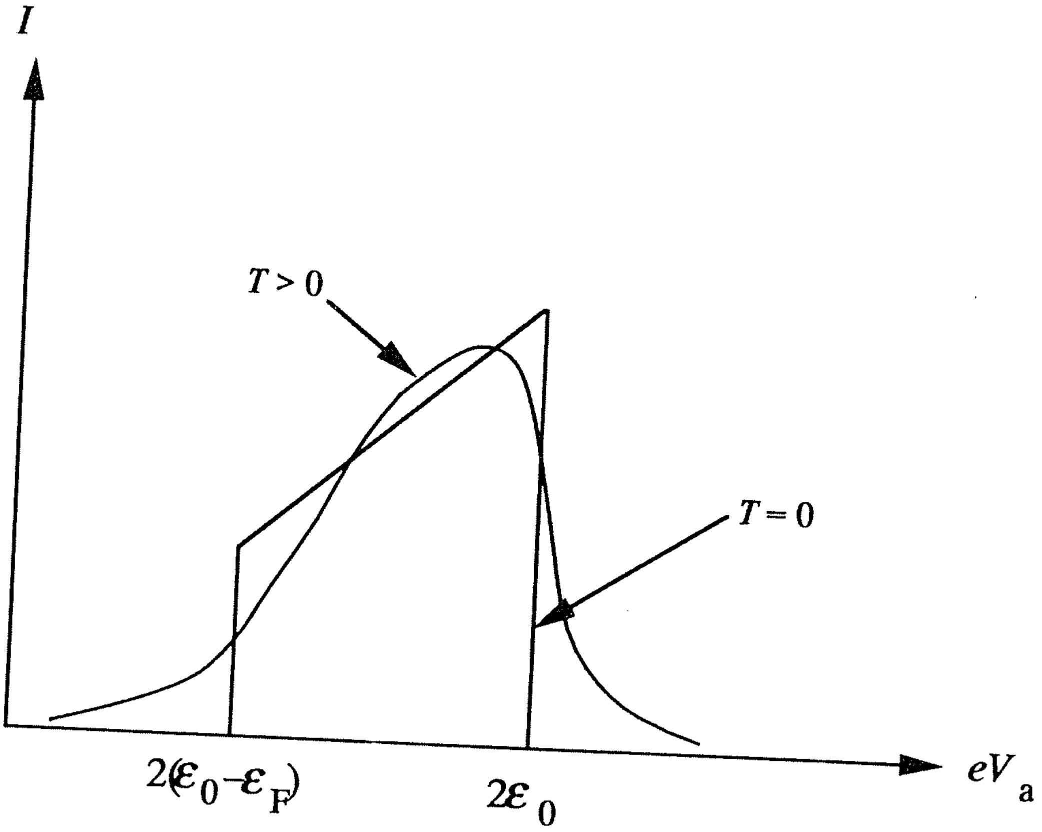


Figure 3.13 The theoretical curves for the simple model of (3.97) are shown for zero temperature and finite temperature.

drop as well. Yet, there is no dissipation within the system being considered! Where does the dissipation occur? It must occur in the contacts, since the current flows through the active tunnelling region in an energy-conserving fashion, as we have assumed. Thermalization of the carriers must occur in the contact. Thus, the tunnelling region determines the current flow for a given voltage drop (or determines the voltage required for a given current flow), but the dissipation occurs at the boundaries. This is quite unusual, but can be correct in small systems, referred to as mesoscopic systems. We can examine this contact effect further.

Let us integrate (3.86) over the transverse dimensions, so that it can be rewritten as

$$\begin{aligned} I &= \frac{e}{\pi\hbar} A \int \frac{\mathrm{d}^2 k_\perp}{(2\pi)^2} \int \mathrm{d}\mathcal{E}_z \frac{T(\mathcal{E}_z)}{1 - T(\mathcal{E}_z)} [f^s(\mathcal{E}_z + \mathcal{E}_\perp) - f^s(\mathcal{E}_z + \mathcal{E}_\perp + eV_a)] \\ &= \frac{e^2 V_a}{\pi\hbar} \frac{mA}{2\pi\hbar^2} \int \mathrm{d}\mathcal{E}_\perp \int \mathrm{d}\mathcal{E}_z \frac{T(\mathcal{E}_z)}{1 - T(\mathcal{E}_z)} \delta(\mathcal{E}_z + \mathcal{E}_\perp - \mathcal{E}_F) \\ &= \frac{e^2 V_a}{\pi\hbar} \frac{mA}{2\pi\hbar^2} \int_0^{\mathcal{E}_F} \mathrm{d}\mathcal{E}_\perp \frac{T(\mathcal{E}_F - \mathcal{E}_\perp)}{1 - T(\mathcal{E}_F - \mathcal{E}_\perp)} \\ &= \frac{e^2 V_a}{\pi\hbar} \frac{m\mathcal{E}_a A}{2\pi\hbar^2} \frac{T(\mathcal{E}_a)}{1 - T(\mathcal{E}_a)}. \end{aligned} \qquad (3.98)$$

Here, $\mathcal{E}_a$ is an average transverse energy. The second fraction in (3.98) is an interesting quantity, in that it is essentially just $k_a^2 A$, where k_a is the wave vector

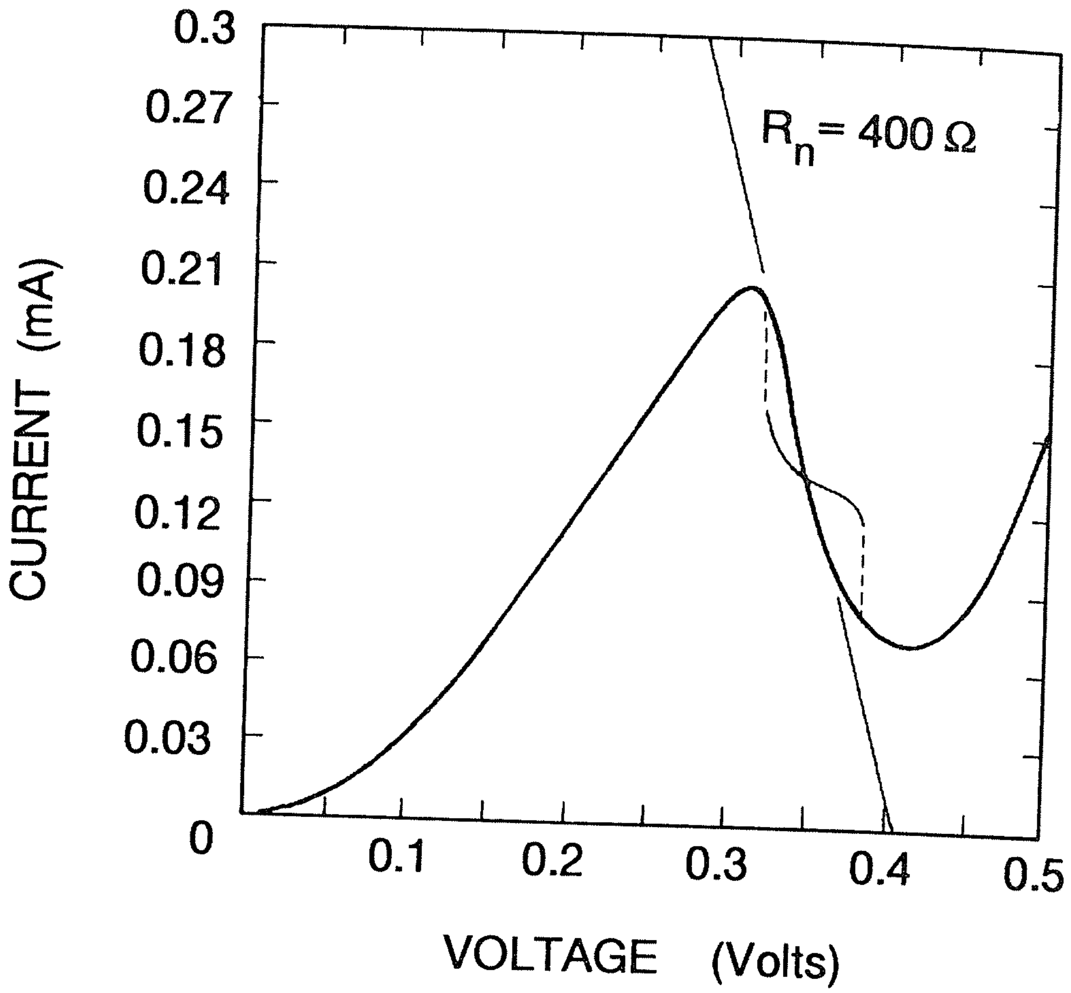


Figure 3.14 The experimental curves obtained for a GaAs–AlGaAs structure with 5 nm barriers and a 5 nm well. The extra current above the drop-off is due to higher resonant states and emission over the top of the barrier; both are forms of leakage. (After Sollner *et al* (1983), by permission.)

corresponding to this average energy. This fraction is just the number of allowed transverse states that can contribute to the current. If we call this latter quantity N_t, then we can write (3.98) as (we use $I = GV_a$ to write only the conductance G)

$$G = \frac{e^2}{\pi\hbar}\sum_{i=1}^{N_t}\frac{T_i}{1-T_i} \tag{3.99}$$

where it is assumed that energy conservation ensures that there is no change in the number of transverse states. If we refer to the transverse states by the term transverse modes, then (3.98) is termed the Landauer formula (the notation used here is a simple version, assuming no mode coupling). It is normally seen only in small mesoscopic systems applications, but it is clear that its applications are even to normal tunnelling structures so long as we recall just what the summation means.

There is an interesting suggestion in (3.98). In large systems, where the number of transverse states is enormous, and where the conductance can vary over a large range, the conductance is a smooth function of the energy. As

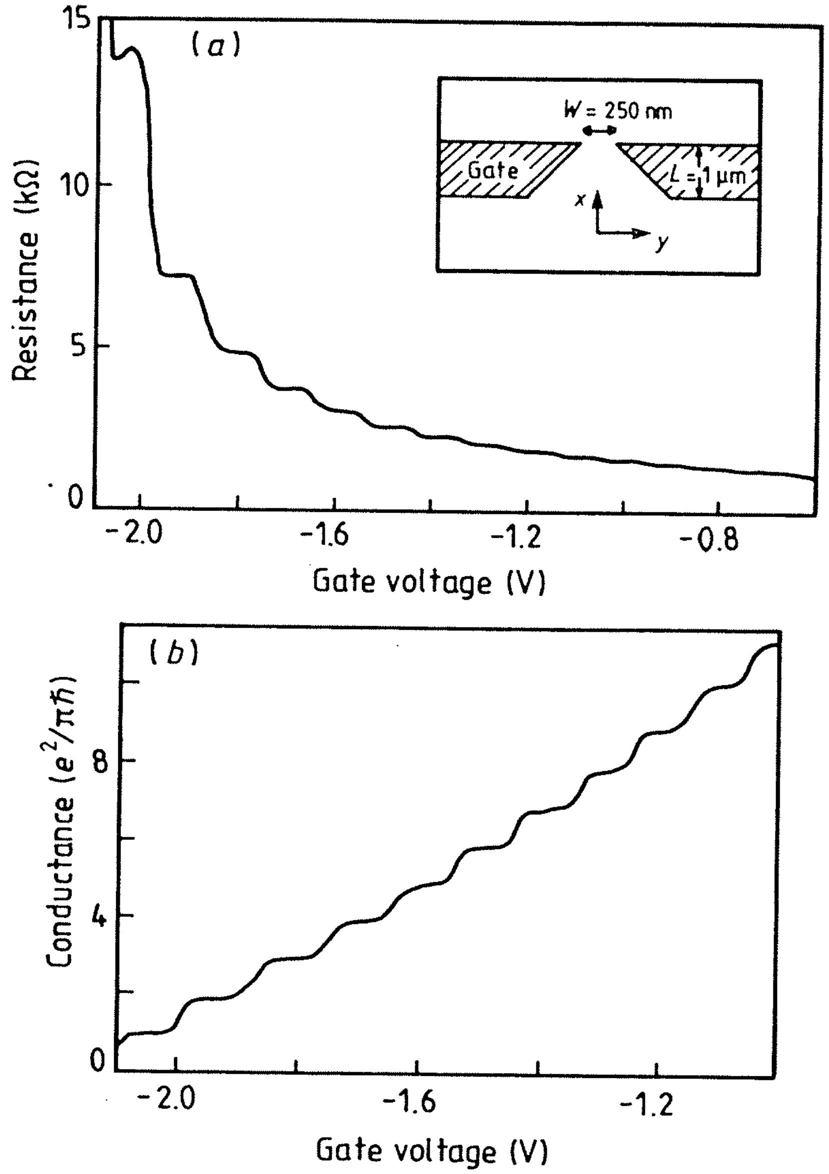


Figure 3.15 Quantized resistance (*a*) and conductance (*b*) can be observed in small conducting systems. Here, the number of transverse states in the small opening between the metal gates is varied by changing the bias on the gates (shown in the inset to (*a*)). (After van Wees *et al* (1988), by permission.)

the Fermi energy, or the bias voltage, is varied, the number of states affected is so large that the conductance is a smooth function of the bias voltage. In small systems, however, the number of transverse modes is quite small, and the conductance should increase in steps of $e^2/\pi\hbar$—as the bias, or the number of transverse modes, is varied. This variation has only been recognized in the past few years, and we show one of the early experiments in figure 3.15. Here, the structure is composed of a GaAs/GaAlAs heterostructure in which the electrons at the interface (on the GaAs side of the interface) sit in a triangular potential, as in section 2.6. Their motion normal to the interface is quantized; however, they are free to move in the plane of the interface and form what is known as a quasi-two-dimensional electron gas. On the surface, metal gates are so placed that when biased they deplete the electrons under the gate. Thus the structure shown in the inset will allow the electrons to move between the two large-area

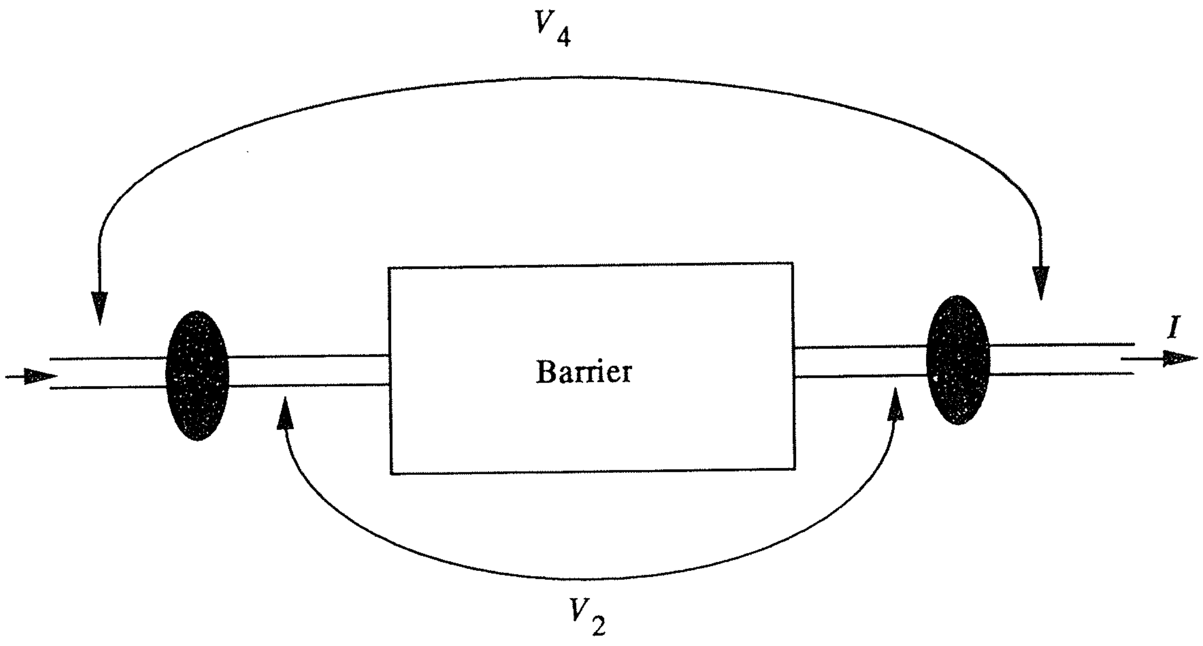


Figure 3.16 The two-terminal and four-terminal resistances are defined in terms of the voltages V_2 and V_4, respectively. They differ by the contact resistances.

regions, but only a very few transverse states exist in the opening. This number can be varied by adjusting the bias on the metal gates, and the conductance shows steps as indicated in the figure. In this measurement, the tunnelling probability is unity as there is no barrier; in fact, as the transverse states are populated and carry current, their transmission coefficient changes from zero to one.

When the transmission is near unity, why do we not see the denominator term playing a larger part? The answer is that the measurement is a 'two-terminal' measurement. Consider, for the moment, only a single transverse state, so that (3.99) can be written as

$$G = \frac{e^2}{\pi\hbar}\frac{T_i}{1 - T_i}. \tag{3.100}$$

We may assert that this is the resistance just across the 'tunnelling' region, and must be modified by the contact resistance for a measurement in which the potential drop is measured at the current leads (a 'two-terminal' measurement; see figure 3.16). If we rewrite this equation in terms of resistances, then

$$R_4 = \frac{\pi\hbar}{e^2}\left(\frac{1}{T_i} - 1\right) \tag{3.101}$$

where the subscript refers to a measurement in which the potential is measured at the barriers and at contacts that are independent of the current leads. The difference lies in the fact that the contacts are areas where equilibration occurs. If we recognize that the original form of the current density (3.77) implied a two-terminal definition, we can say that

$$R_2 = \frac{\pi\hbar}{e^2}\frac{1}{T_i} \tag{3.102}$$

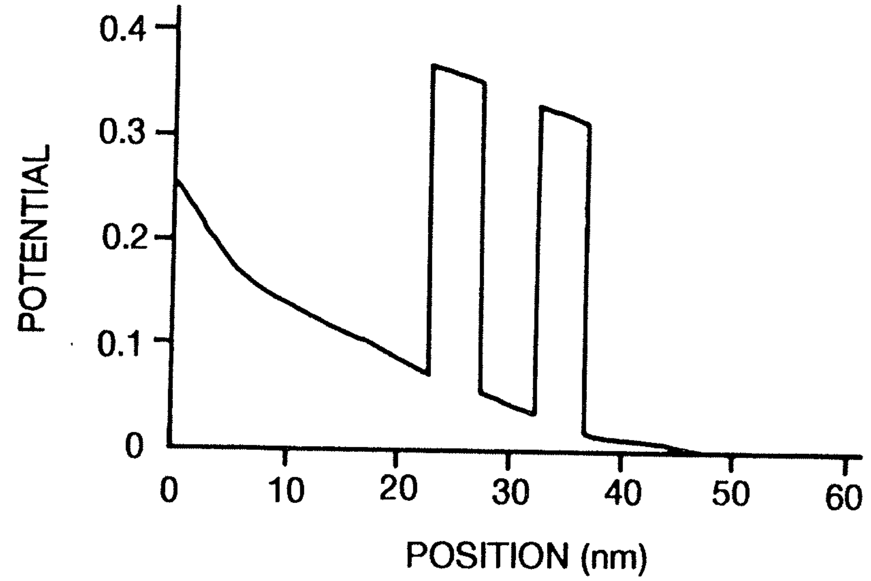


Figure 3.17 The potential profile for a resonant tunnelling diode, in which a depletion region (to the left of the barriers) creates a contact resistance to balance the current-carrying preferences of the barriers and the contacts.

and the difference is given by

$$R_2 = R_4 + R_c \qquad R_c = \frac{\pi\hbar}{e^2}. \tag{3.103}$$

The last form defines the contact resistance R_c.

Contact resistances are a function of all basic dissipative structures, even though the dissipation in the present problem is actually in the contact. Nevertheless, when the contacts want to carry a current different from that of the 'barrier' regions, for a given voltage drop, then additional resistance occurs in the structure. This is shown in figure 3.17 for a model of a resonant tunnelling diode, in which the potential throughout the device can be obtained self-consistently from Poisson's equation. The curvature to the left of the barriers is due predominantly to carrier depletion here which leads to a 'contact' resistance in the structure.

How are we to interpret the difference between the two-terminal and the four-terminal conductances, and therefore how are we to interpret the Landauer formula. If we are truly in the boundary regions, where the distribution function is a Fermi–Dirac distribution, then we can use the two-terminal formula, provided that we compute the total transmission *over the entire region between the boundaries*, with the full variation of the self-consistent potential with position in that region. On the other hand, if we separate the current contacts and the potential contacts, a four-terminal formula may be used, as long as it is interpreted carefully. Effects such as those in figure 3.9 must be carefully included in the region over which the transmission coefficient is being calculated (or measured). Even with a four-terminal measurement, it must be ascertained that the actual contact resistance differences are just those expected and no unusual effects have been overlooked.

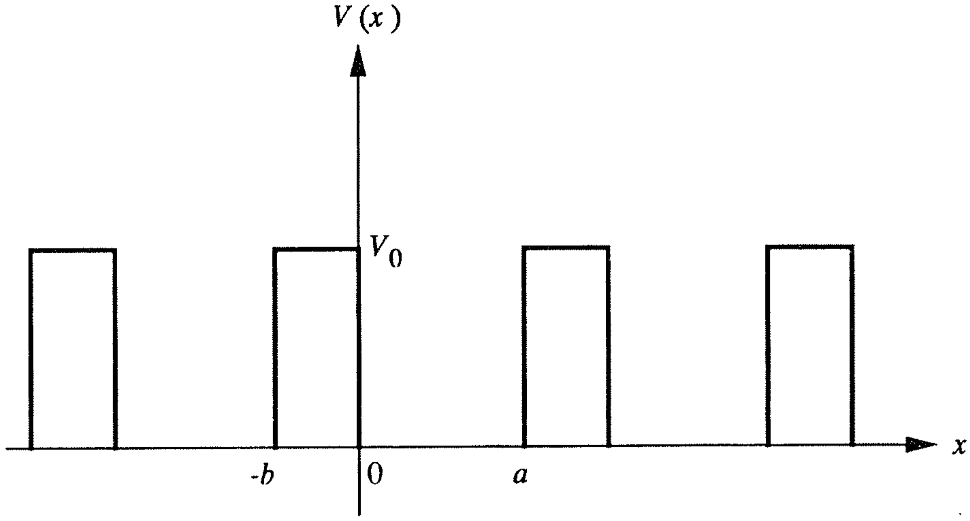


Figure 3.18 A simple periodic potential.

3.7 PERIODIC POTENTIALS

At this point, we want to turn our attention to an array of quantum wells, which are spaced by barriers sufficiently thin that the wave functions can tunnel through in order to couple the wells together. In this sense, we create a *periodic* potential. Due to the extreme complexity of the true periodic potential, for purposes of calculation it is preferable to simplify the model considerably. For that purpose, we will assume square barriers and wells, as shown in figure 3.18. Although the potential model is only an approximation, it enables us to develop the essential features, which in turn will not depend crucially upon the details of the model. The importance of this model is in the energy band structure of crystalline media, such as semiconductors in which the atoms are arranged in a periodic array, and the atomic potentials create a periodic potential in three dimensions in which the electrons must move. The important outcomes of the model are the existence of ranges of allowed energies, called bands, and ranges of forbidden energies, called gaps. We have already, in the previous sections, talked about band gaps in p–n junctions. Here, we review just how periodic potentials give rise to such bands and gaps in the energy spectrum.

The (atomic) potential is represented by the simple model shown in figure 3.18, and such details as repulsive core potentials will be ignored. Our interest is in the filtering effect such a periodic structure has on the energy spectrum of electron waves. The periodic potential has a basic lattice constant (periodicity) of $d = a + b$. We are interested in states in which $\mathcal{E} \ll V_0$. The Schrödinger equation now becomes

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\Psi_1}{\mathrm{d}x^2} - E\Psi_1 = 0 \qquad 0 < x < a \tag{3.104a}$$

and

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\Psi_1}{\mathrm{d}x^2} - E\Psi_1 = -V_0\Psi_1 \qquad -b < x < 0. \tag{3.104b}$$

Of course, shifts of the x-axis by the amount d bring in other regions in which (3.104) is found to be the appropriate equation. Nevertheless, there will be a point at which we will *force* the periodicity onto the solutions. We also expect that the wave function will be periodic with the same periodicity as the potential, or

$$\Psi_1(x) = \mathrm{e}^{\mathrm{i}Kx}u(x) \tag{3.105}$$

where $u(x)$ has the periodicity of the lattice. A wave of the form (3.105) is termed a Bloch function. If we insert (3.104) into (3.103), the resulting equation is

$$\frac{\mathrm{d}^2u_1}{\mathrm{d}x^2} + 2\mathrm{i}k\frac{\mathrm{d}u_1}{\mathrm{d}x} + (k^2 - K^2)u_1 = 0 \qquad 0 < x < a \tag{3.106a}$$

and

$$\frac{\mathrm{d}^2u_2}{\mathrm{d}x^2} + 2\mathrm{i}k\frac{\mathrm{d}u_2}{\mathrm{d}x} + (\gamma^2 + K^2)u_2 = 0 \qquad -b < x < 0. \tag{3.106b}$$

Here, k and γ have their normal meanings as defined in (3.1). These can now be solved by normal means to yield

$$u_1 = A\mathrm{e}^{-\mathrm{i}(K-k)x} + B\mathrm{e}^{-\mathrm{i}(K+k)x} \qquad 0 < x < a \tag{3.107a}$$

$$u_2 = C\mathrm{e}^{-(\mathrm{i}K-\gamma)x} + D\mathrm{e}^{-(\mathrm{i}K+\gamma)x} \qquad -b < x < 0. \tag{3.107b}$$

These solutions again represent waves, in each case (either propagating or evanescent), one propagating in each direction.

There are now four unknowns, the coefficients that appear in (3.107). However, there are only two boundaries in effect. Hence, we require that both the wave function and its derivative be continuous at each boundary. However, it is at this point that we will force the periodicity onto the problem via the choice of matching points. This is achieved by choosing the boundary conditions to satisfy

$$u_1(0) = u_2(0) \tag{3.108}$$

$$u_1(a) = u_2(-b) \tag{3.109}$$

$$\frac{\mathrm{d}u_1(0)}{\mathrm{d}x} = \frac{\mathrm{d}u_2(0)}{\mathrm{d}x} \tag{3.110}$$

$$\frac{\mathrm{d}u_1(a)}{\mathrm{d}x} = \frac{\mathrm{d}u_2(-b)}{\mathrm{d}x}. \tag{3.111}$$

The choice of the matching points, specifically the choice of $-b$ instead of a on u_2, causes the periodicity to be imposed upon the solutions. These four equations lead to four equations for the coefficients, and these form a homogeneous set of equations. There are no *forcing* terms in the equations. Thus, the coefficients can differ from zero only if the determinant of the coefficients vanishes. This leads to the determinantal equation

$$\begin{vmatrix} 1 & 1 & -1 & -1 \\ \mathrm{e}^{-\mathrm{i}(K-k)a} & \mathrm{e}^{-\mathrm{i}(K+k)a} & -\mathrm{e}^{(\mathrm{i}K-\gamma)b} & -\mathrm{e}^{(\mathrm{i}K+\gamma)b} \\ -\mathrm{i}(K-k) & -\mathrm{i}(K+k) & \mathrm{i}K-\gamma & \mathrm{i}K+\gamma \\ -\mathrm{i}(K-k)\mathrm{e}^{-\mathrm{i}(K-k)a} & -\mathrm{i}(K+k)\mathrm{e}^{-\mathrm{i}(K+k)a} & (\mathrm{i}K-\gamma)\mathrm{e}^{(\mathrm{i}K-\gamma)b} & (\mathrm{i}K+\gamma)\mathrm{e}^{(\mathrm{i}K+\gamma)b} \end{vmatrix} = 0. \tag{3.112}$$

Evaluating this determinant leads to

$$\frac{\gamma^2 - k^2}{2k\gamma} \sinh(\gamma b)\sin(ka) + \cosh(\gamma b)\cos(ka) = \cos[K(b+a)]. \tag{3.113}$$

In one of the previous sections, it was pointed out that the true measure of a tunnelling barrier was not its height, but the product γb. Here we will let $V_0 \to \infty$, but keep the product $V_0 b = Q$ finite, which also requires taking the simultaneous limit $b \to 0$. Since γb varies as the square root of the potential, this quantity approaches zero, so (3.113) can be rewritten as

$$\frac{\gamma^2 b}{2k} \sin(ka) + \cos(ka) = \cos(Ka). \tag{3.114}$$

The right-hand side of (3.114) is constrained to lie in the range $[-1, 1]$, so the left-hand side is restricted to values of k, a that yield a value in this range. Now, these latter constants are not constrained to have these values, but it is only when they do that the determinant vanishes. This means that the wave functions have values differing from zero only for those values of k, a for which (3.114) is satisfied. This range can be found graphically, as shown in figure 3.19 (in the figure, only the positive values of Ka are shown, as the figure is completely symmetrical about $Ka = 0$, as can be seen by examining the above equations). The ranges of k, a for which (3.114) is satisfied are known as the allowed states. Other values are known as forbidden states. The allowed states group together in bands, given by the case for which the left-hand side traverses the range $[-1, 1]$. Each allowed band is separated from the next by a forbidden gap region, for which the left-hand side has a magnitude greater than unity.

In this model, k is a function of the energy of the single electron, so the limits on the range of this parameter are simply the limits on the range of allowed energies. If this is the case, then the results should agree with the results for free electrons and for bound electrons. In the former case, the pre-factor of the first term in (3.114) vanishes, and we are left with $k = K$. Thus, the energy is just given by the wave vector in the normal manner. On the other hand, when the pre-factor goes to infinity, we are left with

$$\sin(ka) = 0 \qquad k = \frac{n\pi}{a} \tag{3.115}$$

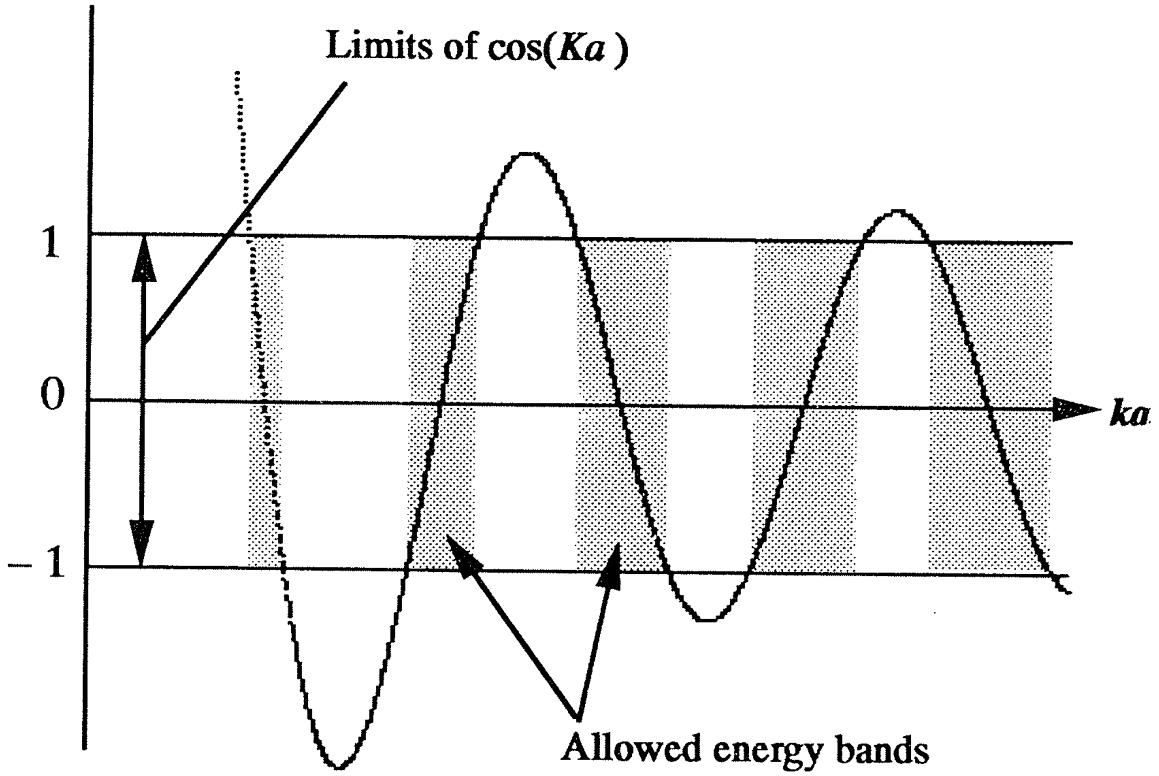


Figure 3.19 The allowed and forbidden values of ka. The shaded areas represent the allowed range of energy values.

which produces the bound-state energies of (2.48) (recall that the well width was $2a$ in the previous chapter). Thus, the approach used here does reproduce the limiting cases that we have already treated. The periodic potential breaks up the free electrons (for weak potentials) by opening gaps in the spectrum of allowed states. On the other hand, for strong potentials, the periodic tunnelling couples the wells and broadens the bound states into bands of states. These are the two limiting approaches, but the result is the same.

The ranges of values for k that lie within the limits projected by K are those of the allowed energy bands (each region of allowed solutions in figure 3.19 corresponds to one allowed energy band). In figure 3.20, we show these solutions, with all values of k restricted to the range $-\pi/a < K < \pi/a$. In solid-state physics, this range of K is termed the *first Brillouin zone* and the energy bands as shown in figure 3.20 are termed the *reduced zone scheme* (as opposed to taking K over an infinite range). We note that the momentum K (or more properly $\hbar K$) is the horizontal axis, and the energy is the vertical axis, which provides a traditional *dispersion relation* of the frequency $\omega = \mathcal{E}/\hbar$ as a function of the wave vector K.

3.8 SINGLE-ELECTRON TUNNELLING

As a last consideration in this chapter, we want to consider tunnelling through the insulator of a capacitor (which we take to be an oxide such as SiO_2 found in MOS structures). The tunnelling through the capacitor oxide is an example of a very

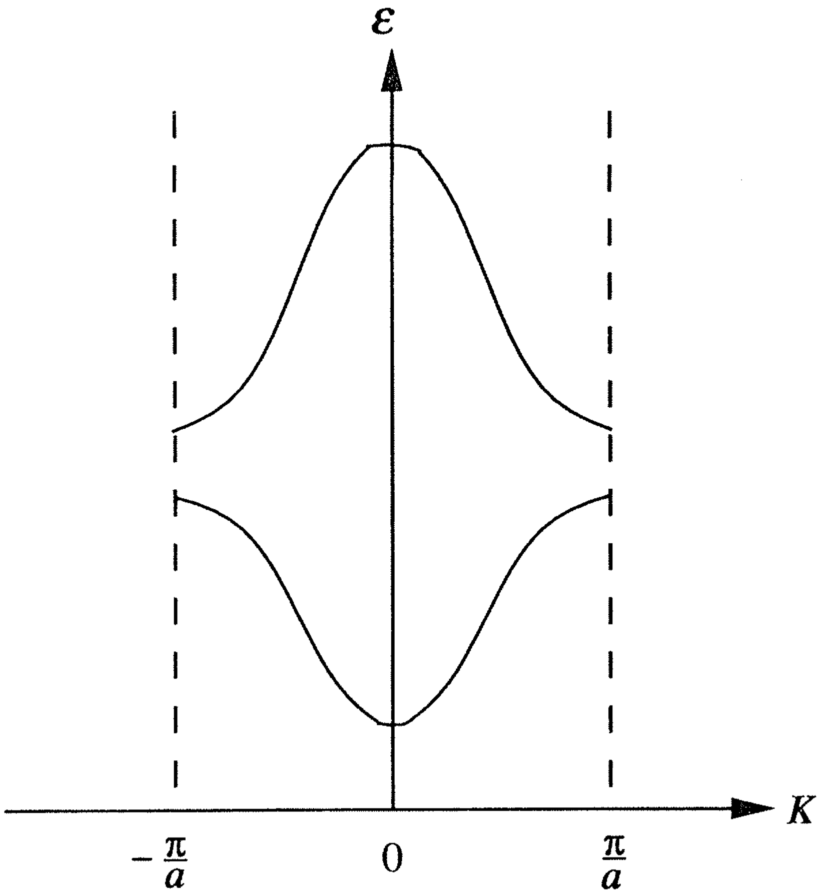


Figure 3.20 The energy band structure that results from the solution diagram of figure 3.19.

simple physical system (the single capacitor) that can exhibit quite complicated behaviour when it is made small. The capacitor is formed by placing an insulator between two metals or by the oxide in a metal–oxide–semiconductor structure, as discussed in section 2.6. Consider, for example, the tunnelling coefficient for such an insulator, in which the barrier height is approximately 3 eV, and the thickness of the insulator (assumed to be SiO_2) is about 3 nm. Although the tunnelling coefficient is small (we may estimate it to be of the order of 10^{-6}), the actual current density that can flow due to tunnelling is of the order of a few picoamperes per square centimetre. If the barriers are semiconductors, rather than metals, then the current can be two orders of magnitude larger, and, of course, the tunnelling coefficient will become much larger under a bias field which distorts the shape of the potential barrier. Thus, in general, oxide insulators of this thickness are notoriously leaky due to tunnelling currents, even though the tunnelling probability is quite low for a single electron (there are of course a great number of electrons attempting to tunnel, so even though the probability of one electron tunnelling is quite low, the number making it through is significant).

What if the area of the capacitor is made small, so that the capacitance is also quite small? It turns out that this can affect the operation of tunnelling through the oxide significantly as well. When an electron tunnels through the oxide, it

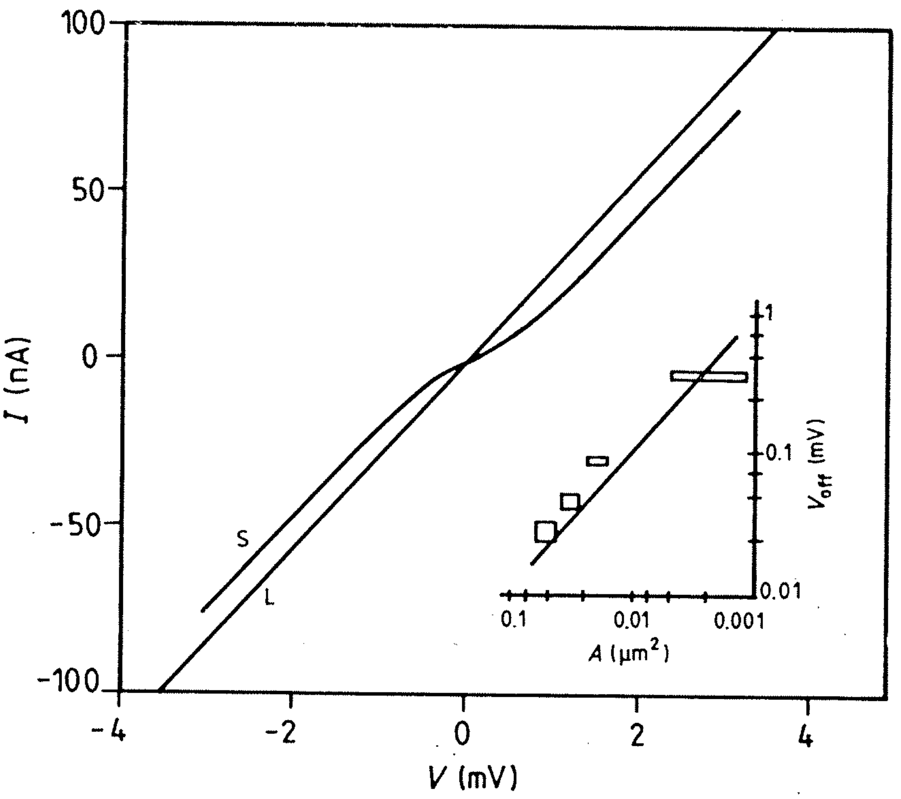


Figure 3.21 Single-electron tunnelling currents in small capacitors. The voltage offset is due to the Coulomb blockade. (After Fulton and Dolan (1987), by permission.)

lowers the energy stored in the capacitor by the amount

$$\delta\mathcal{E} = \frac{e^2}{2C}. \tag{3.116}$$

For example, the voltage across the capacitor changes by the amount

$$\delta V = \frac{e}{C}. \tag{3.117}$$

What this means is that the tunnelling current cannot occur until a voltage equivalent to (3.117) is actually applied across the capacitor. If the voltage on the capacitor is less than this, no tunnelling current occurs because there is not sufficient energy stored in the capacitor to provide the tunnelling transition. When the capacitance is large, say $> 10^{-12}$ F, this voltage is immeasurably small in comparison with the thermally induced voltages ($k_B T/e$). On the other hand, suppose that the capacitance is defined with a lateral dimension of only 50 nm. Then, the area is 2.5×10^{-15} m^2, and our capacitor discussed above has a capacitance of 2.8×10^{-17} F, and the required voltage of (3.117) is 5.7 mV. These capacitors are easily made, and the effects easily measured at low temperatures. In figure 3.21, we show measurements by Fulton and Dolan (1987) on such structures. The retardation of the tunnelling current until a voltage according to (3.117) is reached is termed the *Coulomb blockade*. The name arises from the need to have sufficient Coulomb energy before the tunnelling transition can occur. The Coulomb blockade causes the offset of the current in the small- (S) capacitor case. This offset scales with area, as shown in the inset to the figure, and hence with C as expected in (3.117).

3.8.1 Bloch oscillations

The results discussed above suggest an interesting experiment. If we pass a constant current through the small capacitor, the charge stored on the capacitor can increase linearly with time. Thus, the charge on the capacitor, due to the current, is given by

$$Q(t) = \int_0^t I \,\mathrm{d}t = It. \tag{3.118}$$

When the voltage reaches the value given by (3.117), an electron tunnels across the oxide barrier, and reduces the voltage by the amount given by this latter equation; for example, the tunnelling electron reduces the voltage to zero. The time required for this to occur is just the period T, defined by

$$T = \frac{Q}{I} = \frac{e}{I} = \frac{2\pi}{\omega_\mathrm{B}} \tag{3.119}$$

which defines the Bloch frequency. As we will see, this relates to the time required to cycle through a periodic band structure, such as those discussed in the previous section. Many people have tried to measure this oscillation, but (to date) only indirect inferences as to its existence have been found.

The voltage that arises from the effects described above can be stated as Q/C, where Q is measured by (3.118) modulo e. Here, $Q(t)$ is the instantaneous charge that arises due to the constant current bias, while e is the electronic charge. The charge on the capacitor, and therefore the voltage across the capacitor, rises linearly until the energy is sufficient to cover the tunnelling transition. At this point the charge drops by e, and the voltage decreases accordingly.

This behaviour is very reminiscent of that in periodic potentials, where a Bloch band structure and Brillouin zones are encountered. Consider the band structure in figure 3.20, for example. If we apply a constant electric field to the solid represented by this band structure, then the momentum responds according to

$$\hbar\frac{\mathrm{d}k}{\mathrm{d}t} = eF \tag{3.120}$$

and

$$k = -\frac{\pi}{a} + \left[\frac{eFt}{\hbar}\right]_{\mathrm{mod}(2\pi/a)}. \tag{3.121}$$

The meaning of (3.121) is that the magnitude of the wave vector k increases linearly with electric field, and when it reaches the zone boundary at π/a

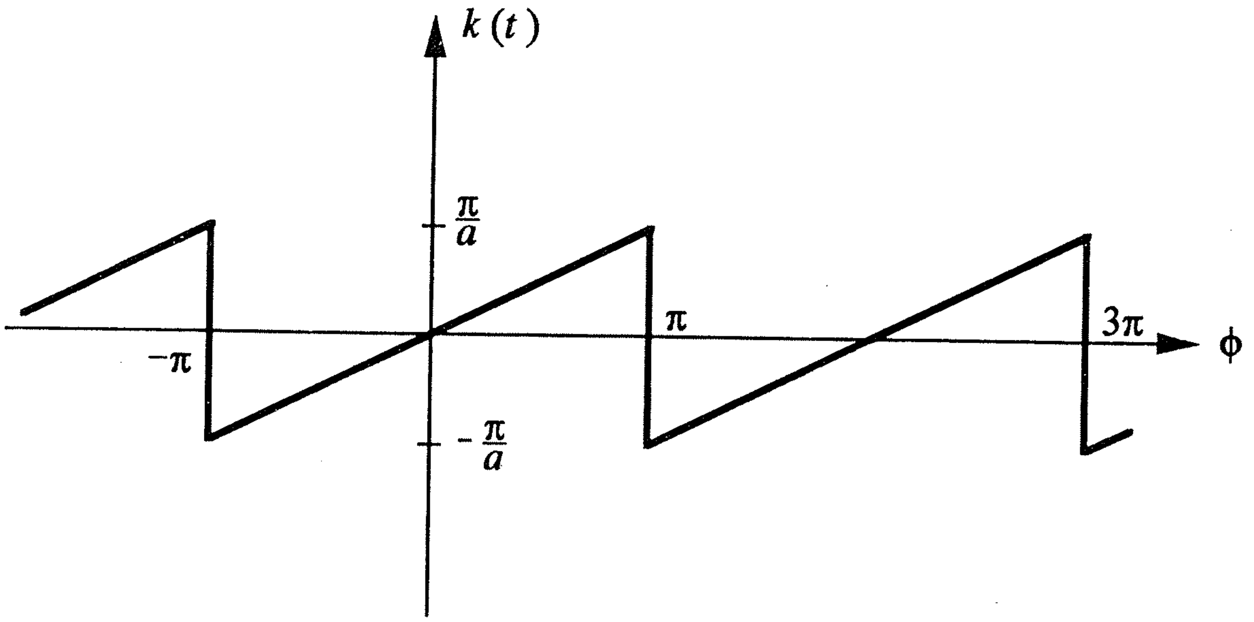


Figure 3.22 The variation in wave vector (or charge) in a periodic potential under the action of a constant electric field. The phase is $\phi = \omega_B t = (eFat)/\hbar$.

it is Bragg reflected back to $-\pi/a$, from where it is again continuously accelerated across the Brillouin zone. (Of course, this is in the reduced zone scheme for the momentum.) This behaviour is shown in figure 3.22, where $\phi = (eFat)/\hbar = \omega_B t$ is defined to be the phase, and ω_B is the Bloch frequency. If we connect the phase with It/e, and offset the charge by the amount $-e/2$, then this figure also describes the behaviour of the charge in the capacitor as described in the previous paragraph. Can we say that the charge is Bragg reflected at $\pm e/2$?

3.8.2 Periodic potentials

The drop in charge, given by the tunnelling of the electron through the small capacitor, does not occur with sudden sharpness, when we operate at a non-zero temperature. Thus, it is possible to approximate the result of (3.118), and figure 3.22, by the expression for the charge on the capacitor as

$$Q(t) = \frac{e}{2}\sin(\omega_B t) \qquad (3.122)$$

which symmetrizes the charge about zero (for zero current bias), and the change occurs now when the instantaneous charge reaches half-integer charge (dropping to the negative of this value so that the net tunnelling charge is a single electron). Now, we want to create a Hamiltonian system, which we can quantize, to produce the effective periodic potential structures of the previous section. For this, we define the *phase* of the charge to be

$$\phi = \omega_B t. \qquad (3.123)$$

We take this phase to have the equivalent coordinate of position given in the previous section (we will adjust it below by a constant), and this means that we can extend the treatment to cases in which there is not a constant current bias applied. Rather, we assert that the phase behaves in a manner that describes some equivalent position. The position is not particularly important for periodic potentials and does not appear at all in figure 3.20 for the band structure. We now take the *momentum* coordinate to correspond to

$$-\mathrm{i}e\frac{\partial}{\partial\phi} = Q. \tag{3.124}$$

Now, this choice is not at all obvious, but it is suggested by the comparison above between the time behaviour of the charge, under a constant current bias, and the time behaviour of the crystal momentum, under a constant electric field bias. The independent variable that describes the state of the capacitor is the charge Q. From the charge, we determine the energy stored in the capacitor, which is just $Q^2/2C$. If we think of the capacitance C playing the role of the mass in (3.1), we can think of the charge as being analogous to the momentum $\hbar k$. Then the energy on the capacitor is just like the kinetic energy in a parabolic band for free electrons. The relationship (3.122) reflects a periodic potential which will open gaps in the free-electron spectrum, and these gaps occur when $Q = \pm e/2$ (a total charge periodicity of e), just as they occur at $k = \pi/a$ for electrons in a periodic potential. In fact, the zone edges occur for the free electrons when $ka = n\pi$. Thus, the quantity kx plays the role of a *phase* with boundaries at $x = \pm a$.

Since we now have a momentum, and a coordinate resembling a 'position', we can develop the commutator relationship (1.22), but for the 'correct' answer, we need to scale the phase by the factor $\hbar/e$. Then,

$$[Q, (\hbar/e)\phi] = -\mathrm{i}\hbar. \tag{3.125}$$

This suggests that the correct position variable, which is now conjugate to the charge, is just $(\hbar/e)\phi$.

The time derivative of the momentum is just Newton's law, and this can lead us to the proper potential energy term to add to the kinetic energy to obtain the total energy. We use

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = F = -\frac{\partial V}{\partial x}. \tag{3.126}$$

Thus, the potential energy is just

$$V(\phi) \sim -\frac{\hbar\omega_{\mathrm{B}}}{2}\int \cos(\phi)\,\mathrm{d}\phi \sim \frac{\hbar\omega_{\mathrm{B}}}{2}[1 - \sin(\phi)] \tag{3.127}$$

and the constant term has been artificially adjusted, as will be discussed below. Now, we want to compare this with the periodic potential shown in figure 3.18.

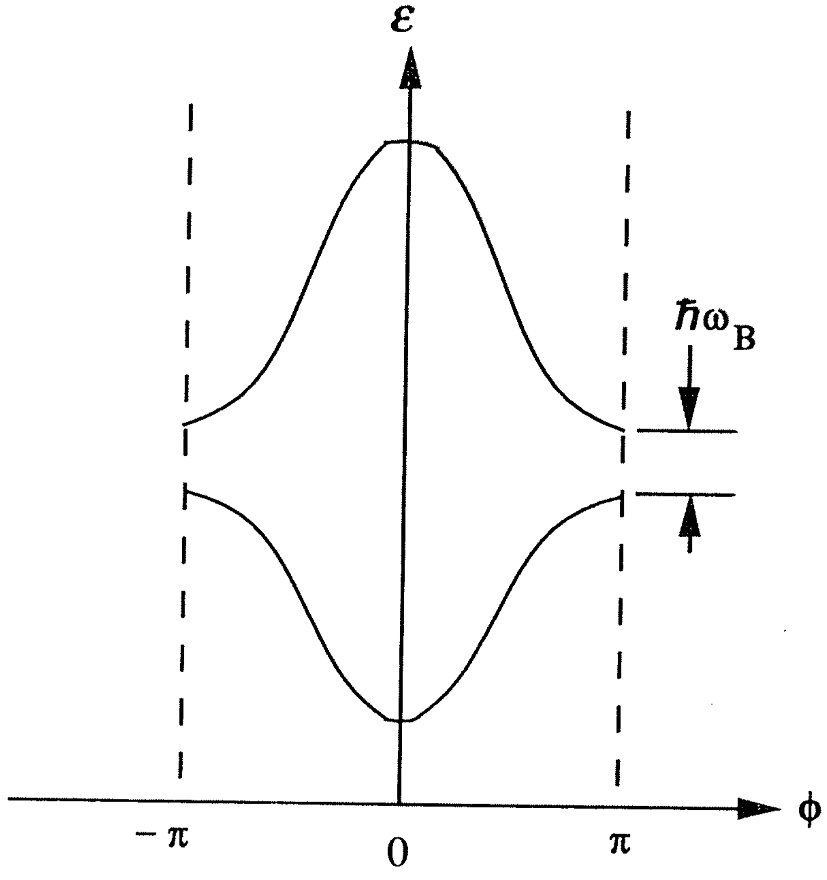


Figure 3.23 The energy band spectrum for the single tunnelling capacitor.

It is possible to expand the potential in figure 3.18 in a Fourier series, and it is obvious that (3.127) is just the lowest-order term in that expansion. It is also possible to expand the charge behaviour of figure 3.22 in a Fourier series and (3.122) is the lowest term in that expansion. Thus, the potential of (3.127) and the charge of (3.122) both correspond to the simplest periodic potential, which is just the lowest Fourier term of any actual potential. The constant term in (3.127) has been defined as just half of the height of the potential, so the sum of the constant and the sine term corresponds to the peak of the potential, and the difference corresponds to the zero-potential region of figure 3.18. For reference, the value of phase $\phi = \pi/2$ corresponds to $x = 0$ in figure 3.18. Now, in periodic potentials, there is a symmetry in the results, which must be imposed onto this problem, and this arises from the fact that we should have used $\pm Q$ in (3.126), which leads to the adjusted potential

$$V(\phi) = \frac{\hbar\omega_{\mathrm{B}}}{2}[1 \pm \sin(\phi)] \tag{3.128}$$

which shifts the $x = 0$ point to $\phi = \pm\pi/2$. The leading term in the potential just offsets the energy, and can be ignored. The Hamiltonian is then

$$H = \frac{Q^2}{2C} \pm \frac{\hbar\omega_{\mathrm{B}}}{2}\sin(\phi). \tag{3.129}$$

This Hamiltonian is in a mixed position and momentum representation. The energy can be written out if we use just a momentum representation, and this is achieved by using (3.122) to eliminate the phase, as

$$\mathcal{E} = \frac{Q^2}{2C} \pm \frac{\hbar\omega_{\mathrm{B}}}{2}\frac{2Q}{e}. \tag{3.130}$$

This energy is shown in figure 3.23. The zone boundaries are at the values of charge $Q = \pm e/2$. At these two points, gaps open in the 'free-electron' energy (the first term in the above equation). These gaps are of $\hbar\omega_{\mathrm{B}}$. We note also that the energy bands have all been offset upward by the constant potential term, which we ignored in (3.130). This is also seen in the Kronig–Penney model treated in the previous section.

The simple capacitor seems to exhibit the very complicated behaviour expected from a periodic potential merely when it is made sufficiently small that tunnelling through the oxide can occur (and the capacitance is sufficiently small that the energy is large compared with the thermal energy). In reality, no such periodic potential exists, but the very real behaviour of the charge, which is represented in figure 3.22, gives rise to physical behaviour equivalent to that of a periodic potential. Thus, we can use the equivalent band structure of figure 3.23 to investigate other physical effects, all of which have their origin in the strange periodic behaviour of the charge on the capacitor.

REFERENCES

Brillouin L 1926 *C. R. Acad. Sci.* **183** 24

Fulton and Dolan 1987 *Phys. Rev. Lett.* **59** 109

Kramers H A 1926 *Z. Phys.* **39** 828

Landauer R 1957 *IBM J. Res. Dev.* **1** 223

Landauer R 1970 *Phil. Mag.* **21** 863

Sollner T L C G, Goodhue W D, Tannenwald P E, Parker C D and Peck D D 1983 *Appl. Phys. Lett.* **43** 588

van Wees B J, van Houten H, Beenakker C W J, Williamson J G, Kouwenhouven L P, van der Marel D and Foxon C T 1988 *Phys. Rev. Lett.* **60** 848

Wentzel G 1926 *Z. Phys.* **38** 518

PROBLEMS

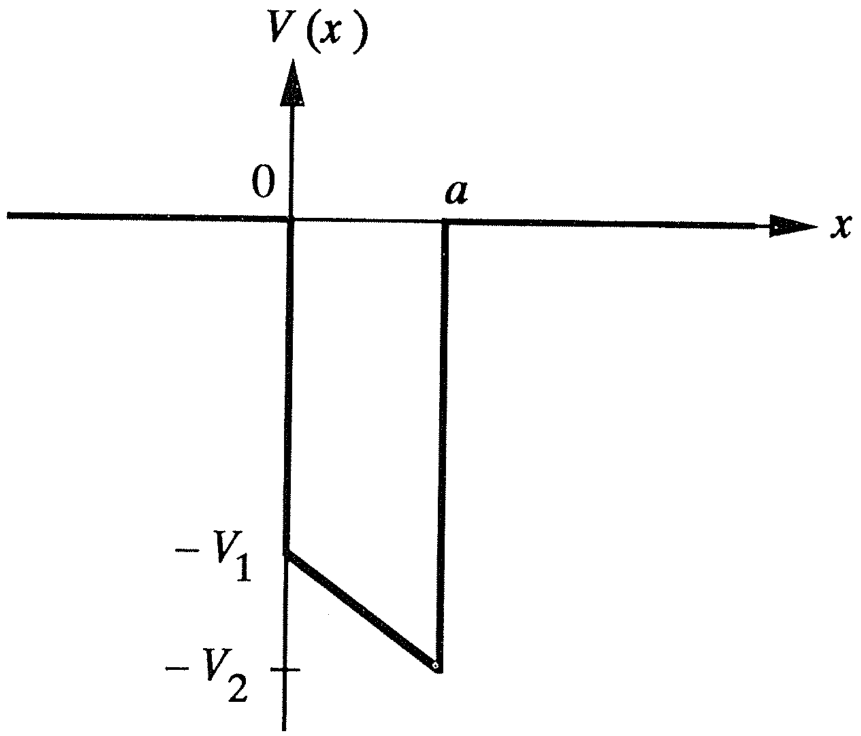
1. For a potential barrier with $V(x) = 0$ for $x > |a/2|$, and $V(x) = 0.3$ eV for $x < |a/2|$, plot the tunnelling probability for $\mathcal{E}$ in the range 0–0.5 eV. Take the value $a = 5$ nm and use the effective mass of GaAs, $m^* = 6.0 \times 10^{-32}$ kg.

2. For a potential barrier with $V(x) = 0$ for $x > |a/2|$, and $V(x) = 0.4$ eV for $x < |a/2|$, plot the tunnelling probability for $\mathcal{E}$ in the range 0–0.5 eV. Take the value $a = 5$ nm and use the effective mass of GaAs, $m^* = 6.0 \times 10^{-32}$ kg.

3. Consider the potential barrier discussed in problem 1. Suppose that there are two of these barriers forming a double-barrier structure. If they are separated by 4 nm, what are the resonant energy levels in the well? Compute the tunnelling probability for transmission through the entire structure over the energy range 0–0.5 eV.

4. Suppose that we create a double-barrier resonant tunnelling structure by combining the barriers of problems 1 and 2. Let the barrier with $V_0 = 0.3$ eV be on the left, and the barrier with $V_0 = 0.4$ eV be on the right, with the two barriers separated by a well of 4 nm width. What are the resonant energies in the well? Compute the tunnelling probability through the entire structure over the energy range 0–0.5 eV. At an energy of 0.25 eV, compare the tunnelling coefficient with the ratio of the tunnelling coefficients (at this energy) for the barrier of problem 2 over that of problem 1 (i.e. the ratio $T_{\mathrm{min}}/T_{\mathrm{max}}$).

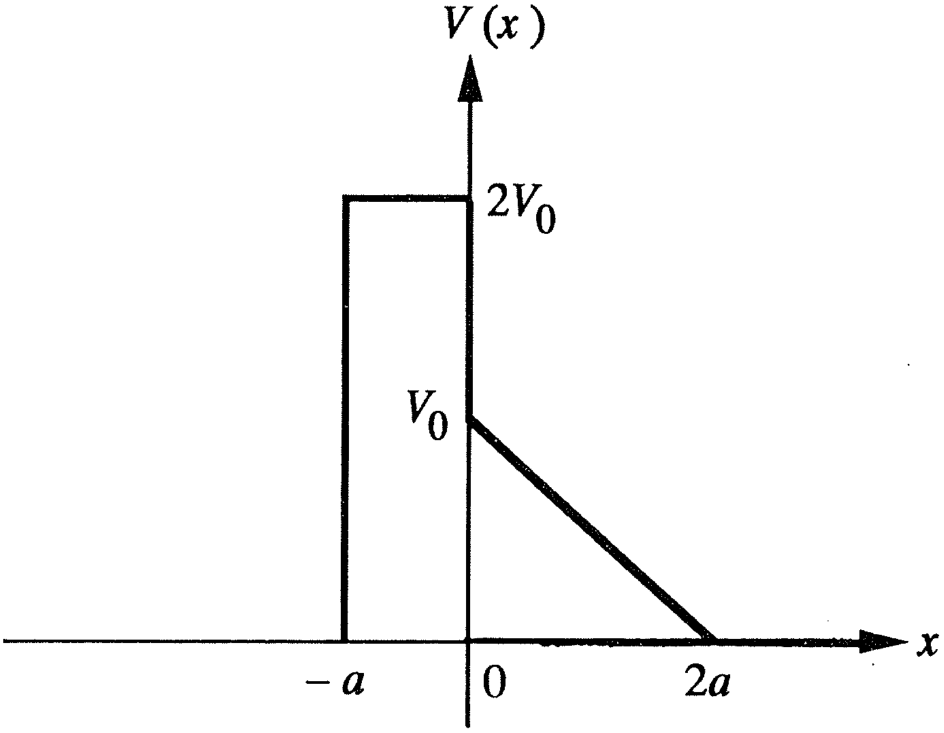
5. Let us consider a trapezoidal potential well, such as that shown in the figure below. Using the WKB method, find the bound states within the well. If $V_1 = 0.3$ eV, $V_2 = 0.4$ eV, and $a = 5$ nm, what are the bound-state energies?



6. A particle is contained within a potential well defined by $V(x) \to \infty$

for $x < 0$ and $V(x) = \alpha x$ for $x > 0$. Using the WKB formula, compute the bound-state energies. How does the lowest energy level compare to that found in (2.78) ($\alpha = eE$)?

7. Consider the tunnelling barrier shown below. Using the WKB form for the tunnelling probability $T(\mathcal{E})$, calculate the tunnelling coefficient for $\mathcal{E} = V_0/2$.



8. A particle moves in the potential well $V(x) = ax^4$. Calculate the bound states with the WKB approximation.

9. In the WKB approximation, show that the tunnelling probability for a double barrier (well of width b, barriers of width $2a$, as shown in figure 3.4, and a height of each barrier of V_0) is given by

$$T = \frac{4}{\left(4\theta^2 + 1/(4\theta^2)\right)\cos^2 L + 4\sin^2 L}$$

where

$$\theta = \exp\left(\int_b^{b+2a} \gamma(x)\,\mathrm{d}x\right)$$

and

$$L = \int_0^b k(x)\,\mathrm{d}x.$$

What value must b have so that only a single resonant level exists in the well?

10. In (3.114), the values for which the right-hand side reach -1 must be satisfied by the left-hand side having $\cos(ka) = -1$, which leads to the energies being those of an infinite potential well. Show that this is the case. Why? The importance of this result is that the top of every energy band lies at an energy defined by the infinite potential well, and the bands form by spreading *downward* from these energies as the coupling between wells is increased.