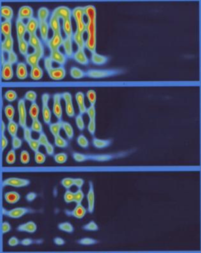
# QUANTUM MECHANICS

## AN INTRODUCTION FOR DEVICE PHYSICISTS AND ELECTRICAL ENGINEERS



SECOND EDITION

DAVID K FERRY

IoP

# Quantum Mechanics

## An Introduction for Device Physicists and Electrical Engineers

## Second Edition

## David K Ferry

This textbook provides a complete course in quantum mechanics for students of semiconductor device physics and electrical engineering. It provides the necessary background to quantum theory for those starting work on micro- and nanoelectronic structures and is particularly useful for those going on to work with semiconductors and lasers.

This book was developed from a course the author has taught for many years with a style and order of presentation of material specifically designed for this audience. It introduces the main concepts of quantum mechanics which are important in everyday solid-state physics and electronics. Each topic includes examples which have been carefully chosen to draw upon relevant experimental research. It also includes problems with solutions to test understanding of theory. For the second edition significant new material, suggested by colleagues, has been added to each chapter, providing updated connections with relevant experiments and device concepts. New references and new problems are included.

The book is intended for undergraduate and graduate students who have knowledge of semiconductor materials, linear vector spaces, and electromagnetic field theory.

**About the author**
David K Ferry is Regents' Professor of Engineering at Arizona State University. He is a Fellow of the IEEE and the APS. He received the 1999 IEEE Cledo Brunetti Award. He is the author, or co-author of more than 350 refereed publications, including three textbooks.

Front cover illustration: a fully self-consistent calculation of the electron density in a quantum wire in the region around a quantum point contact. The calculation was done by Richard Akis and Lucian Shifren, Department of Electrical Engineering, Arizona State University.

ISBN 0-7503-0725-0

9 780750 307253

Institute of Physics Publishing
Bristol and Philadelphia
www.iop.org

# Quantum Mechanics

## An Introduction for Device Physicists and Electrical Engineers

**Second Edition**

David K Ferry

*Arizona State University,*
*Tempe, USA*

IoP

Institute of Physics Publishing
Bristol and Philadelphia

*British Library Cataloguing-in-Publication Data*

A catalogue record for this book is available from the British Library.

ISBN 0 7503 0725 0

*Library of Congress Cataloging-in-Publication Data are available*

Commissioning Editor: James Revill
Publisher: Nicki Dennis
Production Editor: Simon Laurenson
Production Control: Sarah Plenty
Cover Design: Victoria Le Billon
Marketing Executive: Colin Fenton

Published by Institute of Physics Publishing, wholly owned by The Institute of Physics, London

Institute of Physics Publishing, Dirac House, Temple Back, Bristol BS1 6BE, UK

US Office: Institute of Physics Publishing, The Public Ledger Building, Suite 1035, 150 South Independence Mall West, Philadelphia, PA 19106, USA

Typeset in TEX using the IOP Bookmaker Macros
Printed in the UK by J W Arrowsmith Ltd, Bristol

# Contents

# Preface to the first edition

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

# Preface to the second edition

Many of my friends have used the first edition of this book, and have suggested a number of changes and additions, not to mention the many errata necessary. In the second edition, I have tried to incorporate as many additions and changes as possible without making the text over-long. As before, there remains far more material than can be covered in a single one-semester course, but the additions provide further discussion on many topics and important new additions, such as numerical solutions to the Schrödinger equation. We continue to use this book in such a one-semester course, which is designed for fourth-year electrical engineering students, although more than half of those enrolled are first-year graduate students taking their first quantum mechanics course.

I would like to express my thanks in particular to Dragica Vasileska, who has taught the course several times and has been particularly helpful in pointing out the need for additional material that has been included. Her insight into the interpretations has been particularly useful.

**David K Ferry**
Tempe, AZ, 2000

# Chapter 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 300 years of classical mechanics.

In contrast with this, the ideas of quantum mechanics are barely more than a century old. They had their first beginnings in the 1890s with Planck's development of a theory for black-body radiation. This form of radiation is emitted by all bodies according to their temperature. However, before Planck, there were two competing views. In one, the low-frequency view, this radiation increased as a power of the frequency, which led to a problem at very high frequencies. In the other, the high-frequency view, the radiation decreased rapidly with frequency, which led to a problem at low frequencies. Planck unified these views through the development of what is now known as the Planck black-body radiation law:

$$I(f)\,\mathrm{d}f \sim \frac{f^3}{\exp\left(\frac{hf}{k_\mathrm{B}T}\right)-1}\,\mathrm{d}f \tag{1.1}$$

where $f$ is the frequency, $T$ is the temperature, $I$ is the intensity of radiation, and $k_\mathrm{B}$ is Boltzmann's constant $(1.38 \times 10^{-23}\ \mathrm{J\ K}^{-1})$. In order to achieve this result, Planck had to assume that matter radiated and absorbed energy in small, but non-zero quantities whose energy was defined by

$$E = hf \tag{1.2}$$

where $h$ is now known as Planck's constant, given by $6.62 \times 10^{-23}$ J s. While Planck had given us the idea of quanta of energy, he was not comfortable with this idea, but it took only a decade for Einstein's theory of the photoelectric effect (discussed later) to confirm that radiation indeed was composed of quantum particles of energy given by (1.2). Shortly after this, Bohr developed his quantum model of the atom, in which the electrons existed in discrete shells with well defined energy levels. In this way, he could explain the discrete absorption and emission lines that were seen in experimental atomic spectroscopy. While his model was developed in a somewhat *ad hoc* manner, the ideas proved correct, although the mathematical details were changed when the more formal quantum theory arrived in 1927 from Heisenberg and Schrödinger. The work of these two latter pioneers led to different, but equivalent, formulations of the quantum principles that we know to be important in modern physics. Finally, another essential concept was introduced by de Broglie. While we have assigned particle-like properties to light waves earlier, de Broglie asserted that particles, like electrons, should have wavelike properties in which their wavelength is related to their momentum by

$$\lambda = \frac{h}{p} = \frac{h}{m\boldsymbol{v}}. \tag{1.3}$$

$\lambda$ is now referred to as the de Broglie wavelength of the particle.

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, $h$, 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 do not 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

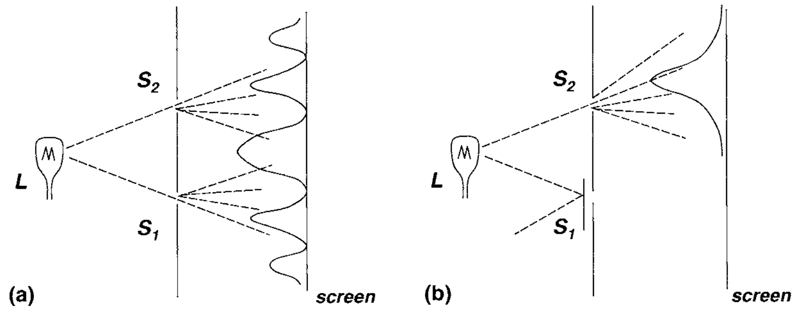


**Figure 1.1.** In panel (*a*), we illustrate how light coming from the source L and passing through the two slits $S_1$ and $S_2$ interferes to cause the pattern indicated on the 'screen' on the right. If we block one of the slits, say $S_1$, then we obtain only the light intensity passing through $S_2$ on the 'screen' as shown in panel (*b*).

small, $6.625 \times 10^{-34}$ 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.

Before proceeding, let us examine an important aspect of light as a wave. If we create a source of coherent light (a single frequency), and pass this through two slits, the wavelike property of the light will create an interference pattern, as shown in figure 1.1. Now, if we block one of the slits, so that light passes through just a single slit, this pattern disappears, and we see just the normal passage of the light waves. It is this interference between the light, passing through two different paths so as to create two different phases of the light wave, that is an essential property of the single wave. When we can see such an interference pattern, it is said that we are seeing the wavelike properties of light. To see the particle-like properties, we turn to the photoelectric effect.

## 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 (1.2) (Einstein 1905)

$$\mathcal{E} = h\nu = \hbar\omega \tag{1.2$'$}$$

where $\hbar = h/2\pi$. The photoelectric effect can be understood through consideration of figure 1.2. 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. 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 $\mathcal{E}_{\mathrm{W}}$, 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 $\mathcal{E} > \mathcal{E}_{\mathrm{W}}$. 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 is 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 (1.3) given earlier

$$p = \frac{h\nu}{c} = \frac{h}{\lambda}. \tag{1.3$'$}$$

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

$$p = \frac{h}{\lambda} = \hbar k \tag{1.4}$$

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

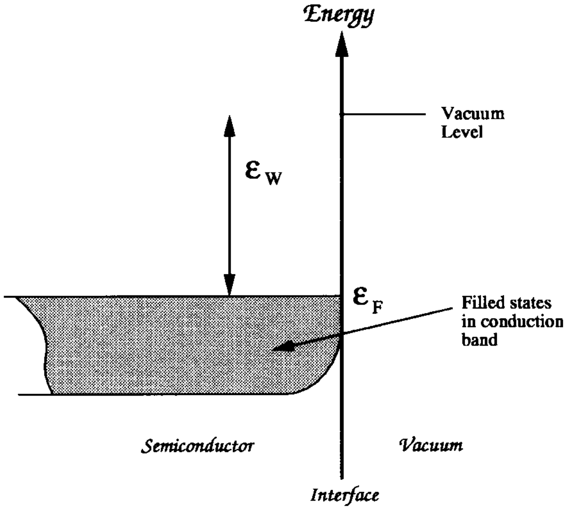


**Figure 1.2.** The energy bands for the surface of a metal. An incident photon with an energy greater than the work function, $\mathcal{E}_W$, can cause an electron to be raised from the Fermi energy, $\mathcal{E}_F$, to above the vacuum level, whereby it can be photoemitted.

of the wave. The two equations (1.2′) and (1.3′) give these relationships. The form of (1.3′) 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 (1.2′) 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 wavelike 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

$$\Psi(x,t) = \Psi_1(x,t) + \Psi_2(x,t). \tag{1.5}$$

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 (1.4) that the momentum of the particles goes immediately into the so-called *wave vector* $\boldsymbol{k}$ of the wave. A special form of (1.5) is

$$\Psi(x,t) = A\mathrm{e}^{\mathrm{i}(k_1 x - \omega t)} + B\mathrm{e}^{\mathrm{i}(k_2 x - \omega t)} \tag{1.6}$$

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 (1.6) 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 $\mathrm{Si_3N_4}$ substrate. Such a loop is shown in figure 1.3, 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

$$k = k_0 - \frac{e}{\hbar}\int E\,\mathrm{d}t. \tag{1.7}$$

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 $\pi$ down to 0)

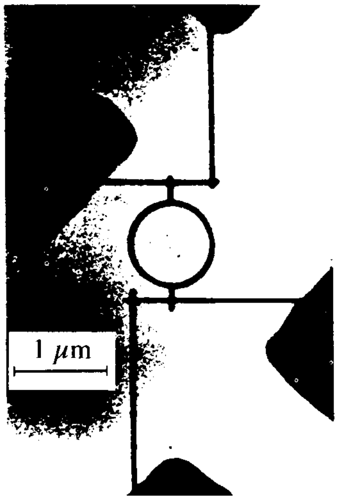


**Figure 1.3.** Transmission electron micrograph of a large-diameter (820 nm) polycrystalline Au ring. The lines are about 40 nm wide and about 38 nm thick. (After Washburn and Webb (1986), by permission.)

and the integral over the bottom half of the ring (from $-\pi$ up to 0):

$$\begin{aligned}\Delta\phi &= -\frac{e}{\hbar}\int \mathrm{d}t\left(\int_{\pi}^{0} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l} + \int_{-\pi}^{0} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l}\right) = -\frac{e}{\hbar}\int \mathrm{d}t\int_{0}^{2\pi} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l} \\ &= -\frac{e}{\hbar}\int \mathrm{d}t\int \boldsymbol{\nabla}\times\boldsymbol{E}\cdot\boldsymbol{n}\,\mathrm{d}A = \frac{e}{\hbar}\int \boldsymbol{B}\cdot\boldsymbol{n}\,\mathrm{d}A = 2\pi\frac{\Phi}{\Phi_0} \qquad (1.8)\end{aligned}$$

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 1.4($a$), the conductance through the ring of figure 1.3 is shown. There is a strong oscillatory behaviour as the magnetic field coupled by the ring is varied. The curve of figure 1.4($b$) 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 $\Phi_0$. 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

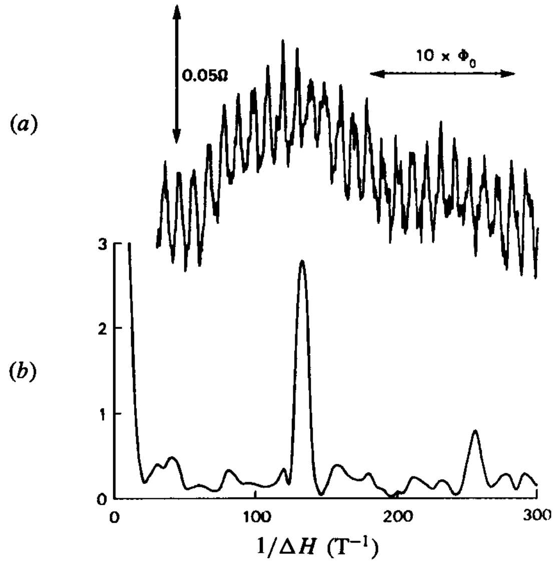


**Figure 1.4.** Conductance through the ring of figure 1.3. In (*a*), the conductance oscillations are shown at a temperature of 0.04 K. The Fourier transform is shown in (*b*) and gives clearly evidence of the dominant $h/e$ period of the oscillations. (After Washburn and Webb (1986), by permission.)

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 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.

The preceding paragraphs describe how we can 'measure' the phase interference between the electron *waves* passing through two separate arms of the system. In this regard, these two arms serve as the two *slits* for the optical waves of figure 1.1. Observation of the interference phenomena shows us that the electrons must be considered as waves, and not as particles, for this experiment. Once more, we have a confirmation of the *correspondence* between waves and particles as two views of a coherent whole. In the preceding experiment, the magnetic field was used to vary the phase in both arms of the interferometer and induce the oscillatory behaviour of the conductance on the magnetic field. It is also possible to vary the phase in just one arm of the interferometer by the use of a tuning gate (Fowler 1985). Using techniques which will be discussed in the following chapters, the gate voltage varies the propagation wave vector $k$ in one arm of the interferometer, which will lead to additional oscillatory conductance as this voltage is tuned, according to (1.7) and (1.8), as the electric field itself is varied instead of using the magnetic field. A particularly ingenious implementation of this interferometer has been developed by Yacoby *et al* (1994), and will be discussed in later chapters once we have discussed the underlying physics.

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

$$|\Psi(x,t)|^2 \tag{1.9}$$

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

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

While (1.10) 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):

$$\lim_{L\to\infty}\int_{-L/2}^{L/2}|\Psi(x,t)|^2\,\mathrm{d}x = 1. \tag{1.11}$$

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.

There are additional constraints which we wish to place upon the wave function. The first is that the system is linear, and satisfies superposition. That is, if there are two physically realizable states, say $\psi_1$ and $\psi_2$, then the total wave function must be expressable by the linear summation of these, as

$$\Psi(x,t) = c_1\psi_1(x,t) + c_2\psi_2(x,t). \tag{1.12}$$

Here, $c_1$ and $c_2$ are arbitrary complex constants, and the summation represents a third, combination state that is physically realizable. Using (1.12) in the probability requirement places additional load on these various states. First, each $\psi_i$ must be normalized independently. Secondly, the constants $c_i$ must now satisfy (1.10) as

$$\begin{aligned}\int_{-\infty}^{\infty}|\Psi(x,t)|^2\,\mathrm{d}x = 1 = |c_1|^2\int_{-\infty}^{\infty}|\psi_1(x,t)|^2\,\mathrm{d}x + |c_2|^2\int_{-\infty}^{\infty}|\psi_1(x,t)|^2\,\mathrm{d}x\\ = |c_1|^2 + |c_2|^2.\end{aligned} \tag{1.13}$$

In order for the last equation to be correct, we must apply the third requirement of

$$\int_{-\infty}^{\infty}\psi_1^*(x,t)\psi_2(x,t)\,\mathrm{d}x = \int_{-\infty}^{\infty}\psi_2^*(x,t)\psi_1(x,t)\,\mathrm{d}x = 0 \tag{1.14}$$

which is that the individual states are *orthogonal* to one another, which must be the case for our use of the composite wave function (1.12) to find the probability.

### 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

$$\langle x\rangle = \int_{-\infty}^{\infty}x|\Psi(x,t)|^2\,\mathrm{d}x = \int_{-\infty}^{\infty}\Psi^*(x,t)x\Psi(x,t)\,\mathrm{d}x. \tag{1.15}$$

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 (1.10) to define the inner product of the wave function and its complex conjugate, then this may be described in the short-hand notation

$$(\Psi, \Psi) = \int_{-\infty}^{\infty} \Psi^*(x,t)\Psi(x,t)\,\mathrm{d}x = 1 \tag{1.16}$$

and

$$\langle x \rangle = (\Psi, x\Psi). \tag{1.17}$$

Before proceeding, it is worthwhile to consider an example of the expectation value of the wave function. Consider the Gaussian wave function

$$\Psi(x,t) = A\exp(-x^2/2)\mathrm{e}^{-\mathrm{i}\omega t}. \tag{1.18}$$

We first normalize this wave function as

$$\int_{-\infty}^{\infty} |\Psi(x,t)|^2\,\mathrm{d}x = A^2 \int_{-\infty}^{\infty} \exp(-x^2)\,\mathrm{d}x = A^2\sqrt{\pi} = 1 \tag{1.19}$$

so that $A = \pi^{-1/4}$. Then, the expectation value of position is

$$\begin{aligned}\langle x \rangle &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-x^2/2)x\exp(-x^2/2)\,\mathrm{d}x \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} x\mathrm{e}^{-x^2}\,\mathrm{d}x = 0.\end{aligned} \tag{1.20}$$

Our result is that the average position is at $x = 0$. On the other hand, the expectation value of $x^2$ is

$$\begin{aligned}\langle x^2 \rangle &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-x^2/2)x^2\exp(-x^2/2)\,\mathrm{d}x \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2\mathrm{e}^{-x^2}\,\mathrm{d}x = \frac{1}{2}.\end{aligned} \tag{1.21}$$

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

$$x\Psi(x,t) = \underline{x}\Psi(x,t) \tag{1.22}$$

where $\underline{x}$ is the eigenvalue of $x$ operating on $\Psi(x,t)$. It is clear that the use of (1.22) in (1.7) means that the eigenvalue $\underline{x} = \langle x \rangle$.

We 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 1.5. 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 $x_{\text{peak}}$. 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 1.5, and this is the width of the wave function, which relates to the standard deviation of the wave function. Thus, we can define

$$(\Delta x)^2 = (\Psi, (x - \langle x \rangle)^2 \Psi). \tag{1.23}$$

For our example wave function of (1.18), we see that the uncertainty may be expressed as

$$\Delta x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{\frac{1}{2} - 0} = \frac{1}{\sqrt{2}}. \tag{1.24}$$

The quantity $\Delta x$ 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 $\Delta x$ 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

$$\Psi(x,0) = \delta(x - x_{\text{peak}}). \tag{1.25}$$

The time variable has been set to zero here for convenience, but it is easy to extend (1.25) to the time-varying case. Clearly, equation (1.25) describes the wave

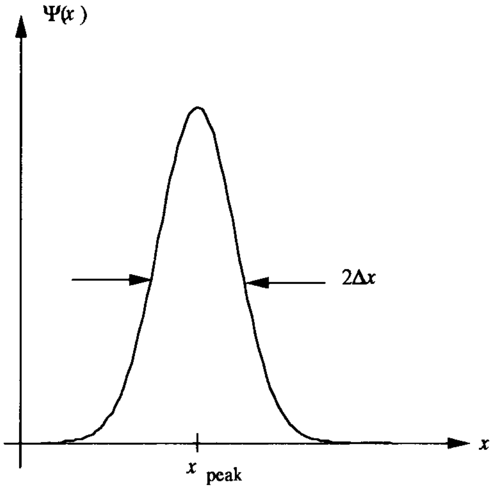


**Figure 1.5.** The positional variation of a typical wave function.

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 1.5 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:

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}kx}\, \mathrm{d}k. \tag{1.26}$$

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 (1.4). That is, the spatial frequency is described by the *wave vector* itself, which in turn is related to the momentum through (1.4). 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

$$\langle p\rangle = \hbar(\phi, k\phi) = \hbar\int_{-\infty}^{\infty}\phi^* k\phi\,\mathrm{d}k. \tag{1.27}$$

As an example of momentum wave functions, we consider the position wave function of (1.18). We find the momentum wave function from

$$\begin{aligned}\phi(k,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Psi(x,t)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x = \frac{1}{\sqrt{2}\pi^{3/4}}\int_{-\infty}^{\infty}\mathrm{e}^{-x^2/2-\mathrm{i}kx}\,\mathrm{d}x\\ &= \frac{1}{\sqrt{2}\pi^{3/4}}\mathrm{e}^{-k^2/2}\int_{-\infty}^{\infty}\exp\left(-\frac{(x+\mathrm{i}k)^2}{2}\right)\,\mathrm{d}x = \frac{1}{\pi^{1/4}}\mathrm{e}^{-k^2/2}.\end{aligned} \tag{1.28}$$

This has the same form as (1.18), so that we can immediately use (1.20) and (1.21) to infer that $\langle k\rangle = 0$ and $\langle k^2\rangle = \frac{1}{2}$.

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 (1.27) and introduce the Fourier transforms corresponding to the functions $\phi$. Then, we may write (1.27) as

$$\begin{aligned}\langle p\rangle &= \frac{\hbar}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}k\int_{-\infty}^{\infty}\mathrm{d}x'\,\Psi^*(x')\mathrm{e}^{\mathrm{i}kx'}k\int_{-\infty}^{\infty}\mathrm{d}x\,\Psi(x)\mathrm{e}^{-\mathrm{i}kx}\\ &= \frac{\hbar}{2\mathrm{i}\pi}\int_{-\infty}^{\infty}\mathrm{d}k\int_{-\infty}^{\infty}\mathrm{d}x'\,\Psi^*(x')\mathrm{e}^{\mathrm{i}kx'}\int_{-\infty}^{\infty}\mathrm{d}x\,\mathrm{e}^{-\mathrm{i}kx}\frac{\partial}{\partial x}\Psi(x)\\ &= -\mathrm{i}\hbar\int_{-\infty}^{\infty}\mathrm{d}x'\int_{-\infty}^{\infty}\mathrm{d}x\,\Psi^*(x')\delta(x-x')\frac{\partial}{\partial x}\Psi(x)\\ &= -\mathrm{i}\hbar\int_{-\infty}^{\infty}\mathrm{d}x\,\Psi^*(x)\frac{\partial}{\partial x}\Psi(x).\end{aligned} \tag{1.29}$$

In arriving at the final form of (1.29), 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 delta function:

$$\delta(x-x') = \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}k\,\mathrm{e}^{\mathrm{i}k(x-x')}. \tag{1.30}$$

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

$$p = -\mathrm{i}\hbar\frac{\partial}{\partial x}. \tag{1.31}$$

### 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

$$[x, p] = xp - px \neq 0. \tag{1.32}$$

The left-hand side of (1.32) 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

$$-(\Psi, [x, p]\Psi) = +\mathrm{i}\hbar\left\{\left(\Psi, x\frac{\partial}{\partial x}\Psi\right) - \left(\Psi, \frac{\partial}{\partial x}x\Psi\right)\right\} = -\mathrm{i}\hbar. \tag{1.33}$$

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 (1.22) 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 (1.25), 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 $\Delta x$ in position, as indicated in figure 1.5, and there will be a corresponding uncertainty $\Delta p$ 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

$$\Psi(x) = \frac{1}{(2\pi)^{1/4}\sigma^{1/2}} \exp\left[-\frac{x^2}{4\sigma^2}\right]. \tag{1.34}$$

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

$$\langle x \rangle = \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] x \, \mathrm{d}x = 0 \tag{1.35}$$

as expected. Similarly, the uncertainty in the position is found from (1.23) as

$$\begin{aligned}(\Delta x)^2 &= \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] x^2 \, \mathrm{d}x \\ &= \frac{\sigma}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] \mathrm{d}x = \sigma^2 \end{aligned} \tag{1.36}$$

and $\Delta x = \sigma$.

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

$$\begin{aligned}\phi(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Psi(x) \mathrm{e}^{-\mathrm{i}kx} \, \mathrm{d}x \\ &= \frac{1}{\sigma^{1/2}(2\pi)^{3/4}} \mathrm{e}^{-\sigma^2 k^2} \int_{-\infty}^{\infty} \exp\left[-\frac{(x - 2\mathrm{i}\sigma^2 k)^2}{4\sigma^2}\right] \mathrm{d}x \\ &= \left(\frac{2}{\pi}\right)^{1/4} \sqrt{\sigma} \mathrm{e}^{-\sigma^2 k^2}. \end{aligned} \tag{1.37}$$

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

$$(\Delta p)^2 = \hbar^2 \sigma \sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-2\sigma^2 k^2} k^2 \, \mathrm{d}k = \frac{\hbar^2}{4\sigma^2}. \tag{1.38}$$

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

$$\Delta A \Delta B \geq \tfrac{1}{2} |\langle [A, B] \rangle| \tag{1.39}$$

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 (1.39).

To understand how a *classical* measurement problem can give a result much like an uncertainty relationship, consider the simple time-varying exponential $\mathrm{e}^{-t/\tau}$. We can find the frequency content of this very simple time variation as

$$F(\omega) = \frac{1}{1 + (\omega\tau)^2}. \tag{1.40}$$

Hence, if we want to reproduce this simple exponential with our electronics, we require a bandwidth $(\Delta\omega)$ that is at least of order $1/\tau$. That is, we require

$$\Delta\omega > \frac{1}{\tau} \Rightarrow \Delta E \Delta t > \hbar \tag{1.41}$$

where we have used $(1.2')$ to replace the angular frequency with the energy of the wave and have taken $\Delta t = \tau$. While this has significant resemblance to the quantum uncertainty principle, it is in fact a *classical* result whose only connection to quantum mechanics is through the Planck relationship. The fact that time is *not* an operator in our approach to quantum mechanics, but is simply a measure of the system progression, means that there cannot be a quantum version of (1.41).

### 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, $x_{\mathrm{peak}}$, 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 (1.26). 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

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}(kx-\omega t)} \, \mathrm{d}k. \tag{1.42}$$

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

$$\omega(k) = \omega(k_0) + (k - k_0) \left. \frac{\partial \omega}{\partial k} \right|_{k=k_0} + \cdots. \tag{1.43}$$

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 (1.43) 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 (1.43) into (1.42), and define the difference variable $u = k - k_0$. Then, (1.42) becomes

$$\Psi(x, t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}(k_0 x - \omega_0 t)} \int_{-\infty}^{\infty} \phi(u + k_0) \mathrm{e}^{\mathrm{i}(ux - \omega' ut)} \, \mathrm{d}u \tag{1.44}$$

where $\omega_0$ is the leading term in (1.43) and $\omega'$ is the partial derivative in the second term of (1.43). The higher-order terms of (1.43) 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

$$\begin{aligned}
\Psi(x, t) &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\mathrm{i}k_0(x - \omega' t)} \mathrm{e}^{\mathrm{i}(k_0 x - \omega_0 t)} \int_{-\infty}^{\infty} \phi(u + k_0) \mathrm{e}^{\mathrm{i}k_0(x - \omega' t)} \mathrm{e}^{\mathrm{i}u(x - \omega' t)} \, \mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\mathrm{i}(\omega_0 - \omega' k_0)t} \int_{-\infty}^{\infty} \phi(u + k_0) \mathrm{e}^{\mathrm{i}(u + k_0)(x - \omega' t)} \, \mathrm{d}u \\
&= \mathrm{e}^{-\mathrm{i}(\omega_0 - \omega' k_0)t} \Psi(x - \omega' t, 0).
\end{aligned} \tag{1.45}$$

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 $\omega'$. This is not surprising. The quantity $\omega'$ 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

$$v_{\mathrm{g}} = \omega' = \left.\frac{\partial \omega}{\partial k}\right|_{k=k_0}. \tag{1.46}$$

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:

$$v_{\mathrm{g}} = \frac{\hbar k_0}{m} = \frac{\partial \omega}{\partial k}. \tag{1.47}$$

If we integrate the last two terms of (1.47) 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:

$$\mathcal{E} = \hbar\omega = \frac{\hbar^2 k^2}{2m} = \frac{p^2}{2m}. \tag{1.48}$$

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 (1.42) to take account of the centred wave packet for the momentum representation to obtain

$$\Psi(x,t) = \sqrt{\frac{\sigma}{2\pi}}\left(\frac{2}{\pi}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0 x}\int_{-\infty}^{\infty} \mathrm{e}^{-\sigma^2 u^2 + \mathrm{i}ux - \mathrm{i}\omega t}\,\mathrm{d}u. \tag{1.49}$$

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

$$\omega = \frac{\hbar k^2}{2m} = \frac{\hbar}{2m}(u+k_0)^2 = \frac{\hbar u^2}{2m} + uv_{\mathrm{g}} + \frac{\hbar k_0^2}{2m}. \tag{1.50}$$

If (1.50) is inserted into (1.49), we recognize a new form for the ‘static’ *effective* momentum wave function:

$$\phi(k) = \sqrt{\sigma}\left(\frac{2}{\pi}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0(x-v_\mathrm{g}t/2)} \exp\left[-u^2\left(\sigma^2 + \mathrm{i}\frac{\hbar t}{2m}\right)\right] \tag{1.51}$$

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 (1.49) using the short-hand notation

$$\sigma' = \sqrt{\sigma^2 + \mathrm{i}\frac{\hbar t}{2m}} \tag{1.52$a$}$$

and

$$x' = x - v_\mathrm{g}t. \tag{1.52$b$}$$

This gives

$$\begin{aligned}\Psi(x',t) &= \sqrt{\frac{\sigma}{2\pi}}\left(\frac{2}{\pi}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0(x-v_\mathrm{g}t/2)} \int_{-\infty}^{\infty} \mathrm{e}^{-\sigma'^2u^2+\mathrm{i}ux'}\,\mathrm{d}u \\ &= \frac{\sqrt{\sigma}}{(2\pi)^{1/4}\sigma'}\mathrm{e}^{\mathrm{i}k_0(x-v_\mathrm{g}t/2)} \exp\left[-\left(\frac{x'}{2\sigma'}\right)^2\right].\end{aligned} \tag{1.53}$$

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 $\sigma'$ are complex, the result is

$$(\Psi,\Psi) = \frac{\sigma}{|\sigma'|} = \frac{1}{\sqrt{1+\hbar^2t^2/(4m^2\sigma^4)}} \equiv \frac{1}{S}. \tag{1.54}$$

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

$$\langle x\rangle = \frac{(\Psi, x\Psi)}{(\Psi,\Psi)} = v_\mathrm{g}t. \tag{1.55}$$

Similarly, the standard deviation in position is found to be

$$\langle(\Delta x)^2\rangle = \sigma^2S^2 = \sigma^2\left[1 + \frac{\hbar^2t^2}{4m^2\sigma^4}\right]. \tag{1.56}$$

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

$$\Delta x\Delta p = \frac{\hbar}{2}\sqrt{1+\frac{\hbar^2t^2}{4m^2\sigma^4}}. \tag{1.57}$$

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

Fowler A B 1985 *US Patent* 4550330

Landau L D and Lifshitz E M 1958 *Quantum Mechanics* (London: Pergamon)

Longair M S 1984 *Theoretical Concepts in Physics* (Cambridge: Cambridge University Press)

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

Yacoby A, Heiblum M, Umansky V, Shtrikman H and Mahalu D 1994 *Phys. Rev. Lett.* **73** 3149–52

## Problems

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

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

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 (1.32) 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 $k$?

7. The longest wavelength that can cause the emission of electrons from silicon is 296 nm. (*a*) What is the work function of silicon? (*b*) If silicon is irradiated with light of 250 nm wavelength, what is the energy and momentum of the emitted electrons? What is their wavelength? (*c*) If the incident photon flux is $5\ \mathrm{mW\ cm^{-2}}$, what is the photoemission current density?

8. For particles which have a thermal velocity, what is the wavelength at 300 K of electrons, helium atoms, and the $\alpha$-particle (which is ionized $^4\mathrm{He}$)?

9. Consider that an electron is confined within a region of 10 nm. If we assume that the uncertainty principle provides a RMS value of the momentum, what is their confinement energy?

10. A wave function has been determined to be given by the spatial variation

$$\Psi(x) = \begin{cases} 2A & -a < x < 0 \\ 2A(a - x) & 0 < x < a \\ 0 & \text{elsewhere.} \end{cases}$$

Determine the value of $A$, the expectation value of $x$, $x^2$, $p$ and $p^2$. What is the value of the uncertainty in position–momentum?

11. A wave function has been determined to be given by the spatial variation

$$\Psi(x) = \begin{cases} 2A\sin\left(\frac{\pi}{a}\right) & -a < x < a \\ 0 & \text{elsewhere.} \end{cases}$$

Determine the value of $A$, the expectation value of $x$, $x^2$, $p$ and $p^2$. What is the value of the uncertainty in position–momentum?

# Chapter 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).

It must be noted that the choice of a wave-function-based approach to quantum mechanics is not the only option. Indeed, two separate formulations of the new quantum mechanics appeared almost simultaneously. One was developed by Werner Heisenberg, at that time a lecturer in Göttingen (Germany), during 1925. In this approach, a calculus of non-commuting operators was developed. This approach was quite mathematical, and required considerable experience to work through in any detail. It remained until later to discover that this calculus was actually representable by normal matrix calculus. The second formulation was worked out by Erwin Schrödinger, at the time a Professor in Vienna, over the winter vacation of 1927. In Schrödinger's formulation, a wave equation was found to provide the basic understanding of quantum mechanics. Although not appreciated at the time, Schrödinger presented the connection between the two approaches in subsequent papers. In a somewhat political environment, Heisenberg received the 1932 Nobel prize for 'discovering' quantum mechanics, while Schrödinger was forced to share the 1933 prize with Paul Dirac for advances in atomic physics. Nevertheless, it is Schrödinger's formulation which is almost universally used today, especially in textbooks. This is especially true for students with a background in electromagnetic fields, as the concept of a wave equation is not completely new to them.

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 $\boldsymbol{F} = m\boldsymbol{a}$. As mentioned above, we cannot 'derive' this equation, because the equation itself is the basic postulate of wave mechanics, as formulated by Schrödinger (1926).

Our prime rationale in developing the wave equation is that the 'wave function' should in fact be a wave. That is, we prefer the spatial and temporal variations to have the form

$$\psi \sim \mathrm{e}^{\mathrm{i}(kx-\omega t)} \tag{2.1}$$

in one dimension. To begin, we can rewrite (1.42) as

$$\Psi(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}(kx-\omega t)}\, \mathrm{d}k. \tag{2.2}$$

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 (2.2) with respect to time, to yield

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

which can be rewritten as

$$\mathrm{i}\,\hbar \frac{\partial \Psi}{\partial t} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathcal{E} \mathrm{e}^{\mathrm{i}(kx-\omega t)}\, \mathrm{d}k. \tag{2.4}$$

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

$$\mathcal{E} = T + V = \frac{p^2}{2m} + V(x,t). \tag{2.5}$$

The momentum does operate on the momentum representation functions, but by using our position space operator form (1.31), the energy term can be pulled out of the integral in (2.4), and we find that the kinetic energy may be rewritten using (1.4) as

$$\frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m} \tag{2.6}$$

but we note that the factor of $k^2$ can be obtained from (2.2) as

$$\begin{aligned}\frac{\partial^2 \Psi(x,t)}{\partial x^2} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \frac{\partial^2}{\partial x^2} \mathrm{e}^{\mathrm{i}(kx-\omega t)}\, \mathrm{d}k \\ &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) k^2 \mathrm{e}^{\mathrm{i}(kx-\omega t)}\, \mathrm{d}k \end{aligned} \tag{2.7}$$

so that we can collect the factors inside the integrals to yield

$$\mathrm{i}\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + V(x,t)\Psi(x,t). \tag{2.8}$$

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.

This new wave equation is a complex equation, in that it involves the complex factor $\mathrm{i} = \sqrt{-1}$. We expect, therefore, that the wave function $\Psi(x,t)$ is a complex quantity itself, which is why we use the squared magnitude in the probability definitions. This wave function thus has a magnitude and a phase, both of which are important quantities. We will see below that the magnitude is related to the density (charge density when we include the charge itself) while the phase is related to the 'velocity' with which the density moves. This leads to a *continuity equation* for probability (or for charge), just as in electromagnetic theory.

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 (2.8), which gives (the exponential factor is omitted as it cancels equally from all terms)

$$-a\frac{\partial S}{\partial t}+\mathrm{i}\hbar\frac{\partial a}{\partial t}=\frac{a}{2m}\left(\frac{\partial S}{\partial x}\right)^2-\frac{\mathrm{i}\hbar a}{2m}\frac{\partial^2 S}{\partial x^2}-\frac{\mathrm{i}\hbar}{m}\frac{\partial S}{\partial x}\frac{\partial a}{\partial x}-\frac{\hbar^2}{2m}\frac{\partial^2 a}{\partial x^2}+Va. \quad (2.9)$$

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

$$\frac{\partial S}{\partial t}+\frac{1}{2m}\left(\frac{\partial S}{\partial x}\right)^2+V-\frac{\hbar^2}{2am}\frac{\partial^2 a}{\partial x^2}=0 \quad (2.10)$$

and

$$\frac{\partial a}{\partial t}+\frac{a}{2m}\frac{\partial^2 S}{\partial x^2}+\frac{1}{m}\frac{\partial S}{\partial x}\frac{\partial a}{\partial x}=0. \quad (2.11)$$

In (2.10), 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 (2.2) leads immediately to (2.5), 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 (2.10), 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, (2.11), can be rearranged by multiplying by $a$, for which (in vector notation for simplicity of recognition)

$$\frac{\partial a^2}{\partial t}+\boldsymbol{\nabla}\cdot\left(\frac{a^2}{m}\boldsymbol{\nabla}S\right)=0. \quad (2.12)$$

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, 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 $\Psi$ 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

$$P(x,t) = \Psi^*(x,t)\Psi(x,t) = |\Psi(x,t)|^2. \tag{2.13}$$

This, of course, leads to the normalization of (1.10), which just expresses the fact that the sum of the probabilities must be unity. If (2.13) 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 (2.13) to the local charge density, then there must be a corresponding current density $\boldsymbol{J}$, such that $(\rho = -eP)$

$$e\frac{\partial P}{\partial t} = \boldsymbol{\nabla} \cdot \boldsymbol{J} \tag{2.14}$$

although we use only the $x$-component here. Now, the complex conjugate of (2.8) is just

$$-\mathrm{i}\hbar\frac{\partial \Psi^*}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2 \Psi^*}{\partial x^2} + V(x,t)\Psi^*(x,t). \tag{2.15}$$

We now use (2.13) in (2.14), with (2.8) and (2.15) inserted for the partial derivatives with respect to time, as (we neglect the charge, and will find the probability current)

$$\mathrm{i}\hbar\frac{\partial P}{\partial t} = -\frac{\hbar^2}{2m}[\Psi^*\nabla^2\Psi - (\nabla^2\Psi^*)\Psi] \tag{2.16}$$

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

$$\boldsymbol{J}_\Psi = \frac{\hbar}{2m\mathrm{i}}[\Psi^*(\boldsymbol{\nabla}\Psi) - (\boldsymbol{\nabla}\Psi^*)\Psi]. \tag{2.17}$$

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 (2.17) 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 (2.17) as

$$\boldsymbol{J}_\Psi = \frac{1}{2m}(\boldsymbol{p} + \boldsymbol{p}^*)|\Psi|^2. \tag{2.18}$$

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 (2.18) is simply the product of the velocity and the probability, which yields the probability current.

The result (2.18) differs from the earlier form that appears in (2.12). 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 (2.18) would yield zero identically, while (2.12) would give a non-zero result. However, (2.12) was obtained by separating the real and imaginary parts of (2.9), 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, (2.9) was obtained for a very special form of the wave function. On the other hand, (2.18) results from a quite general wave function, and while the specific result depended upon a plane wave, the approach was not this limited. If (2.2) is used for the general wave function, then (2.18) 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

$$\langle A \rangle = (\Psi, A\Psi) = (\Psi, \underline{a}\Psi) = \underline{a} \tag{2.19}$$

for a properly normalized wave function. Now,

$$\underline{a}^* = (\Psi, \underline{a}^*\Psi) = (\underline{a}\Psi, \Psi) = (A\Psi, \Psi) = (\Psi, A^+\Psi) \tag{2.20}$$

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, $\underline{a} = \underline{a}^* \implies A = A^+$. 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 (2.15), and then divide by this same wave function, we obtain

$$\frac{\mathrm{i}\hbar}{\chi}\frac{\partial \chi}{\partial t} = -\frac{\hbar^2}{2m\Psi}\frac{\partial^2 \Psi}{\partial x^2} + V(x). \tag{2.21}$$

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

$$\chi(t) = \mathrm{e}^{-\mathrm{i}\mathcal{E}t/\hbar} \tag{2.22}$$

and the *time-independent Schrödinger equation*

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \Psi}{\partial x^2} + V(x)\Psi(x) = \mathcal{E}\Psi(x). \tag{2.23}$$

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

$$\frac{\partial^2 \Psi}{\partial x^2} + k^2\Psi(x) = 0 \tag{2.24}$$

where

$$\frac{\hbar^2 k^2}{2m} = \mathcal{E} \qquad k = \sqrt{\frac{2m\mathcal{E}}{\hbar^2}}. \tag{2.25}$$

The solution to (2.24) is clearly of the form of sines and cosines, but here we will take the exponential terms, and

$$\Psi(x) = A\mathrm{e}^{\mathrm{i}kx} + B\mathrm{e}^{-\mathrm{i}kx}. \tag{2.26}$$

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 (2.22) 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 $k$ as

$$k = \sqrt{\frac{2m[\mathcal{E} - V(x)]}{\hbar^2}}. \tag{2.27}$$

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 (2.26) 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 (2.10), 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

$$\frac{\lambda}{V}\frac{\partial V}{\partial x} \ll 1 \tag{2.28}$$

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 (2.26), with the exponential factors replaced by

$$\exp\left[\pm\int^{x}\sqrt{\frac{2m[\mathcal{E}-V(x')]}{\hbar^2}}\,\mathrm{d}x'\right]. \tag{2.29}$$

However, it is important to note that solutions such as (2.29) 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, (2.23) 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

$$V = V_0\Theta(x) \qquad \text{with} \quad V_0 > 0 \tag{2.30}$$

where $\Theta(x)$ is the Heaviside step function in which $\Theta = 1$ for $x \geq 0$, 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 (2.27), as the wave vector is imaginary for $\mathcal{E} < V_0$. Only one exponent

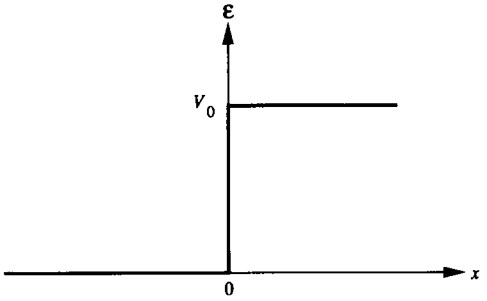


**Figure 2.1.** Schematic view of the potential of (2.30) which is non-zero (and constant) only in the positive half-space.

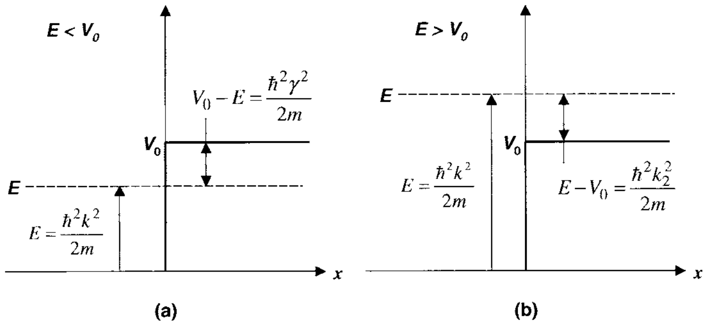

$E < V_0$
$V_0 - E = \frac{\hbar^2\gamma^2}{2m}$
$V_0$
$E$
$E = \frac{\hbar^2 k^2}{2m}$
$x$
(a)
$E > V_0$
$E$
$V_0$
$E = \frac{\hbar^2 k^2}{2m}$
$E - V_0 = \frac{\hbar^2 k_2^2}{2m}$
$x$
(b)

**Figure 2.2.** The various wave vectors are related to the energy of the wave: (*a*) the case for $E < V_0$; (*b*) the case for $E > V_0$.

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 (2.26), composed of an incident wave (the positive-exponent term) and a reflected wave (the negative-exponent term). That is, we write the wave function for $x < 0$ as

$$\psi_1(x) = A\mathrm{e}^{\mathrm{i}kx} + B\mathrm{e}^{-\mathrm{i}kx} \tag{2.31}$$

where the energy and the wave vector $k$ are related by (2.25). This behaviour is shown in figure 2.2(*a*). In the positive half-space, the solution of the Schrödinger equation is given by

$$\psi_2 = C\mathrm{e}^{-\gamma x} \tag{2.32}$$

where

$$\gamma = \sqrt{\frac{2m[V_0 - \mathcal{E}]}{\hbar^2}}. \tag{2.33}$$

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

$$\psi_1(0) = \psi_2(0) \Rightarrow A + B = C. \tag{2.34}$$

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 (2.23) 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

$$\left.\frac{\mathrm{d}\psi_1}{\mathrm{d}x}\right|_{x=0} = \left.\frac{\mathrm{d}\psi_2}{\mathrm{d}x}\right|_{x=0} \Rightarrow \mathrm{i}k(A - B) = -\gamma C. \tag{2.35}$$

This last equation can be rearranged by placing the momentum term in the denominator on the right-hand side. Then adding (2.34) and (2.35) leads to

$$\frac{C}{A} = \frac{2\mathrm{i}k}{\mathrm{i}k - \gamma}. \tag{2.36}$$

This result can now be used in (2.34) to find

$$\frac{B}{A} = \frac{\mathrm{i}k + \gamma}{\mathrm{i}k - \gamma}. \tag{2.37}$$

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 $(x < 0)$

$$\Psi_1(x) = A[\mathrm{e}^{\mathrm{i}kx} + \mathrm{e}^{-\mathrm{i}(kx+\theta)}] \tag{2.38}$$

where

$$\theta = 2\tan^{-1}\left(\frac{\gamma}{k}\right). \tag{2.39}$$

The probability amplitude is given by

$$|\psi_1(x)|^2 = 2A^2[1 + 2\cos(2kx + \theta)] \qquad x < 0. \tag{2.40}$$

As may have been expected, this is a *standing-wave* pattern, with the probability oscillating from 0 to twice the value of $A^2$. 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 $\pi$. 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 (2.31), which includes both an incident wave and a reflected wave. Similarly, the transmitted wave will be assumed to be of the form

$$\psi_2 = C\mathrm{e}^{\mathrm{i}k_2 x} \tag{2.41}$$

where

$$k_2 = \sqrt{\frac{2m}{\hbar^2}(E - V_0)}. \tag{2.42}$$

The relationships between this wave vector and that for the region $x < 0$ are schematically described in figure 2.2(*b*). Again, we will match both the wave function and its derivative at $x = 0$. This leads to

$$\begin{gathered}\psi_1(0) = \psi_2(0) \Rightarrow A + B = C \\ \left.\frac{\mathrm{d}\psi_1}{\mathrm{d}x}\right|_{x=0} = \left.\frac{\mathrm{d}\psi_2}{\mathrm{d}x}\right|_{x=0} \Rightarrow \mathrm{i}k(A - B) = \mathrm{i}k_2 C.\end{gathered} \tag{2.43}$$

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 (2.43) leads to

$$\frac{C}{A} = \frac{2k}{k + k_2} \qquad \frac{B}{A} = \frac{k - k_2}{k + k_2}. \tag{2.44}$$

Here, we see that if $V_0 \to 0$, $k_2 \to k$ 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 (2.17). For the incident and transmitted waves, these currents are simply

$$J_C = \frac{\hbar k_2}{m}\left(\frac{2k}{k+k_2}\right)^2 \qquad J_A = \frac{\hbar k}{m}. \tag{2.45}$$

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

$$T = \frac{J_C}{J_A} = \frac{4kk_2}{(k+k_2)^2} \tag{2.46}$$

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

$$R = -\frac{J_B}{J_A} = \left(\frac{k-k_2}{k+k_2}\right)^2. \tag{2.47}$$

This leads to the result that

$$T + R = 1. \tag{2.48}$$

A critical point arises when $k_2 = 0$, that is, the energy is resonant with the top of the potential barrier. For this energy, the reflection coefficient from (2.47) 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 $\Psi_2 = C + Dx$, 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, (2.43) must become $B = A$. As a result of the first of equations (2.43), 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 (2.36) 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

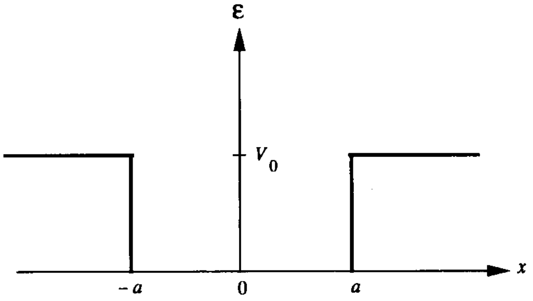


**Figure 2.3.** A potential well is formed by two barriers located at $|x| = a$.

the latter case, as shown in figure 2.3. 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 $\gamma$. 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 (2.24), with the wave vector defined by (2.25). The solution is now given, just as in the free-particle case, by (2.26). At the right-hand boundary, this leads to the situation

$$A\mathrm{e}^{\mathrm{i}ka} + B\mathrm{e}^{-\mathrm{i}ka} = 0 \tag{2.49}$$

and at the left-hand boundary,

$$A\mathrm{e}^{-\mathrm{i}ka} + B\mathrm{e}^{\mathrm{i}ka} = 0. \tag{2.50}$$

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

$$\begin{vmatrix} \mathrm{e}^{\mathrm{i}ka} & \mathrm{e}^{-\mathrm{i}ka} \\ \mathrm{e}^{-\mathrm{i}ka} & \mathrm{e}^{\mathrm{i}ka} \end{vmatrix} = 0. \tag{2.51}$$

This leads to the requirement that

$$\sin(2ka) = 0 \tag{2.52}$$

or

$$k = \frac{n\pi}{2a} \qquad \mathcal{E}_n = \frac{n^2\pi^2\hbar^2}{8ma^2} \qquad n = 1, 2, 3, \ldots. \tag{2.53}$$

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 (2.49). This leads to

$$\frac{B}{A} = -\mathrm{e}^{\mathrm{i}n\pi} = (-1)^{n+1}. \tag{2.54}$$

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

$$\Psi_n(x) = \begin{cases} A\cos(n\pi x/(2a)) & n \text{ odd} \\ A\sin(n\pi x/(2a)) & n \text{ even} \end{cases} \tag{2.55}$$

These can be combined by offsetting the position, so that

$$\Psi_n(x) = A\sin\left[\frac{n\pi}{2a}(x+a)\right]. \tag{2.56}$$

This last solution fits both boundary conditions, and yields the two solutions of (2.55) when the multiple-angle expansion of the sine function is used. Of course, each indexed wave function of (2.56) corresponds to one of the Fourier expansion terms in the Fourier series that represents a square barrier. In fact, (2.24) 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 (2.56), and the general inner product with the range of integration now defined from $-a$ to $a$. This leads to

$$(\Psi_n, \Psi_n) = A^2 \int_{-a}^{a} \sin^2\left[\frac{n\pi}{2a}(x+a)\right] \mathrm{d}x = 1. \tag{2.57}$$

This readily leads to the normalization

$$A = \frac{1}{\sqrt{a}}. \tag{2.58}$$

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,

$$-\Psi(x) = \sum_n \frac{c_n}{\sqrt{a}} \sin\left[\frac{n\pi}{2a}(x+a)\right] \tag{2.59}$$

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

$$\sum_n |c_n|^2 = 1. \tag{2.60}$$

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 (2.55), 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 $\alpha$. Repeated application of this variable replacement shows that $\alpha^2 = 1$, so $\alpha$ can only take on the values $\pm 1$, 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. In this regard, the system is precisely like the single finite barrier that was discussed in section 2.3.2. When the energy is below the height of the barrier, the wave must decay into the region where the barrier exists, as shown in figure 2.4. On the other hand, when the energy is greater than the barrier height, propagating waves exist in all regions, but there is a mismatch in the wave vectors, which leads to quasi-bound states and reflections from the interface. We begin with the case for the energy below the barrier height, which is the case shown in figure 2.4.

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

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

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + k^2\Psi = 0 \qquad k^2 = \frac{2mE}{\hbar^2} \tag{2.61}$$

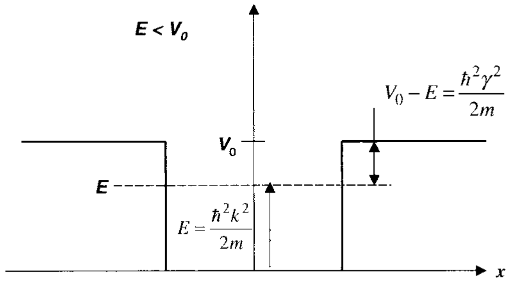

$E < V_0$
$V_0 - E = \frac{\hbar^2\gamma^2}{2m}$
$V_0$
$E$
$E = \frac{\hbar^2 k^2}{2m}$
$x$

**Figure 2.4.** The various wave vectors are related to the energy of the wave for the case of $E < V_0$.

In (2.61), it must be remembered that $V_0$ is the magnitude of the potential barrier, and is a positive quantity. Similarly, in the range $|x| > a$, the Schrödinger equation becomes

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} - \gamma^2\Psi = 0 \qquad \gamma^2 = \frac{2m(V_0 - E)}{\hbar^2}. \tag{2.62}$$

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 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 (2.62) for the damped region will be of the form $C\mathrm{e}^{-\gamma|x|}$, $|x| > a$. 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 $k$ (and hence $\gamma$, since it is not independent of $k$) 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 $\Psi'/\Psi$ continuous. (This is obviously called the logarithmic derivative since it is the derivative of the logarithm of $\Psi$.) 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

$$\frac{-k\sin(kx)}{\cos(kx)} = -k\tan(kx). \tag{2.63}$$

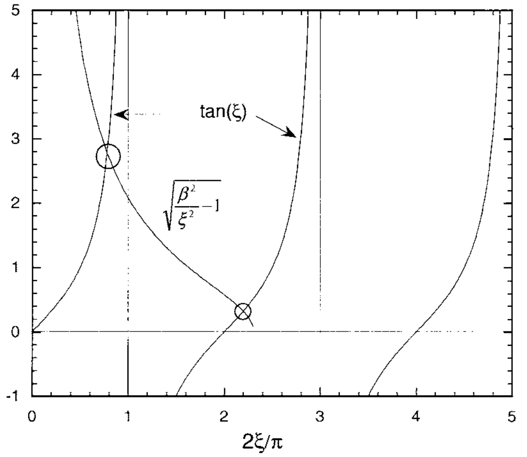


**Figure 2.5.** The graphical solution of (2.65) is indicated by the circled crossings. Here, we have used the values of $a = 5$ nm, $V_0 = 0.3$ eV, and $m = 0.067m_0$, appropriate to a GaAs quantum well between two layers of GaAlAs. The two circled crossings indicate that there are two even-symmetry solutions.

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

$$k\tan(ka) = \gamma. \tag{2.64}$$

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

$$\tan(\xi) = \frac{\gamma}{k} = \sqrt{\frac{\beta^2}{\xi^2} - 1} \qquad \beta^2 = \frac{2mV_0a^2}{\hbar^2}. \tag{2.65}$$

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 $(0, V_0)$ that constitute bound states. In general, the solution must be found graphically. This is shown in figure 2.5, in which we plot the left-hand side of (2.65) 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 $\beta$ is reduced, and there is a smaller range of $\xi$ 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 $\beta$. 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 $\xi$ 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 $\beta$ of the decreasing curve in figure 2.5 moves toward the origin. Thus, the solution point approaches $\xi = 0$, or $k = 0$. By expanding the tangent function for small $\xi$, it is found that the solution is approximately $\beta \simeq \xi$. However, this requires $\mathcal{E} \simeq V_0$, 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 ($\beta$) moves to the right and more allowed energy levels appear for the even functions. It is clear from the construction of figure 2.5 that at least one solution must occur, even if the width is the parameter made smaller, as the $\xi$-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 (2.55). Again, the logarithmic derivative of the propagating waves for $|x| < a$ may be found to be

$$\frac{k\cos(kx)}{\sin(kx)} = k\,\mathrm{cotan}(kx). \tag{2.66}$$

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

$$k\,\mathrm{cotan}(kx) = -\gamma \tag{2.67}$$

or

$$\mathrm{cotan}(\xi) = -\sqrt{\frac{\beta^2}{\xi^2} - 1}. \tag{2.68}$$

Again, a graphical solution is required. This is shown in figure 2.6. The difference between this case and that for the even wave functions is that the left-hand side of (2.68) starts on the opposite side of the $\xi$-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 2.5 and 2.6 that the solution point that does occur lies in between those that occur for the even-symmetry wave functions. Thus, this 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

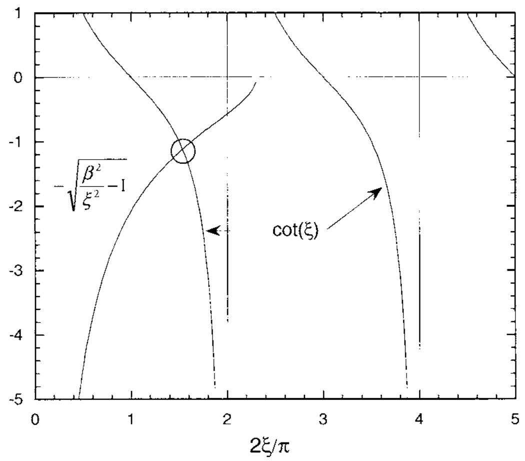


**Figure 2.6.** The graphical solution of (2.68). The parameters here are the same as those used in figure 2.5. Only a single solution (circled) is found for the anti-symmetric solution, and this energy lies between the two found in figure 2.5.

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 (2.53).

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 28% AlAs alloyed into GaAs, has a band gap that is 0.45 eV larger than that of pure GaAs (about 1.85 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

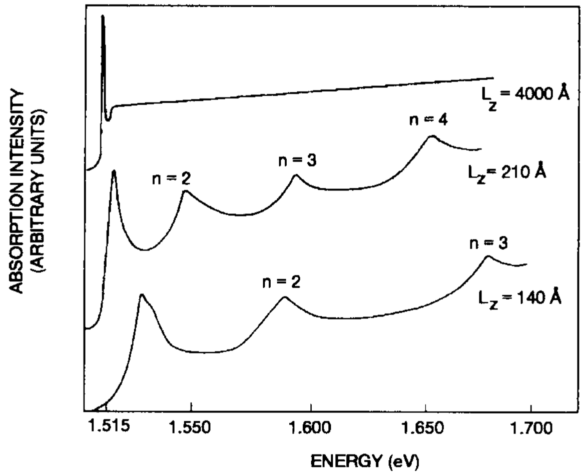


**Figure 2.7.** Absorption observed between bound states of holes and electrons in 21 nm and 14 nm quantum wells formed by placing a layer of GaAs between two layers of AlGaAs. For a well thickness of 400 nm, the absorption is uniform. (After Dingle *et al* (1974), by permission.)

these transitions will lie below the absorption band for the AlGaAs. Such an absorption spectrum is shown in figure 2.7. Transitions at the lowest heavy-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 2.8, 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.* $\mathcal{E} > V_0$

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

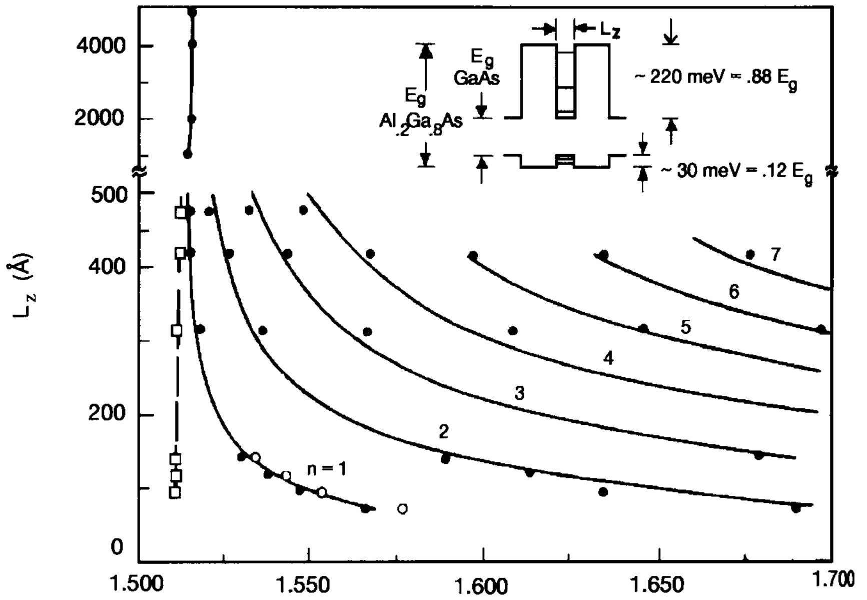


**Figure 2.8.** Variation of the absorption bands for transitions from heavy-hole (solid circles) and light-hole (open circles) levels to electron levels of the quantum wells as a function of well width. The solid curves are calculated positions. (After Dingle *et al* (1974), by permission.)

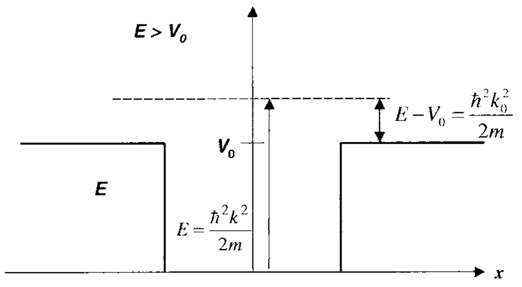


**Figure 2.9.** The various wave vectors are related to the energy of the wave for the case of $E > V_0$.

from the left (negative $x$) to the right (positive $x$).

Throughout the entire space, the Schrödinger equation is given by the form (2.61), with different values of $\boldsymbol{k}$ in the various regions. The value of $\boldsymbol{k}$ given in

(2.61) remains valid in the quantum well region, while for $|x| > a$ (see figure 2.9),

$$k_0^2 = \frac{2m(E - V_0)}{\hbar^2}. \tag{2.69}$$

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

$$F\mathrm{e}^{\mathrm{i}k_0 x}. \tag{2.70}$$

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

$$C\mathrm{e}^{\mathrm{i}kx} + D\mathrm{e}^{-\mathrm{i}kx}. \tag{2.71}$$

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

$$\mathrm{e}^{\mathrm{i}k_0 x} + B\mathrm{e}^{-\mathrm{i}k_0 x} \tag{2.72}$$

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

$$\begin{bmatrix} 0 & \omega & \omega^{-1} & -\omega_0 \\ 0 & \omega & -\omega^{-1} & -(k_0/k)\omega_0 \\ -\omega_0 & \omega^{-1} & \omega & 0 \\ \omega_0 & (k/k_0)\omega^{-1} & -(k/k_0)\omega & 0 \end{bmatrix} \begin{bmatrix} B \\ C \\ D \\ F \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \omega^{-1} \\ \omega^{-1} \end{bmatrix}. \tag{2.73}$$

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:

$$F = \frac{\mathrm{e}^{-2\mathrm{i}k_0 a}}{\cos(2ka) - \mathrm{i}[(k^2 + k_0^2)/2kk_0]\sin(2ka)}. \tag{2.74}$$

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 (2.74). This leads to

$$T = \frac{1}{1 + [(k^2 - k_0^2)/2kk_0]^2 \sin^2(2ka)}. \tag{2.75}$$

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

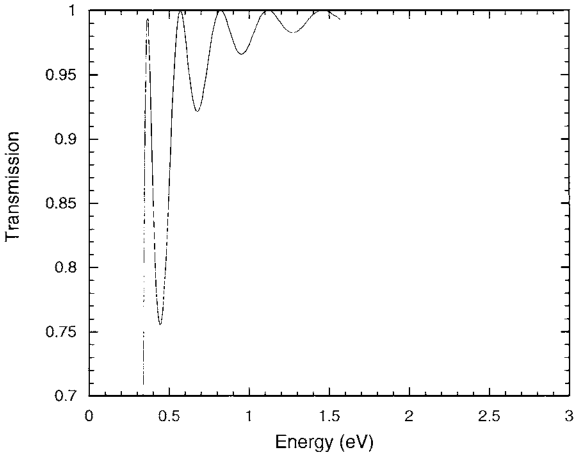


**Figure 2.10.** The transmission, given by (2.75), for a finite potential well. The parameters are those of figure 2.5, appropriate to a GaAs quantum well situated between two GaAlAs layers.

wave vector, and this leads to a larger modulation of the transmission coefficient. An example is shown in figure 2.10.

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 2.11, 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 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

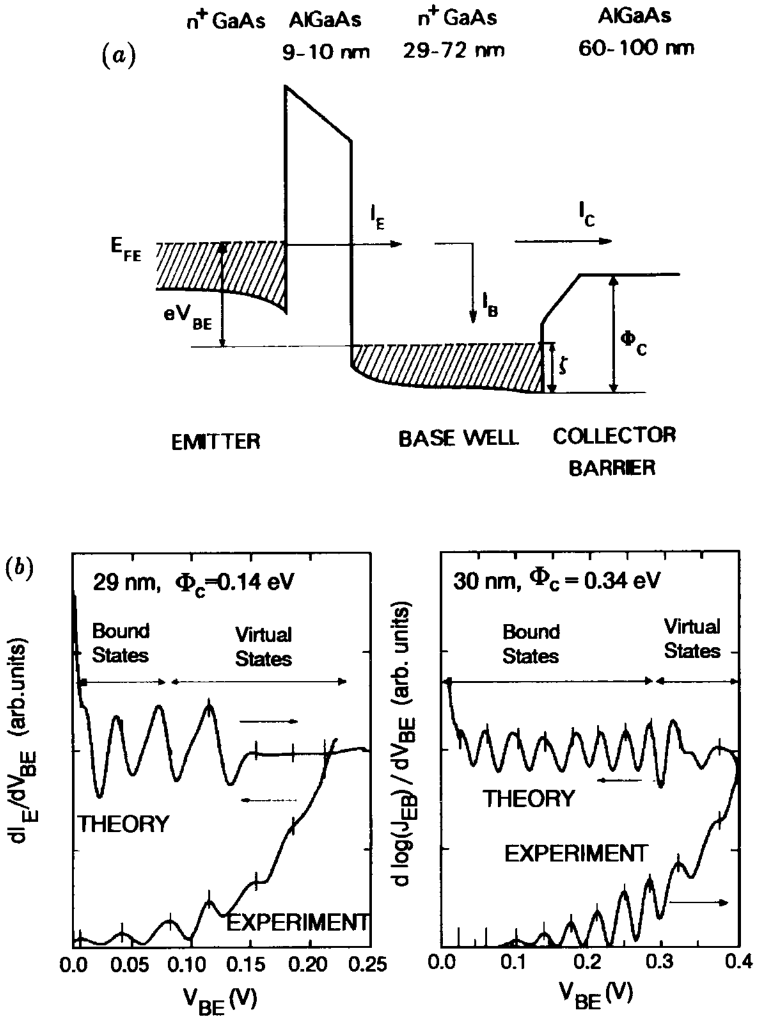


**Figure 2.11.** Transport of ballistic electrons through a double-barrier, ballistic transistor, whose potential profile is shown in (*a*). The quantum resonances of propagation over the well are evident in the density of collected electrons (*b*) for two different sizes. (After Heiblum *et al* (1987), by permission.)

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 unit 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.

Before proceeding, it is worthwhile to discuss the material systems that were used in the preceding examples, both theoretical and experimental. While one might at first glance think it quite difficult to put different materials together effectively, this is allowed today through the efficiency of molecular-beam epitaxy. In this growth process, materials can be grown almost one atomic layer at a time. This is facilitated in the GaAs–AlAs system, since these two materials have almost identical lattice constants. Hence, one can alloy the two materials to create a 'new' semiconductor with a band gap somewhere between that of GaAs and that of AlAs. Moreover, one can change the material system from the alloy to, for example, GaAs within one atomic layer. So, it is quite easy to create designer structures with molecular-beam epitaxial growth of semiconductor materials. The examples discussed in the preceding paragraphs are just some of the structures that can be easily made with this growth technology.

## 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 2.12(*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:

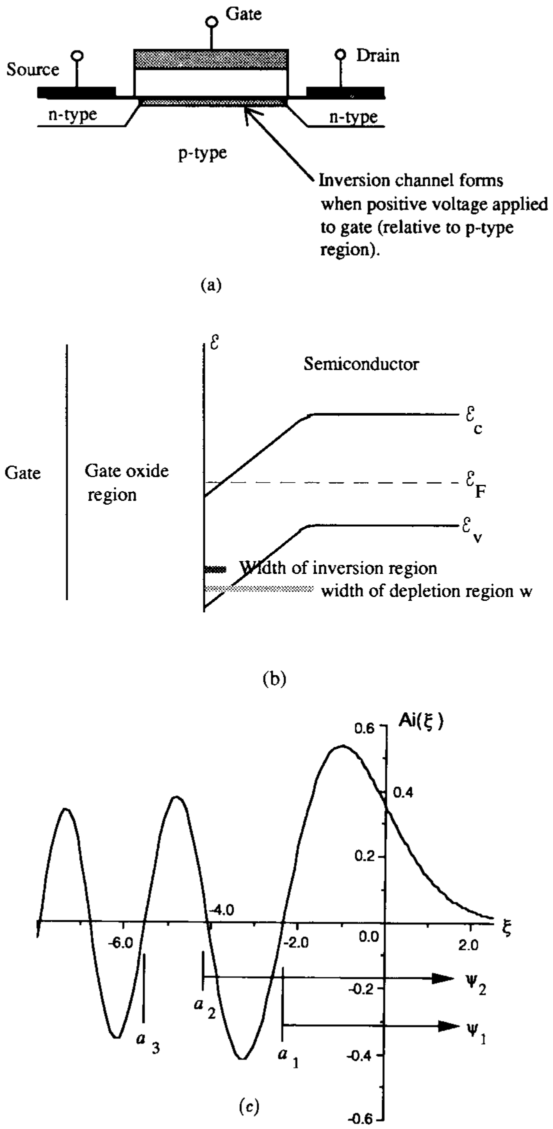


**Figure 2.12.** (*a*) A MOS field-effect transistor, (*b*) the triangular potential, and (*c*) the Airy function and the use of the zeros to match the boundary conditions.

(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

$$E_{\mathrm{s}} = \frac{e}{\varepsilon_{\mathrm{ox}}}(N_{\mathrm{a}}w + n_{\mathrm{s}}) \tag{2.76}$$

where $w$ is the thickness of the layer of ionized acceptors $N_{\mathrm{a}}$ (normal to the surface), the surface electron density $n_{\mathrm{s}}$ 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 (2.76) is discontinuous by the dielectric constant ratio. Thus, just inside the interface, (2.76) 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

$$E_{\mathrm{s}} = \frac{e}{\varepsilon_{\mathrm{s}}}\left(N_{\mathrm{a}}w + \frac{n_{\mathrm{s}}}{2}\right). \tag{2.77}$$

In this approximation, a constant electric field in this region gives rise to a linear potential in the Schrödinger equation (figure 2.12($b$)). 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

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + eE_{\mathrm{s}}x\Psi = \mathcal{E}\Psi \qquad \text{for} \quad x > 0. \tag{2.78}$$

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 $x$.

While the previous discussion has been for the case of a MOS transistor, such as found in silicon integrated circuits, quite similar behaviour arises in the GaAs–AlGaAs heterojunction high-electron-mobility transistor (HEMT). In this case, the AlGaAs plays a role similar to the oxide in the MOS transistor, with the exception that the dopant atoms are placed in this layer. The dopants near the interface are ionized, with the electrons falling into a potential well on the GaAs side of the interface, a process that is facilitated by the barrier created by the difference in the two conduction band edges, as shown in figure 2.13. There are very few ionized dopants in the GaAs, although the interface electric field still satisfies (2.76). This field is created by the positively charged, ionized donors and the negatively charged electrons in the potential well. By placing a metal gate on the surface of the AlGaAs layer, bias applied to the gate can affect the density of the carriers in the quantum well, reducing them to zero. This is a *depletion*

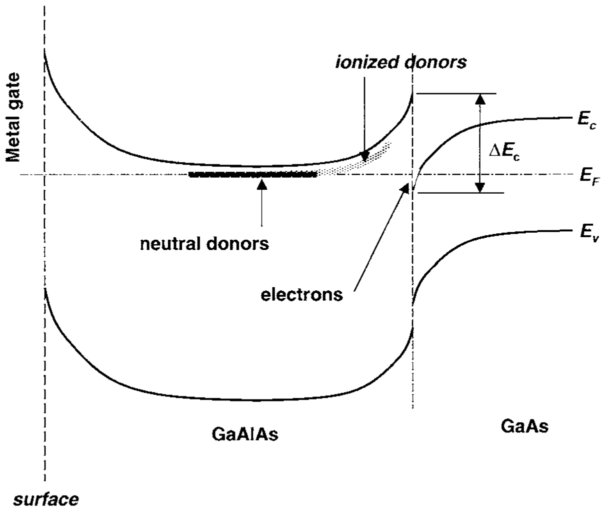


**Figure 2.13.** Band alignment for the AlGaAs–GaAs high-electron-mobility transistor. Ionized donors in the GaAlAs produce the electrons that reside in the triangular quantum well on the GaAs side of the interface.

device, as opposed to the inversion device of the MOS transistor. That is, the gate *removes* carriers from the channel in the HEMT, while the gate pulls carriers into the channel in the MOS transistor. Since the impurities are removed from the region where the carriers reside, the mobility is higher in the HEMT, hence its name. Such devices have found extensive use as analogue microwave amplifiers, either as power amplifiers or as low-noise amplifiers in receivers. Nevertheless, the potential variation near the interface still appears as approximately a triangular potential well, just as in the MOS transistor.

To simplify the solution, we will make a change of variables in (2.78), which will put the equation into a standard form. For this, we redefine the position and energy variables as

$$z = \left(\frac{2meE_s}{\hbar^2}\right)^{1/3} x \qquad z_0 = \frac{2m\mathcal{E}}{\hbar^2}\left(\frac{\hbar^2}{2meE_s}\right)^{2/3}. \tag{2.79}$$

Then, using $\xi = z - z_0$, (2.78) becomes

$$\frac{\partial^2\Psi}{\partial\xi^2} - \xi\Psi = 0. \tag{2.80}$$

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 $\xi$; and (ii) Ai($\xi$) behaves as a damped sinusoid with a period that also varies as $\xi$. For our purposes, this is all we need. The second solution of (2.80), the Airy functions 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 $\xi$, which is the case for 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 $a_{\mathrm{s}}$. Thus, choosing a wave function that puts the first zero $a_1$ at the point $x = 0$ would fit all the boundary conditions for the lowest energy level (figure 2.12). 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)

$$a_{\mathrm{s}} \simeq -\left(\frac{3\pi(4s-1)}{8}\right)^{2/3}. \tag{2.81}$$

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

$$\xi = -\left(\frac{9\pi}{8}\right)^{2/3} = -z_0 = \frac{2m\mathcal{E}}{\hbar^2}\left(\frac{\hbar^2}{2meE_{\mathrm{s}}}\right)^{2/3} \tag{2.82}$$

or

$$\mathcal{E}_1 = \frac{\hbar^2}{2m}\left(\frac{9\pi meE_{\mathrm{s}}}{4\hbar^2}\right)^{2/3} \tag{2.83}$$

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 2.14($a$), 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.

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

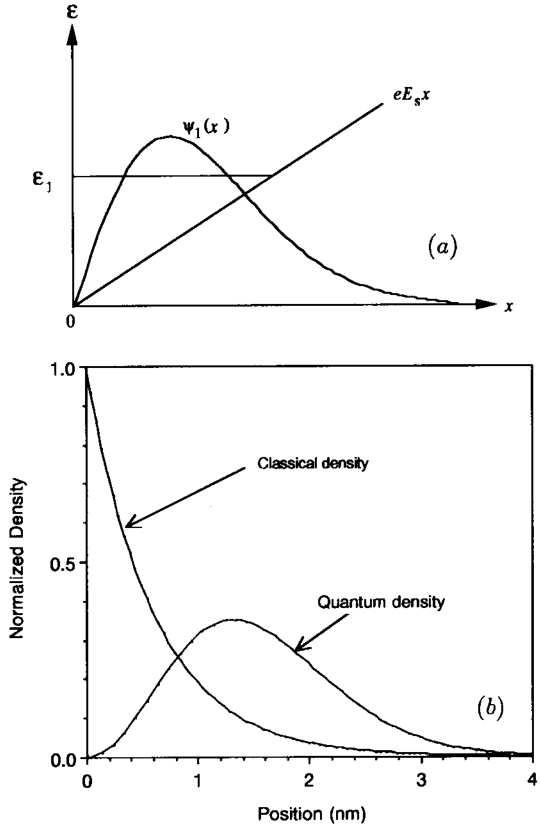


**Figure 2.14.** (*a*) The triangular potential well, the lowest energy level, and the Airy function wave function. (*b*) A comparison of the classical and quantum charge distributions.

semiconductor as

$$n(x) = n_s \exp\left[-\frac{eE_s x}{k_B T}\right]. \tag{2.84}$$

This decays to $1/e$ of the peak in a distance given by $k_BT/eE_s$. 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) $3 \times 10^5$ 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)

$$a'_{\mathrm{s}} = -\left(\frac{3\pi(4s-3)}{8}\right)^{2/3} \tag{2.85}$$

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

$$x \simeq \left(\frac{\hbar^2}{2meE_{\mathrm{s}}}\right)^{1/3}\left[\left(\frac{9\pi}{8}\right)^{2/3} - 2.1(3\pi)^{2/3}\right] \tag{2.86}$$

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 2.14(*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 2.15. 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

$$V(x) = \begin{cases} \infty & |x| > b/2 + a \\ 0 & a + b/2 > |x| > b/2 \\ V_0 & |x| < b/2. \end{cases} \tag{2.87}$$

(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

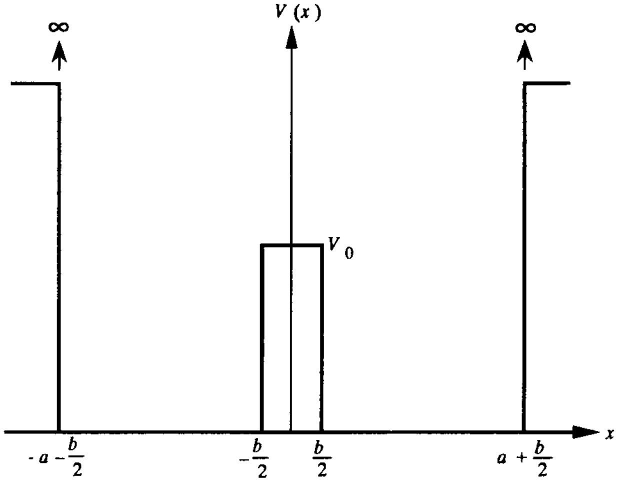


**Figure 2.15.** The double-well potential.

sum of propagating waves, one moving to the right and one moving to the left, 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

$$\tanh(\gamma b)[1 - \cos(2ka)] + \frac{k}{\gamma}\sin(2ka) = 0 \qquad (2.88a)$$

and

$$\tanh(\gamma b)[1 + \cos(2ka)] + \frac{\gamma}{k}\sin(2ka) = 0 \qquad (2.88b)$$

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

$$\sin(ka) = 0 \qquad \text{or} \qquad ka = n\pi \qquad (2.89)$$

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 (2.88) 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 (2.88$a$) about its relevant zero, where the latter is given by (2.89). This then leads to the approximate solutions

$$\sin(ka) = \pm|ka - n\pi|(1 - 2\mathrm{e}^{-2\gamma b}). \tag{2.90}$$

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 (2.90) can be found also by expanding the sine function around the zero point to give approximately

$$ka = n\pi \pm 2\sqrt{3}\mathrm{e}^{-\gamma b}. \tag{2.91}$$

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 2.16, 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

$$\Psi(x) = \sum_n c_n\psi_n(x). \tag{2.92}$$

In every sense, this series is strongly related to the Fourier series, where the expansion basis functions, our eigenfunctions, are determined by the geometry

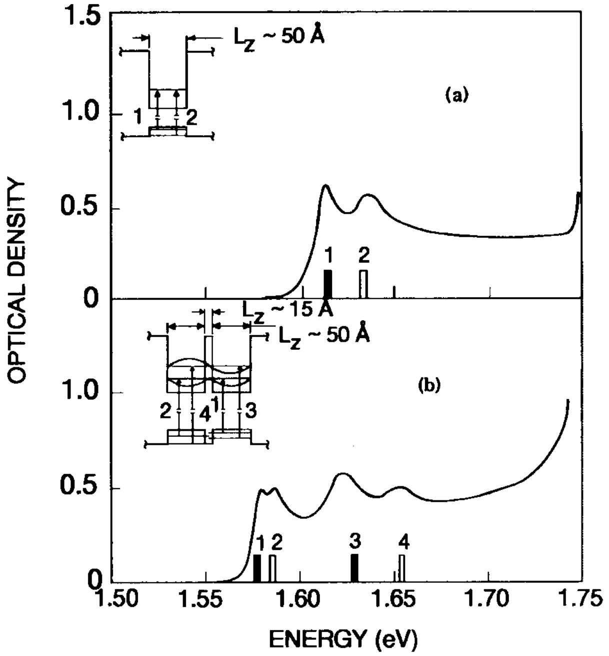


**Figure 2.16.** Optical absorption spectrum of a series of (*a*) 80 isolated GaAs quantum wells, of 5 nm thickness, separated by 18 nm of AlGaAs. In (*b*), the data are for 60 pairs of similar wells separated by a 1.5 nm barrier. The two bound states of the isolated wells each split into two combination levels in the double wells. (After Dingle *et al* (1975), by permission.)

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 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

$$\Psi(x,t) = \sum_n c_n \psi_n(x) \exp\left[-\frac{\mathrm{i}\mathcal{E}_n t}{\hbar}\right]. \tag{2.93}$$

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

$$\langle x\rangle = (\Psi, x\Psi) = \int_{-\infty}^{\infty} \Psi^*(x,t)x\Psi(x,t)\,\mathrm{d}x. \tag{2.94}$$

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 (2.91) 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)

$$\begin{aligned}\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\int x\rho\,\mathrm{d}x = \int x\frac{\partial\rho}{\partial t}\,\mathrm{d}x = -\int x\frac{\partial J}{\partial x}\,\mathrm{d}x \\ &= -\int \frac{\partial(xJ)}{\partial x}\,\mathrm{d}x + \int J\frac{\partial x}{\partial x}\,\mathrm{d}x = \int J\,\mathrm{d}x.\end{aligned} \tag{2.95}$$

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 (2.17) 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 (2.95) is identically the expectation value of the momentum. If (2.17) is used in (2.95), we find (in vector notation and with volume integrations)

$$\begin{aligned}\frac{\mathrm{d}\langle \boldsymbol{x}\rangle}{\mathrm{d}t} &= \int \boldsymbol{J}\,\mathrm{d}\boldsymbol{x} = \frac{\hbar}{2m\mathrm{i}}\int [\Psi^*(\boldsymbol{\nabla}\Psi) - (\boldsymbol{\nabla}\Psi^*)\Psi]\,\mathrm{d}\boldsymbol{x} \\ &= \frac{\hbar}{m\mathrm{i}}\int \Psi^*(\boldsymbol{\nabla}\Psi)\,\mathrm{d}\boldsymbol{x}\end{aligned} \tag{2.96}$$

where we have used $(\boldsymbol{\nabla}\Psi^*)\Psi = \boldsymbol{\nabla}(\Psi^*\Psi) - \Psi^*(\boldsymbol{\nabla}\Psi)$. 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:

$$\begin{aligned}\frac{\mathrm{d}\langle \boldsymbol{p}\rangle}{\mathrm{d}t} &= -\mathrm{i}\,\hbar\frac{\partial}{\partial t}\int \Psi^*(\boldsymbol{\nabla}\Psi)\,\mathrm{d}\boldsymbol{x}\\ &= \int\left(-\mathrm{i}\hbar\frac{\partial\Psi^*}{\partial t}\right)\boldsymbol{\nabla}\Psi\,\mathrm{d}\boldsymbol{x} - \int \Psi^*\boldsymbol{\nabla}\left(\mathrm{i}\hbar\frac{\partial\Psi^*}{\partial t}\right)\,\mathrm{d}\boldsymbol{x}\\ &= \int\left[-\frac{\hbar^2}{2m}\nabla^2\Psi^*\right]\boldsymbol{\nabla}\Psi\,\mathrm{d}\boldsymbol{x} - \int \Psi^*\boldsymbol{\nabla}\left[-\frac{\hbar^2}{2m}\nabla^2\Psi^*\right]\,\mathrm{d}\boldsymbol{x}\\ &\quad + \int[(V\Psi^*)\boldsymbol{\nabla}\Psi - \Psi^*\boldsymbol{\nabla}(V\Psi)]\,\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{2.97}$$

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

$$\begin{aligned}(p^2\Psi, p\Psi) - (\Psi, p^3\Psi) &= (\Psi, (p^+)^2 p\Psi) - (\Psi, p^3\Psi)\\ &= (\Psi, p^3\Psi) - (\Psi, p^3\Psi) = 0.\end{aligned} \tag{2.98}$$

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

$$\frac{\mathrm{d}\langle \boldsymbol{p}\rangle}{\mathrm{d}t} = -\langle \boldsymbol{\nabla}V(x)\rangle. \tag{2.99}$$

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 (2.93), 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 (2.93) to determine the coefficients in the generalized Fourier series, which this latter equation represents as

$$c_n = \int \psi_n^*(x)\Psi(x, 0)\,\mathrm{d}x. \tag{2.100}$$

This can be re-inserted into (2.93) to give the general solution

$$\begin{aligned}\Psi(x, t) &= \sum_n \int \psi_n^*(x')\psi_n(x)\Psi(x', 0)\exp\left[-\frac{\mathrm{i}\mathcal{E}_n t}{\hbar}\right]\mathrm{d}x' \\ &= \int K(x, x'; t, 0)\Psi(x', 0)\,\mathrm{d}x' \end{aligned} \tag{2.101}$$

where the *propagator kernel* is

$$K(x, x'; t, 0) = \sum_n \psi_n^*(x')\psi_n(x)\exp\left[-\frac{\mathrm{i}\mathcal{E}_n t}{\hbar}\right]. \tag{2.102}$$

The kernel (2.102) 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

$$c_n = \int \psi_n^*(x)\Psi(x, t_0)\exp\left[\frac{\mathrm{i}\mathcal{E}_n t_0}{\hbar}\right]\mathrm{d}x. \tag{2.103}$$

Following the same procedure—that is, re-introducing this into (2.90)—the general solution at arbitrary time for the wave function is then

$$\Psi(x, t) = \int K(x, x'; t, t_0)\Psi(x', t_0)\,\mathrm{d}x' \tag{2.104}$$

where

$$K(x, x'; t, t_0) = \sum_n \psi_n^*(x')\psi_n(x)\exp\left[-\frac{\mathrm{i}\mathcal{E}_n(t - t_0)}{\hbar}\right]. \tag{2.105}$$

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 (2.104) 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

$$G_{\mathrm{r}}(x,x';t,t_0) = -\mathrm{i}\Theta(t-t_0)K(x,x';t,t_0) \tag{2.106}$$

where $\Theta$ 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

$$G_{\mathrm{a}}(x,x';t,t_0) = \mathrm{i}\Theta(t_0-t)K(x,x';t,t_0) \tag{2.107}$$

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

$$K(x,x';t,t_0) = \mathrm{i}[G_{\mathrm{r}}(x,x';t,t_0) - G_{\mathrm{a}}(x,x';t,t_0)]. \tag{2.108}$$

We can compute the kernel from the general Schrödinger equation itself. To see this, note that when $t = t_0$, equation (2.105) becomes just a sum over a complete set of basis states, and a property of these orthonormal functions is that

$$K(x,x';t_0,t_0) = \delta(x-x') \tag{2.109}$$

which is expected just by investigating (2.102). This suggests that we can develop a differential equation for the kernel, which has the unique initial condition (2.109). This is done by beginning with the time derivative, as (for a free wave propagation, $V = 0$)

$$\begin{aligned}\frac{\partial K}{\partial t} &= -\sum_n \psi_n^*(x')\psi_n(x)\left[\frac{\mathrm{i}\mathcal{E}_n}{\hbar}\right]\exp\left[-\frac{\mathrm{i}\mathcal{E}_n(t-t_0)}{\hbar}\right]\\ &= -\frac{\mathrm{i}}{\hbar}\sum_n \psi_n^*(x')[H\psi_n(x)]\exp\left[-\frac{\mathrm{i}\mathcal{E}_n(t-t_0)}{\hbar}\right]\\ &= -\frac{\mathrm{i}}{\hbar}HK \end{aligned}\tag{2.110}$$

or

$$\mathrm{i}\hbar\frac{\partial K}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2 K}{\partial x^2}. \tag{2.111}$$

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 (2.109) and vanishing of $K$ at large distances. This leads to (we take $t_0$ as zero for convenience)

$$K(x,x',t) = \sqrt{\frac{m}{2\pi\hbar t}}\exp\left[-\frac{m(x-x')^2}{2\mathrm{i}\hbar t}\right]. \tag{2.112}$$

It may readily be ascertained that this satisfies the condition (2.109) at $t = 0$.

The definition of the kernel (2.105) 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 (2.102) in time, the kernel is just

$$K(x,x',\omega)=\sum_n\frac{\psi_n^*(x')\psi_n(x)}{\mathrm{i}(\omega-\omega_n)} \tag{2.113}$$

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 $\eta$ as in

$$G(x,x',\omega)=\sum_n\frac{\psi_n^*(x')\psi_n(x)}{(\omega_n-\omega\pm\mathrm{i}\eta)} \tag{2.114}$$

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

In the preceding paragraphs, we have developed a complicated formulation of the propagator kernel $K$. It is reasonable to ask why this was done, and the answer lies in the manner in which normal circuits are analysed. That is, when electrical engineers study circuits, the *response function* for those circuits is developed. Usually, this is the response of the linear circuit to an impulsive driving function $(\delta(t-t_0), t_0\to 0)$. If this function is denoted as $f(t)$, then the response to an arbitrary input $g(t)$ is given by

$$R(t)=\int_0^t f(t-\tau)g(\tau)\,\mathrm{d}\tau. \tag{2.115}$$

This *convolution* integral represents the systematic integration of the input function as the summation of the responses to a weighted sum of delta functions in time. In Laplace transform space, this is written as

$$\tilde{R}(s)=\tilde{f}(s)\tilde{g}(s) \tag{2.116}$$

where the tildes represent the Laplace transforms of the individual quantities. Equation (2.104) allows us to perform the same process with the wave function in quantum mechanics. In this latter equation, $\Psi(x',t_0)$ represents an arbitrary initial condition in space and time. The kernel represents the impulse reponse of the quantum system, and the integration over these initial space and time variables is the convolution integral. While it is not immediately obvious that this integration is a convolution, the integration represents the same phenomenon with the exception that the initial wave function $\Psi(x', t_0)$ occurs only at the initial time and place (which corresponds to setting $\tau = 0$ in (2.115) and integrating over the space variable).

Suppose we take the initial Gaussian wave packet, described by (1.18) with the normalization of (1.19), as the initial wave function. This is given by

$$\Psi(x', t_0) = \frac{1}{\pi^{1/4}} \exp(-x'^2/2) \mathrm{e}^{-\mathrm{i}\omega t_0}. \tag{2.117}$$

Here, the argument of the exponential is dimensionless, so we insert a spread value $x_0$ as the normalization. Hence, (2.117) is rewritten as

$$\Psi(x', 0) = \frac{1}{\sqrt{x_0}\pi^{1/4}} \exp(-x'^2/2x_0^2). \tag{2.118}$$

Using the kernel of (2.112), the wave function at an arbitrary time and place is given by

$$\begin{aligned}
\Psi(x,t) &= \sqrt{\frac{m}{2\mathrm{i}\hbar t x_0}} \frac{1}{\pi^{3/4}} \int_{-\infty}^{\infty} \exp\left[-\frac{m(x-x')^2}{2\mathrm{i}\hbar t} - \frac{x'^2}{2x_0^2}\right] \mathrm{d}x' \\
&= \sqrt{\frac{m}{2\mathrm{i}\hbar t x_0}} \frac{1}{\pi^{3/4}} \exp\left[-\frac{x^2}{2x_0^2(1+\mathrm{i}\hbar t/mx_0^2)}\right] \\
&\quad \times \int_{-\infty}^{\infty} \exp\left[-\frac{m}{2\mathrm{i}\hbar t}\left(x'\sqrt{1+\mathrm{i}\hbar t/mx_0^2} - \frac{x}{\sqrt{1+\mathrm{i}\hbar t/mx_0^2}}\right)^2\right] \mathrm{d}x' \\
&= \frac{1}{\sqrt{x_0(1+\mathrm{i}\hbar t/mx_0^2\pi^{1/4}}} \exp\left[-\frac{x^2}{2x_0^2(1+\mathrm{i}\hbar t/mx_0^2)}\right]
\end{aligned} \tag{2.119}$$

which is essentially (1.53) obtained in an easier fashion (here $x_0 = \sigma/\sqrt{2}$). Thus, if we can find the kernel, or the Green's functions, to describe a quantum system, then the evolution of an individual wave packet is found by the simple integration (2.104).

## 2.9 Numerical solution of the Schrödinger equation

If some arbitrary function is known at a series of equally spaced points, one can use a Taylor series to expand the function about these points and evaluate it in the regions between the known points. Consider the set of points shown in figure 2.17 for example. Here, each point is separated from its neighbours by the value $a$ (in position). This allows us to develop a finite-difference scheme for the numerical evaluation of the Schrödinger equation. If we first expand the function in a Taylor

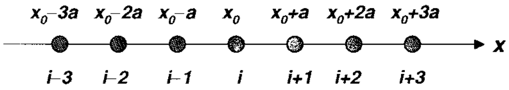


**Figure 2.17.** A one-dimensional, equi-distant mesh for the evaluation of the Schrödinger equation by numerical methods.

series about the points on either side of $x_0$, we get

$$f(x_0+a) = f(x_0) + a\left.\frac{\partial f}{\partial x}\right|_{x=x_0} + \mathrm{O}(a^2) \to f_{i+1} \approx f_i + a\left.\frac{\partial f}{\partial x}\right|_i \qquad (2.120a)$$

$$f(x_0-a) = f(x_0) - a\left.\frac{\partial f}{\partial x}\right|_{x=x_0} + \mathrm{O}(a^2) \to f_{i-1} \approx f_i - a\left.\frac{\partial f}{\partial x}\right|_i . \qquad (2.120b)$$

The factor $\mathrm{O}(a^2)$ is the truncation error. In the two equations on the right-hand side, we have used a short-hand notation for the *node index i*. If we now subtract the two equations on the right-hand side, we obtain an approximate form for the derivative at $x_0$:

$$\left.\frac{\partial f}{\partial x}\right|_i = \frac{f_{i+1} - f_{i-1}}{2a}. \qquad (2.121)$$

We can as easily take an average value in between, and rewrite (2.121) as

$$\left.\frac{\partial f}{\partial x}\right|_{i+1/2} = \frac{f_{i+1} - f_i}{a} \qquad \left.\frac{\partial f}{\partial x}\right|_{i-1/2} = \frac{f_i - f_{i-1}}{a}. \qquad (2.122)$$

These last two forms are important for now developing the second derivative of the function, as

$$\begin{aligned}\left.\frac{\partial^2 f}{\partial x^2}\right|_i &= \frac{\left.\frac{\partial f}{\partial x}\right|_{i+1/2} - \left.\frac{\partial f}{\partial x}\right|_{i-1/2}}{a} = \frac{1}{a}\left(\frac{f_{i+1} - f_i}{a} - \frac{f_i - f_{i-1}}{a}\right)\\ &= \frac{f_{i+1} + f_{i-1} - 2f_i}{a^2}. \end{aligned} \qquad (2.123)$$

Hence, the Schrödinger equation can now be written as

$$-\frac{\hbar^2}{2ma^2}(\Psi_{i+1} + \Psi_{i-1} - 2\Psi_i) + V_i\Psi_i = E\Psi_i. \qquad (2.124)$$

Of course, the trouble with an eigenvalue equation such as the Schrödinger equation is that solutions are only found for certain values of the energy $E$. This means that the energies (eigenvalues) must be found initially before the wave functions can be determined. However, the form (2.124) easily admits to this problem, given a reasonable set of computational routines on a modern computer. Equation (2.124) can be rewritten in the form

$$[S]\Psi] = 0 \qquad (2.125)$$

where $\Psi]$ is a $n \times 1$ column matrix, and $n$ is the number of nodes in the discretization scheme of figure 2.17. The matrix $[S]$ is a tri-diagonal matrix in which all terms are zero except for those on the main diagonal and those once removed from this diagonal. That is, the non-zero terms are

$$S_{ii} = V_i + \frac{\hbar^2}{ma^2} - E \qquad S_{i,i+1} = S_{i,i-1} = -\frac{\hbar^2}{2ma^2}. \tag{2.126}$$

It is immediately obvious from (2.125) that, since the right-hand side is zero, the matrix $[S]$ must be singular. That is,

$$\det[S] = 0 \tag{2.127}$$

is required for any non-trivial solutions of (2.125) to exist. This determinant then gives the $n$ values of the energy $E_n$ that can be found for the $n$ equations. This leads to an important point: we must use many more nodes in the discretization than the number of energy levels we seek. Otherwise, errors arise that are quite large. Even with a large number of nodes, there are errors in the values of the energies because of the truncation used in (2.120). Thus, the numerical solution yields approximate values of the energies and wave functions.

Let us consider an example of an infinite potential well, whose width is 20 nm (we use free electrons). The exact solutions are given in (2.53) in units of the total width of the well as

$$E = \frac{s^2\pi^2\hbar^2}{2mW^2}. \tag{2.128}$$

(The value of $a$ in the referenced equation is one-half the width of the well and $s$ is used here for the integer to avoid confusion with the number of grid points $n$.) Here, we take the discretization variable $a = 1.0\ \mathrm{nm}$, so that $n = 20$. Since the end points $i = 0$ and $i = 20$ both have the wave function zero, we only need to consider the $n - 1$ interior grid points and $[S]$ is a $19 \times 19$ matrix. To compare the computed values, we will give the energies in units of $\pi^2\hbar^2/2mW^2$ so that the actual energy levels are given by $s^2$. We compare the results with the numerically determined values of the lowest six energies in table 2.1. It may be seen that while the results are close, the numerical estimates are somewhat below the actual values.

We now turn to computing the wave functions. Since the wave function has only been evaluated on the nodes, using the energy values $E_n$ in the determinant of $[S]$ gives the set of values the node functions take for each energy. Only $n - 1$ values can be found for each energy level, with the last value needing to be determined from the normalization. It is important to note that the boundary conditions at $i = 0$, and $i = n + 1$ must be imposed at the beginning in establishing (2.125). In figure 2.18, we plot the first and fourth wave functions for this example, showing the exact solutions, which are

$$\Psi_s = \sqrt{\frac{2}{W}}\sin\left(\frac{s\pi x}{W}\right) \tag{2.129}$$

**Table 2.1.** Comparison of numerically computed energies with the actual values.

| $n$ | Actual | Estimate |
|---|---|---|
| 1 | 1 | 0.997 95 |
| 2 | 4 | 3.967 21 |
| 3 | 9 | 8.834 68 |
| 4 | 16 | 15.480 5 |
| 5 | 25 | 23.741 03 |
| 6 | 36 | 33.412 87 |

with the numerically computed eigenvalues. The solid curves are the exact values from (2.129) while the solid points are the numerically computed values at the grid points. It may be seen that these appear to be fairly accurate given the large scale of the figure.

When the problem is time varying, the numerical approach is more complicated. The Schrödinger equation is a diffusion equation, but with complex coefficients. This becomes more apparent if we rewrite (2.8) as

$$\frac{\partial \Psi}{\partial t} = \frac{\mathrm{i}\hbar}{2m}\frac{\partial^2 \Psi}{\partial x^2} + \frac{1}{\mathrm{i}\hbar}V(x,t)\Psi. \tag{2.130}$$

The quantity $\hbar/2m$ has the units of a diffusion constant $(\mathrm{cm}^2\ \mathrm{s}^{-1})$. Nevertheless, we can use the expansion $(2.120a)$ for the time derivative as

$$f_i(t_0+\Delta t) = f_i(t_0)+\Delta t\frac{\partial f_i}{\partial t}\bigg|_{t=t_0} + \mathrm{O}(a^2) \to f_i^{n+1} \approx f_i^n + \Delta t \frac{\partial f_i}{\partial t}\bigg|_n . \tag{2.131}$$

We have left the subscript $i$ to denote the position at which the function is evaluated and used the superscript to denote the time evolution, in which $t = n\Delta t$. Hence, the explicit first-order evaluation of (2.130) is given by

$$\psi_i^{n+1} = \psi_i^n + \frac{\mathrm{i}\hbar\Delta t}{2ma^2}(\psi_{i+1}^n + \psi_{i-1}^n - 2\psi_i^n) - \frac{\mathrm{i}\Delta t}{\hbar}V_i\psi_i^n. \tag{2.132}$$

This is basically the simplest approach and develops the value at grid point $i$ and new time $t + \Delta t$ in terms of the function evaluated at the preceding time step.

There are clearly errors associated with the time step $\Delta t$ and the space step $a$. In fact, if these are too big, then an instability develops in the solution. We can check the linear stability by assuming that the wave function is composed of a set of Fourier modes

$$\psi = \hat{\psi}(t)\mathrm{e}^{\mathrm{i}qx} \tag{2.133}$$

for which (2.132) becomes (we ignore the potential at this time)

$$\hat{\psi}^{n+1}\mathrm{e}^{\mathrm{i}qx} = \hat{\psi}^n\mathrm{e}^{\mathrm{i}qx} + \frac{\mathrm{i}\hbar\Delta t}{2ma^2}(\hat{\psi}^n\mathrm{e}^{\mathrm{i}q(x+a)} + \hat{\psi}^n\mathrm{e}^{\mathrm{i}q(x-a)} - \hat{\psi}^n\mathrm{e}^{\mathrm{i}qx}) \tag{2.134}$$

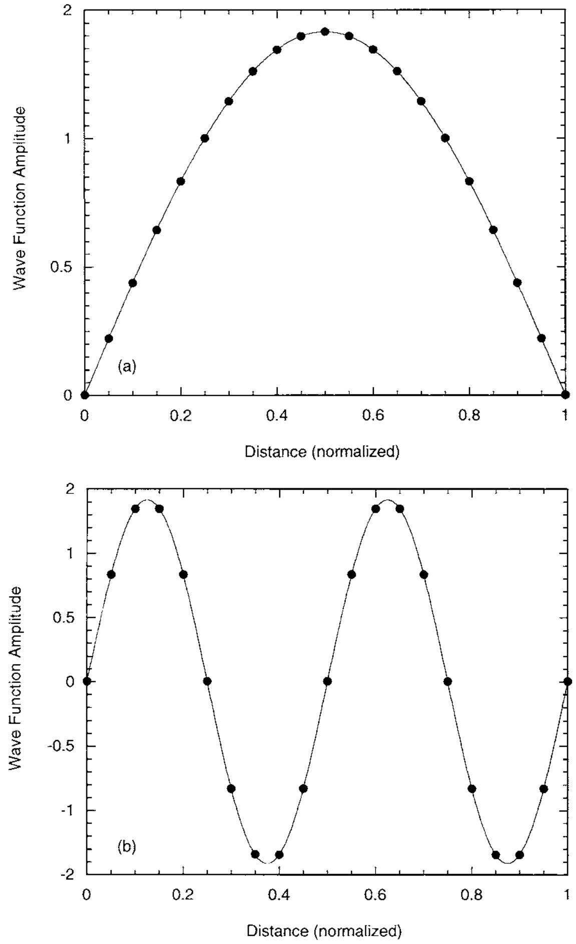


**Figure 2.18.** A comparison of the exact wave function with a numerically computed estimate for an infinite quantum well. In (*a*), $\Psi_1$ is shown, while in (*b*), $\Psi_4$ is plotted.

or

$$\begin{aligned}
\hat{\psi}^{n+1} &= \hat{\psi}^n + \frac{\mathrm{i}\hbar\Delta t}{2ma^2}(\hat{\psi}^n \mathrm{e}^{\mathrm{i}qa} + \hat{\psi}^n \mathrm{e}^{-\mathrm{i}qa} - \hat{\psi}^n) \\
&= \hat{\psi}^n + \frac{\mathrm{i}\hbar\Delta t}{ma^2}[\cos(qa) - 1] \\
&= \hat{\psi}^n \left\{1 - \frac{2\mathrm{i}\hbar\Delta t}{ma^2}\sin^2(qa/2)\right\}.
\end{aligned} \tag{2.135}$$

For stability, the term in curly brackets must be less than one for all values of $q$.

This can only occur if

$$-\frac{2\mathrm{i}\hbar\Delta t}{ma^2} \geq -2 \tag{2.136$a$}$$

or

$$\Delta t \leq \frac{ma^2}{2\hbar}. \tag{2.136$b$}$$

For example, for the previous example, in which $a = 1\ \mathrm{nm}$, the time step must be smaller than $4.3 \times 10^{-15}$ s for a free electron and less than $2.9 \times 10^{-16}$ s for an electron in GaAs. This is a severe limitation for many applications.

An improvement is to go to a second-order method, due to Crank and Nicholson (1947) (see also Richtmyer and Morton 1967). In this, a two-step approach is utilized in which an estimate for $\psi_i^{n+1}$ is obtained using (2.132). Once this estimate is obtained, an improved value is found by including this in the update of the final value of the wave function via

$$\begin{aligned}\psi_i^{n+1} = \psi_i^n &+ \frac{\mathrm{i}\hbar\Delta t}{4ma^2}(\psi_{i+1}^n + \psi_{i-1}^n - 2\psi_i^n) - \frac{\mathrm{i}\Delta t}{2\hbar}V_i\psi_i^n \\ &+ \frac{\mathrm{i}\hbar\Delta t}{4ma^2}(\psi_{i+1}^{n+1} + \psi_{i-1}^{n+1} - 2\psi_i^{n+1}) - \frac{\mathrm{i}\Delta t}{2\hbar}V_i\psi_i^{n+1}.\end{aligned} \tag{2.137}$$

In essence, this is a form of a predictor–corrector algorithm to estimate the time variation. The intrinsic advantage is that the growth factor (the bracketed term in (2.135)) becomes

$$\{\ \} \to \frac{1 - \frac{2\mathrm{i}\hbar\Delta t}{ma^2}\sin^2(qa/2)}{1 + \frac{2\mathrm{i}\hbar\Delta t}{ma^2}\sin^2(qa/2)} \tag{2.138}$$

and the magnitude is less than unity for all values of the coefficients. This means that any combination of $a$ and $\Delta t$ can be used, and the solution is, in principle, stable. This is, of course, an oversimplification, and the solutions must be checked for stability (convergence) with variable values of, for example, $\Delta t$ for a given value of $a$.

## References

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

Brandt S and Dahmen H D 1985 *The Picture Book of Quantum Mechanics* (New York: Wiley)

Cassidy D C 1991 *Uncertainty—the Life and Science of Werner Heisenberg* (New York: Freeman)

Crank J and Nicholson P 1947 *Proc. Camb. Phil. Soc.* **43** 50

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

Eisberg R M 1961 *Fundamentals of Modern Physics* (New York: Wiley)

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** 816–19

Landau R H and Páez M J 1997 *Computational Physics; Problem Solving with Computers* (New York: Wiley)

Moore W 1989 *Schrödinger, Life and Thought* (Cambridge: Cambridge University Press)

Pais A 1991 *Neils Bohr's Time* (Oxford: Clarendon)

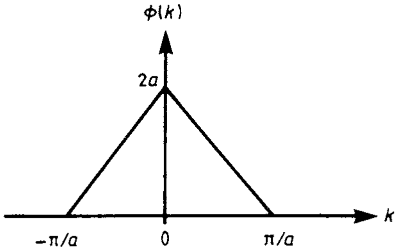
Potter D 1973 *Computational Physics* (Chichester: Wiley)

Richtmyer R D and Morton K W 1967 *Difference Methods for Initial-Value Problems* (New York: Interscience)

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 $\Delta x$ and $\Delta k$?



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.

12. For an infinite potential well of width 20 nm, compute the energies associated with the transitions between the five lowest levels. Give the answers in eV.

13. For a finite potential well of width 20 nm and height 0.3 eV (use an effective mass appropriate to GaAs electrons, $m = 0.067m_0$), compute the energy needed to ionize an electron (move it from the lowest energy level to the top of the well). If a 0.4 eV photon excites the electron out of the well, what is its kinetic energy? What is its wavelength?

14. For a finite potential well of width 20 nm and height 0.3 eV (use an effective mass appropriate to GaAs electrons, $m = 0.067m_0$), carry out a numerical evaluation of the energy levels. Use enough grid points in the 'forbidden' region to assure that the wave function is thoroughly damped. How many bound states are contained in the well?

15. Consider a potential well with $V_0 \to \infty$ for $x < 0$, and $V(x) = Fx$ for $x > 0$, with $F = 0.02\ \text{eV nm}^{-1}$. Using a numerical procedure, compute the ten lowest energy levels in the quantum well.

# Chapter 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.

Since Esaki's discovery of the tunnel diode, tunnelling has become important in a variety of situations. In reverse biased p–n junctions, Zener breakdown occurs when electrons in the valence band can tunnel across the gap into the conduction band when a sufficiently high electric field has been applied to bring these two bands to the same energy levels (on opposite sides of the junction). Similarly, resonant tunnelling diodes have been fabricated in heterostructures such as GaAs–AlGaAs, and we will discuss these in some detail in a later section. Finally, as semiconductor devices become smaller, particularly the metal–oxide–semiconductor field-effect transistor (MOSFET), where a thin layer of silicon dioxide is used as the gate insulator, this thin oxide becomes susceptible to leakage currents via tunnelling through the oxide.

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 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 $\boldsymbol{k}$ in the region $|x| > a$, and the decaying wave vector $\boldsymbol{\gamma}$ 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

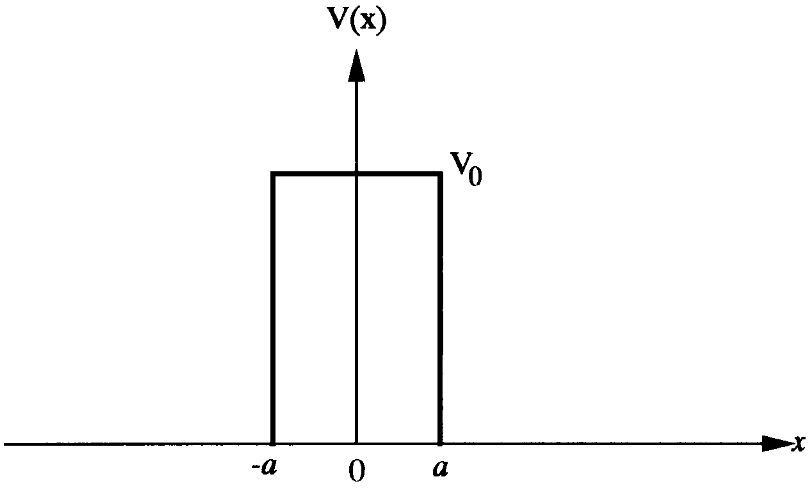


**Figure 3.1.** The simple rectangular tunnelling barrier.

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

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

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

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

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

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

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

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

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

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

Here, the elements are defined by the relations

$$\begin{aligned} M_{11} &= \left(\frac{\mathrm{i}k+\gamma}{2\mathrm{i}k}\right)\left(\frac{\mathrm{i}k+\gamma}{2\gamma}\right)\mathrm{e}^{2(\mathrm{i}k-\gamma)a} - \left(\frac{\mathrm{i}k-\gamma}{2\gamma}\right)\left(\frac{\mathrm{i}k-\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{2(\mathrm{i}k+\gamma)a} \\ &= \left[\cosh(2\gamma a) - \frac{\mathrm{i}}{2}\left(\frac{k^2-\gamma^2}{k\gamma}\right)\sinh(2\gamma a)\right]\mathrm{e}^{2\mathrm{i}ka} && (3.8) \\ M_{21} &= \left(\frac{\mathrm{i}k+\gamma}{2\gamma}\right)\left(\frac{\mathrm{i}k-\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{-2\gamma a} - \left(\frac{\mathrm{i}k+\gamma}{2\mathrm{i}k}\right)\left(\frac{\mathrm{i}k-\gamma}{2\gamma}\right)\mathrm{e}^{2\gamma a} \\ &= -\frac{\mathrm{i}}{2}\left(\frac{k^2+\gamma^2}{k\gamma}\right)\sinh(2\gamma a) && (3.9) \\ M_{22} &= M_{11}^* \qquad M_{12} = M_{21}^*. && (3.10) \end{aligned}$$

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

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

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 $\gamma \to -\mathrm{i}k'$. 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 (2.75) 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

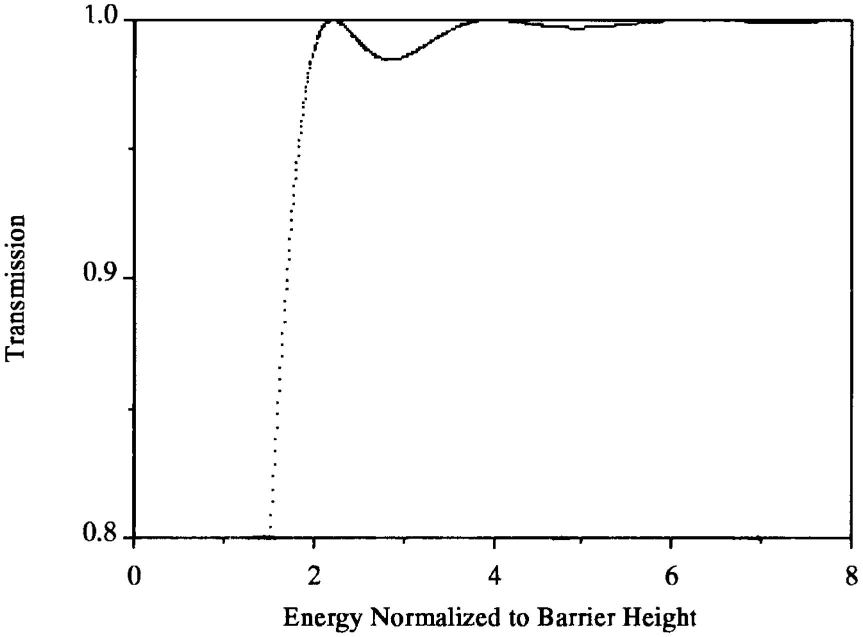


**Figure 3.2.** Tunnelling (transmission) probability for a simple barrier (for generic values).

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

$$E\mathrm{e}^{\mathrm{i}k_1 a} + F\mathrm{e}^{-\mathrm{i}k_1 a} = C\mathrm{e}^{\gamma a} + D\mathrm{e}^{-\gamma a} \tag{3.17$a$}$$

$$\mathrm{i}k_1[E\mathrm{e}^{\mathrm{i}k_1 a} - F\mathrm{e}^{-\mathrm{i}k_1 a}] = \gamma[C\mathrm{e}^{\gamma a} - D\mathrm{e}^{-\gamma a}]. \tag{3.17$b$}$$

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

$$\begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} \left(\dfrac{\mathrm{i}k_1 + \gamma}{2\gamma}\right)\mathrm{e}^{(\mathrm{i}k_1 - \gamma)a} & -\left(\dfrac{\mathrm{i}k_1 - \gamma}{2\gamma}\right)\mathrm{e}^{-(\mathrm{i}k_1 + \gamma)a} \\ -\left(\dfrac{\mathrm{i}k_1 - \gamma}{2\gamma}\right)\mathrm{e}^{(\mathrm{i}k_1 + \gamma)a} & \left(\dfrac{\mathrm{i}k_1 + \gamma}{2\gamma}\right)\mathrm{e}^{-(\mathrm{i}k_1 - \gamma)a} \end{bmatrix} \begin{bmatrix} E \\ F \end{bmatrix}. \tag{3.18}$$

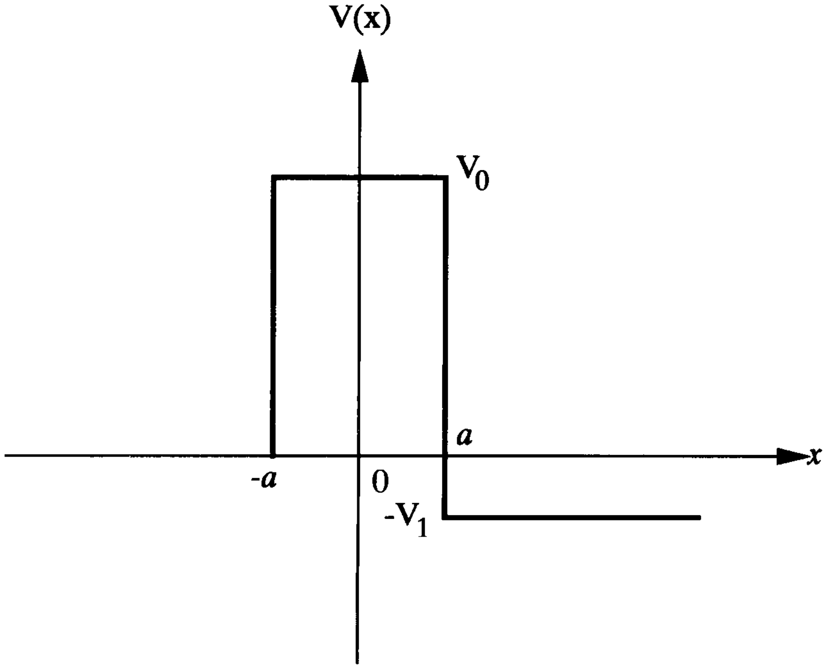


**Figure 3.3.** A more complex tunnelling barrier.

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

$$\begin{aligned} M_{11} &= \left(\frac{\mathrm{i}k+\gamma}{2\mathrm{i}k}\right)\left(\frac{\mathrm{i}k_1+\gamma}{2\gamma}\right)\mathrm{e}^{(\mathrm{i}k+\mathrm{i}k_1-2\gamma)a} \\ &\quad -\left(\frac{\mathrm{i}k_1-\gamma}{2\gamma}\right)\left(\frac{\mathrm{i}k-\gamma}{2\mathrm{i}k}\right)\mathrm{e}^{(\mathrm{i}k+\mathrm{i}k_1+2\gamma)a} \\ &= \left[\frac{1}{2}\left(1+\frac{k_1}{k}\right)\cosh(2\gamma a)-\frac{\mathrm{i}}{2}\left(\frac{kk_1-\gamma^2}{k\gamma}\right)\sinh(2\gamma a)\right]\mathrm{e}^{\mathrm{i}(k+k_1)a} \end{aligned} \tag{3.19}$$

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

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 (2.45)–(2.48), 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 principle 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

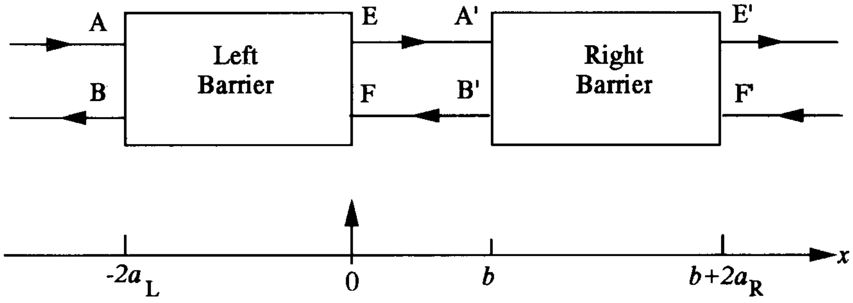


**Figure 3.4.** Two generic barriers are put together to form a double-barrier structure.

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 $k$s and $\gamma$), 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 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

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

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

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

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

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

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

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

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 $\gamma$. 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 $\mathrm{e}^{\mathrm{i}2ka}$ 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 $\mathrm{e}^{-\mathrm{i}2ka}$, and the left-hand barrier contributes a factor of $\mathrm{e}^{\mathrm{i}2ka}$, so the two cancel each other. In order to simplify the mathematical details, we write the remainder of (3.8) as

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

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

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

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

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

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

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

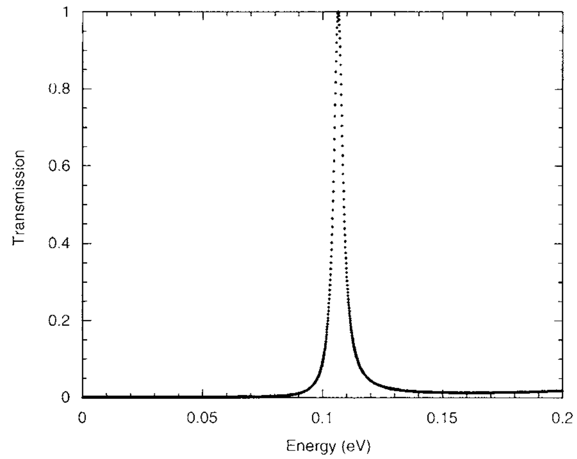


**Figure 3.5.** The transmission through a double barrier system composed of AlGaAs barriers and a GaAs well. Here, the barrier height is 0.25 eV and the thicknesses are 4 nm. The well is taken to be 2 nm thick so that only a single resonant level is present in the well.

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

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

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 $\theta$, 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.

In figure 3.5, we plot the transmission for a double quantum well in which the barriers are 0.25 eV high and 4 nm thick. The well is taken to be 2 nm thick. Here, it is assumed that the well is GaAs and the barriers are AlGaAs, so that there is only a single resonant level in the well. It is clear that the transmission rises to unity at the value of this level. We will return later to the shape of the resonant transmission peak, and how it may be probed in some detail experimentally.

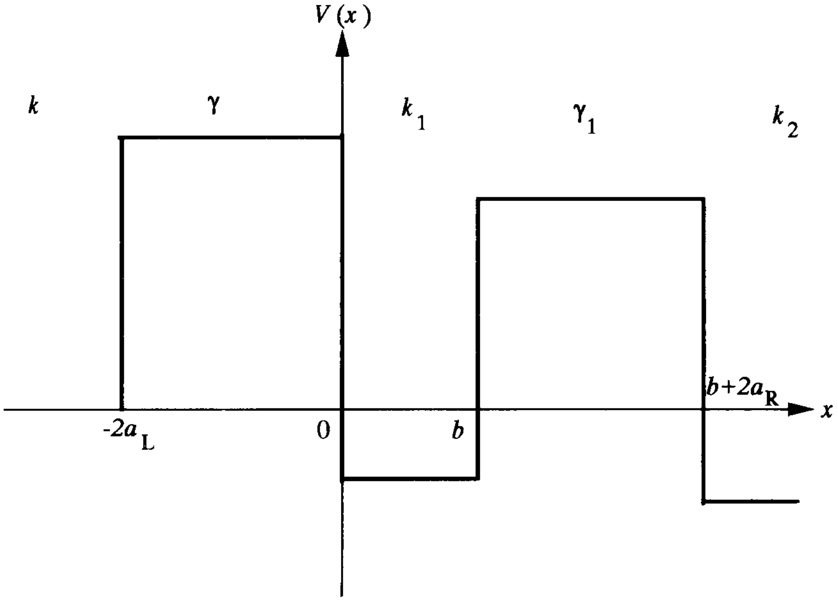


**Figure 3.6.** The potential structure for a general double barrier. The definition of the various constants is given for each region.

### 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 3.6. 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

$$M_i = m_i \mathrm{e}^{\mathrm{i}\theta_i} \tag{3.34}$$

where $i \equiv \mathrm{L11}, \mathrm{L12}, \mathrm{R11}, \mathrm{R21}$. This leads to

$$m_{\mathrm{L11}} = \sqrt{\frac{1}{4}\left(1+\frac{k_1}{k}\right)^2 \cosh^2(2\gamma a_{\mathrm{L}}) + \frac{1}{4}\left(\frac{kk_1-\gamma^2}{k\gamma}\right)^2 \sinh^2(2\gamma a_{\mathrm{L}})} \tag{3.35}$$

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

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

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

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

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

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

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

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:

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

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 $k_1(a_{\mathrm{L}} - a_{\mathrm{R}})/2$. 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

$$\begin{aligned} |M_{\mathrm{T}11}|^2 &= (m_{\mathrm{L}11}m_{\mathrm{R}11} - m_{\mathrm{L}12}m_{\mathrm{R}21})^2 \\ &= m_{\mathrm{L}11}^2 m_{\mathrm{R}11}^2 \left(1 - \frac{m_{\mathrm{L}12}m_{\mathrm{R}21}}{m_{\mathrm{L}11}m_{\mathrm{R}11}}\right)^2 . \end{aligned} \tag{3.44}$$

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

$$\begin{aligned} \left(\frac{m_{\mathrm{L}12}}{m_{\mathrm{L}11}}\right)^2 &= \frac{1 + \dfrac{\gamma^2(k_1 - k)^2}{(\gamma^2 + k^2)(\gamma^2 + k_1^2)\sinh^2(2\gamma a_{\mathrm{L}})}}{1 + \dfrac{\gamma^2(k_1 + k)^2}{(\gamma^2 + k^2)(\gamma^2 + k_1^2)\sinh^2(2\gamma a_{\mathrm{L}})}} \\ &\simeq 1 - \frac{4kk_1\gamma^2}{(\gamma^2 + k^2)(\gamma^2 + k_1^2)\sinh^2(2\gamma a_{\mathrm{L}})} \simeq 1 - T_{\mathrm{L}} \end{aligned} \tag{3.45}$$

and

$$\frac{m_{\mathrm{L}12}}{m_{\mathrm{L}11}} \simeq 1 - \frac{T_{\mathrm{L}}}{2} \tag{3.46}$$

where we have used (3.21) in the limit $2\gamma a_{\mathrm{L}} \gg 1$. 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)

$$T = \frac{T_{\mathrm{L}}T_{\mathrm{R}}}{[1 - (1 - T_{\mathrm{L}}/2)(1 - T_{\mathrm{R}}/2)]^2} \simeq \frac{4T_{\mathrm{L}}T_{\mathrm{R}}}{(T_{\mathrm{L}} + T_{\mathrm{R}})^2}. \tag{3.47}$$

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

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

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

$$T = T_{\mathrm{L}}T_{\mathrm{R}}/2 \tag{3.48}$$

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.

### 3.3.3 Shape of the resonance

It is apparent from the shape of the transmission curve in figure 3.5 that there is a very sharp resonance that coincides with the resonant energy level in the quantum well. We can analyse this resonance more by considering it to have a Lorentzian shape around the maximum, that is

$$\frac{1}{T(E)} \approx 1 + \left(\frac{E - E_0}{\Delta E}\right)^2 . \tag{3.49}$$

In fact, this form holds quite accurately over orders of magnitude in transmission (Price 1999). In essence, this means that we can write $|M_{11}|^2 = PQ$, with $P = Q^*$. From (3.27), we identify

$$M_{11} = Q = Q_0 \left[1 - \mathrm{i}\frac{E - E_0}{\Delta E}\right] \tag{3.50}$$

and

$$M_{11}^* = P = P_0 \left[1 + \mathrm{i}\frac{E - E_0}{\Delta E}\right] . \tag{3.51}$$

Hence, $PQ$ must have two zeros located at $E = E_0 \pm \Delta E$. One of these is a zero of $P$ and the other is a zero of $Q$. Thus, when the energy $E$ is swept through the resonance energy $E_0$, the phase angle $\arg(P)$ is swept through an amount approaching $\pi$. Price (1999) has pointed out that this has an analogy with formal wave propagation theory, so that one can assign a *transit time* for passage of an electron through the double barrier structure as

$$t_{\mathrm{transit}} = \hbar\frac{\mathrm{d}(\arg(P))}{\mathrm{d}E} \to \frac{\hbar/\Delta E}{1 + \left(\frac{E-E_0}{\Delta E}\right)^2} . \tag{3.52}$$

In figure 3.5, the width at half-maximum is about 4.5 meV, so at the resonance, the transit time is approximately 0.15 ps. Away from resonance, the transit time is smaller, so that the transit time *lengthens* at the resonance energy, an effect first noted by Bohm (1951).

While it is quite difficult to measure the transit time of the electron through the resonant structure, creative methods have been found to measure the phase shift. In chapter 1, we discussed the experiments of Yacoby *et al* (1994) to create an Aharonov–Bohm loop in a semiconductor nanostructure. In subsequent work, they embedded a semiconductor quantum dot in one leg of the ring (Yacoby *et al* 1995). The structure is shown in figure 3.7. The quantum dot is a two-dimensional electron gas at the interface of a GaAlAs/GaAs heterostructure, which is further confined by lateral in-plane gates. Propagation through the quantum dot is facilitated by two *tunnelling* quantum point contacts which, along with the dot itself, are located in one leg of the Aharonov–Bohm (A–B) loop. Measurements of the transmitted current through the A–B ring give the total phase through each

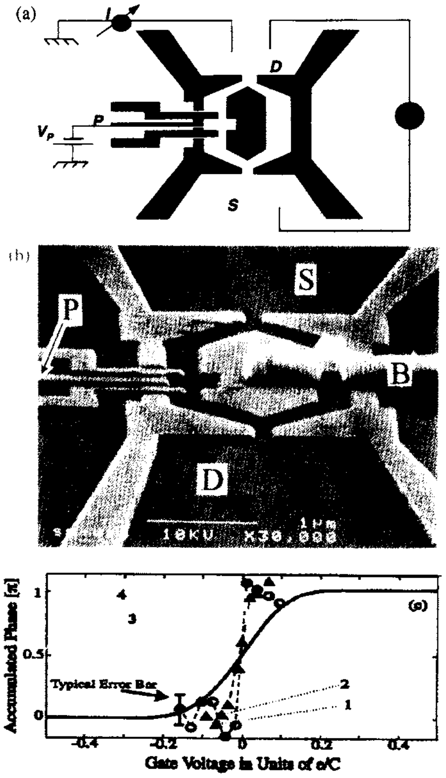


**Figure 3.7.** (*a*) A schematic description of the modified Aharonov–Bohm ring's circuit. The shaded regions are metallic gates. (*b*) An SEM micrograph of the structure. The white regions are the metal gates. The central metallic island is biased via an air bridge (B) extending to the right. The dot is tuned by the plunger (P). Summary of the experimentally measured phases within two transmission peaks (○ and ▲; the broken lines are only guides to the eye). The expected behaviour of the phase in a one-dimensional resonant tunnelling model is shown by the solid line. (After Yacoby *et al* (1995), by permission.)

branch in terms of the interference between the two legs. This phase interference is measured by varying the magnetic field through the ring. When the size of the dot is tuned by a *plunger* gate (P in figure 3.7), the resonant energy is swept through the Fermi energy of the remaining parts of the system. By this gate voltage tuning the resonance shape is swept through the transmission energy, and a change in phase by a factor of $\pi$ is found, as shown in figure 3.7(*c*). This shift is expected from the previous discussion, but the transition is sharper than the authors expected. Several reasons were presented by Yacobi *et al* (1995) for this, but it may simply be that the sharpness of the resonant transmission peak results from a simple two-barrier effect, with the remaining properties of the quantum dot irrelevant to the overall experiment. That is, the phase behaviour of the electrons in the dot is irrelevant to resonant transmission through discrete dot levels coupled by the quantum point contacts.

## 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 (WKB) approach (Wentzel 1926, Kramers 1926, Brillouin 1926).

Consider figure 3.8, 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

$$k(x) = \sqrt{\frac{2m}{\hbar^2}[E - V(x)]} \qquad \mathcal{E} > V(x) \tag{3.53}$$

and

$$\gamma(x) = \sqrt{\frac{2m}{\hbar^2}[V(x) - E]} \qquad \mathcal{E} < V(x). \tag{3.54}$$

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 (3.53) and (3.54) as the lowest approximation, but seek higher approximations. To proceed, we assume that the wave function is generically definable as

$$\Psi(x) \sim \mathrm{e}^{\mathrm{i}u(x)} \tag{3.55}$$

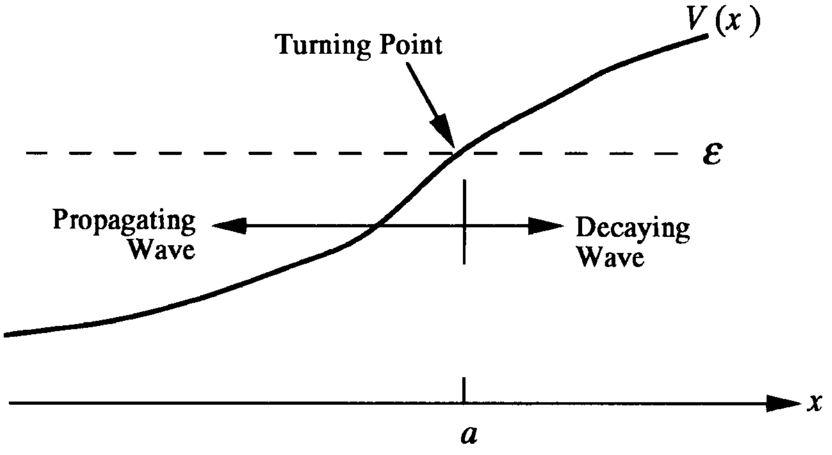


**Figure 3.8.** A simple variation of potential and the corresponding energy surface.

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 (2.9) when the variation of the pre-factor of the exponent is ignored. This gives

$$\mathrm{i}\frac{\partial^2 u}{\partial x^2} - \left(\frac{\partial u}{\partial x}\right)^2 + k^2(x) = 0 \tag{3.56}$$

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

$$\mathrm{i}\frac{\partial^2 u}{\partial x^2} = 0. \tag{3.57}$$

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 $i$th iteration of the solution as the solution of

$$\left(\frac{\partial u_i}{\partial x}\right)^2 = k^2(x) + \mathrm{i}\frac{\partial^2 u_{i-1}}{\partial x^2}. \tag{3.58}$$

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 (3.58)) into the equation for the first-order approximation leads to

$$\frac{\partial u_1}{\partial x} = \sqrt{k^2(x) + \mathrm{i}\frac{\partial k}{\partial x}} \simeq \pm k(x) + \mathrm{i}\frac{1}{2k(x)}\frac{\partial k}{\partial x}. \tag{3.59}$$

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 (3.59) 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 (3.59) can now be integrated over the position, with an arbitrary initial position as the reference point. This gives

$$u_1 \simeq \pm \int^x k(x')\,\mathrm{d}x' + \frac{\mathrm{i}}{2}\ln k(x) + \ln C_1 \tag{3.60}$$

which leads to

$$\Psi(x) \sim \frac{C_1}{\sqrt{k(x)}} \exp\left[\pm\mathrm{i}\int^x k(x')\,\mathrm{d}x'\right]. \tag{3.61}$$

The equivalent solution for the decaying wave function is

$$\Psi(x) \sim \frac{C_1}{\sqrt{\gamma(x)}} \exp\left[\pm\int^x \gamma(x')\,\mathrm{d}x'\right]. \tag{3.62}$$

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.

To see how this occurs, we remind ourselves of (2.14) and (2.16). There, it was necessary to define a current which was continuous in the time-independent situation. Using (2.18), we then require the continuity of

$$|J| = |p\psi^2|. \tag{3.63}$$

This continuity of current was used both in the last chapter, and in this chapter, to determine the transmission coefficient. Now, if we are to infer a connection formula for the wave function, then we must use a generalization of (3.63) to say that

$$\frac{1}{\sqrt{|p|}}\psi \tag{3.64}$$

must be continuous. It is the recognition of this fact that leads to the forms (3.61) and (3.62) for the wave functions on either side of the interface of figure 3.8. These connection formulas are the heart of the WKB technique, but they have their source in the decomposition of the wave function discussed in section 2.2.

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

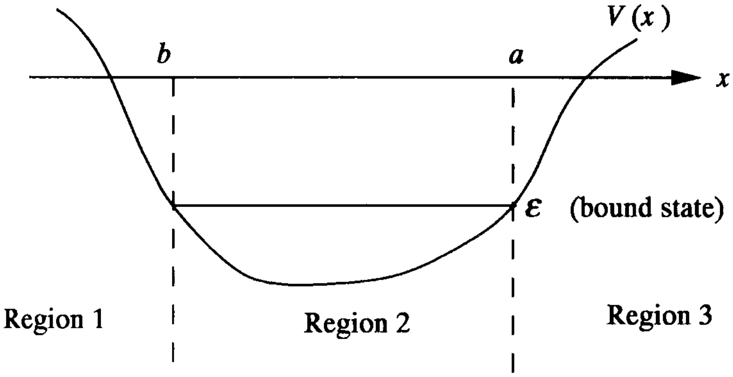


**Figure 3.9.** An arbitrary potential well in which to apply the WKB method.

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 3.8, the turning point is to the right of the classical region (where $\mathcal{E} > V$). For this case, the connection formulas are given by

$$\frac{2}{\sqrt{k}}\cos\left(\int_x^a k\,\mathrm{d}x' - \frac{\pi}{4}\right) \leftrightarrow \frac{1}{\sqrt{\gamma}}\exp\left(-\int_a^x \gamma\,\mathrm{d}x'\right) \tag{3.65}$$

$$\frac{2}{\sqrt{k}}\sin\left(\int_x^a k\,\mathrm{d}x' - \frac{\pi}{4}\right) \leftrightarrow \frac{1}{\sqrt{\gamma}}\exp\left(\int_a^x \gamma\,\mathrm{d}x'\right). \tag{3.66}$$

The alternative case is for the mirror image of figure 3.8, 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)

$$\frac{1}{\sqrt{\gamma}}\exp\left(-\int_x^b \gamma\,\mathrm{d}x'\right) \leftrightarrow \frac{2}{\sqrt{k}}\cos\left(\int_b^x k\,\mathrm{d}x' - \frac{\pi}{4}\right) \tag{3.67}$$

$$-\frac{1}{\sqrt{\gamma}}\exp\left(\int_x^b \gamma\,\mathrm{d}x'\right) \leftrightarrow \frac{2}{\sqrt{k}}\sin\left(\int_b^x \gamma\,\mathrm{d}x' - \frac{\pi}{4}\right). \tag{3.68}$$

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 3.9. 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

$$\Psi_1(x) \simeq \frac{1}{\sqrt{\gamma}} \exp\left( - \int_x^b \gamma \, \mathrm{d}x' \right) \qquad x < b. \tag{3.69}$$

At $x = b$, this must match to the cosine wave if we use (3.67). Thus, we know that in region 2, the wave function is given by

$$\Psi_2(x) \simeq \frac{2}{\sqrt{k}} \cos\left( \int_b^x k \, \mathrm{d}x' - \frac{\pi}{4} \right) \qquad b < x < a. \tag{3.70}$$

We now want to work our way across to $x = a$, and this is done quite simply with simple manipulations of (3.70), as

$$\begin{aligned}\Psi_2(x) &\simeq \frac{2}{\sqrt{k}} \cos\left( \int_b^x k \, \mathrm{d}x' + \frac{\pi}{4} - \frac{\pi}{2} \right) = \frac{2}{\sqrt{k}} \sin\left( \int_b^x k \, \mathrm{d}x' + \frac{\pi}{4} \right) \\ &= \frac{2}{\sqrt{k}} \sin\left( \int_b^a k \, \mathrm{d}x' - \int_x^a k \, \mathrm{d}x' + \frac{\pi}{4} \right) \\ &= -\frac{2}{\sqrt{k}} \cos\left( \int_b^a k \, \mathrm{d}x' \right) \sin\left( \int_x^a k \, \mathrm{d}x' - \frac{\pi}{4} \right) \\ &\quad + \frac{2}{\sqrt{k}} \sin\left( \int_b^a k \, \mathrm{d}x' \right) \cos\left( \int_x^a k \, \mathrm{d}x' - \frac{\pi}{4} \right).\end{aligned} \tag{3.71}$$

We also know that the solution for the matching at the interface $x = a$ must satisfy (3.65), 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 (3.71). This can *only* be achieved by requiring that

$$\cos\left( \int_b^a k \, \mathrm{d}x' \right) = 0 \tag{3.72}$$

or

$$\int_b^a k \, \mathrm{d}x' = (2n+1)\frac{\pi}{2} \qquad n = 0, 1, 2, \ldots. \tag{3.73}$$

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

If we compare (3.73) with the result for a sharp potential as the infinite quantum well of (2.53), 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 (3.73) 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

$$T \sim \exp\left( -2\int_b^a \gamma\,\mathrm{d}x \right) \tag{3.74}$$

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 (3.74).

## 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 3.10. 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

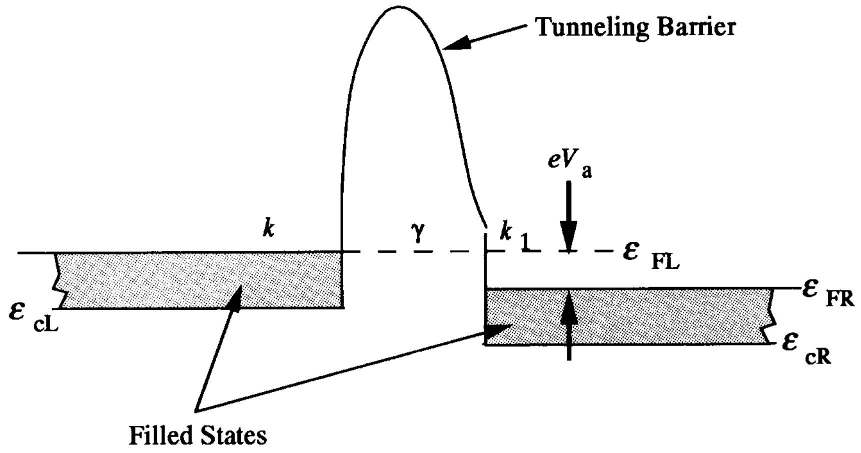


**Figure 3.10.** Tunnelling occurs from filled states on one side of the barrier to the empty states on the opposite side. The current is the net flow of particles from one side to the other.

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 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)

$$\mathcal{E} = \frac{\hbar^2 k_z^2}{2m} = \frac{\hbar^2 k_{1z}^2}{2m} + \text{constant} \tag{3.75}$$

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

$$\nu_z(k_z)\,\mathrm{d}k_z = \nu_z(k_{1z})\,\mathrm{d}k_{1z}. \tag{3.76}$$

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

$$J_{\mathrm{LR}} = 2e \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \nu_z(k_z) T(k_z) f(\mathcal{E}_{\mathrm{L}}) \tag{3.77}$$

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

$$J_\mathrm{RL} = 2e \int \frac{\mathrm{d}^3 k_1}{(2\pi)^3} \nu_z(k_{1z}) T(k_{1z}) f(\mathcal{E}_\mathrm{R}). \tag{3.78}$$

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

$$J = 2e \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \nu_z(k_z) T(k_z)[f(\mathcal{E}_\mathrm{L}) - f(\mathcal{E}_\mathrm{L} + eV_\mathrm{a})] \tag{3.79}$$

where we have related the energy on the left to that on the right through the bias, as shown in figure 3.10, and expressed in (3.75). 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 (3.79). First, we note that the energy is a scalar quantity and can therefore be decomposed into its $z$-component and its transverse component, as

$$\mathcal{E} = \mathcal{E}_z + \mathcal{E}_\perp \tag{3.80}$$

and

$$\mathrm{d}^3 k = \mathrm{d}^2 k_\perp \, \mathrm{d}k_z. \tag{3.81}$$

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

$$\mathrm{d}k_z = \left(\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}k_z}\right)^{-1} \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}\mathcal{E}_z} \, \mathrm{d}\mathcal{E}_z. \tag{3.82}$$

The second term on the right-hand side is unity, so it drops out. The first term may be evaluated from (3.75) as

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}k_z} = \frac{\hbar^2 k_z}{m} = \hbar \nu_z. \tag{3.83}$$

The velocity term here will cancel that in (3.79), and we can write the final representation of the current as

$$J = \frac{e}{\pi\hbar} \int \frac{\mathrm{d}^2 k_\perp}{(2\pi)^2} \int \mathrm{d}\mathcal{E}_z \, T(\mathcal{E}_z)[f(\mathcal{E}_z + \mathcal{E}_\perp) - f(\mathcal{E}_z + \mathcal{E}_\perp + eV_\mathrm{a})]. \tag{3.84}$$

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 3.11 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 (2.10), 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)

$$f_{\mathrm{L}}(-k_z) = Rf_{\mathrm{L}}(k_z) + Tf_{\mathrm{R}}(-k_{1z}) \tag{3.85}$$

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

$$f_{\mathrm{L}}(k_z) - f_{\mathrm{L}}(-k_z) = Tf_{\mathrm{L}}(k_z) - Tf_{\mathrm{R}}(-k_{1z}). \tag{3.86}$$

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

$$f_{\mathrm{R}}(k_{1z}) - f_{\mathrm{R}}(-k_{1z}) = Tf_{\mathrm{L}}(k_z) - Tf_{\mathrm{R}}(-k_{1z}). \tag{3.87}$$

The first thing that is noted from (3.86) and (3.87) 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 (3.84).

To proceed further, we want to dissect the distribution functions in a manner suggested by (3.86) and (3.87). Here, we break each of the two functions into its

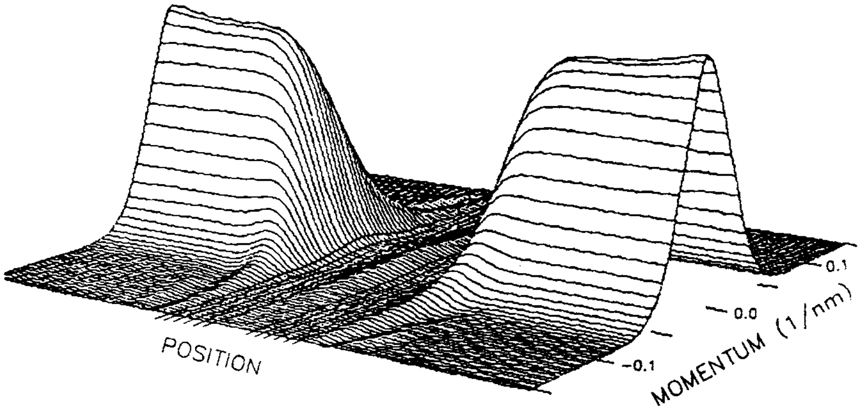


**Figure 3.11.** A quantum charge distribution, with a single tunnelling barrier located in the centre. The charge is plotted as a function of position along one axis and as a function of the $z$-component of momentum along the other. The double-barrier structure is indicated by the heavier lines parallel to the momentum axis that are drawn at the centre of the density.

symmetric and anti-symmetric parts, as

$$f(k_z) = f^{\mathrm{s}}(k_z) + f^{\mathrm{a}}(k_z) \tag{3.88}$$

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

$$f^{\mathrm{s}}(k_z) = \tfrac{1}{2}[f(k_z) + f(-k_z)] \tag{3.89$a$}$$

$$f^{\mathrm{a}}(k_z) = \tfrac{1}{2}[f(k_z) - f(-k_z)]. \tag{3.89$b$}$$

Equations (3.86) and (3.87) 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

$$f_{\mathrm{L}}^{\mathrm{a}}(k_z) = f_{\mathrm{R}}^{\mathrm{a}}(k_{1z}) = f^{\mathrm{a}}(k_z). \tag{3.90}$$

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

$$2f^{\mathrm{a}}(k_z) = T[f_{\mathrm{L}}^{\mathrm{s}}(k_z) - f_{\mathrm{R}}^{\mathrm{s}}(k_{1z})] + 2Tf^{\mathrm{a}}(k_z) \tag{3.91}$$

and

$$f^{\mathrm{a}}(k_z) = \frac{1}{2}\frac{T}{1-T}[f_{\mathrm{L}}^{\mathrm{s}}(k_z) - f_{\mathrm{R}}^{\mathrm{s}}(k_{1z})]. \tag{3.92}$$

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 3.11 is symmetric, but its effect is reflected in the ratio of the transmission to the reflection coefficients that appears in (3.92). 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 3.11 should properly be included. The final equation for the current is then

$$J = \frac{e}{\pi\hbar}\int \frac{\mathrm{d}^2 k_\perp}{(2\pi)^2}\int \mathrm{d}\mathcal{E}_z \frac{T(\mathcal{E}_z)}{1-T(\mathcal{E}_z)}[f^{\mathrm{s}}(\mathcal{E}_z+\mathcal{E}_\perp) - f^{\mathrm{s}}(\mathcal{E}_z+\mathcal{E}_\perp+eV_{\mathrm{a}})]. \tag{3.93}$$

(The factor of 2 in (3.92) cancels when the two distributions in (3.84) 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 3.12(*a*). When a small bias is applied, as shown in figure 3.12(*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 3.12(*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 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,

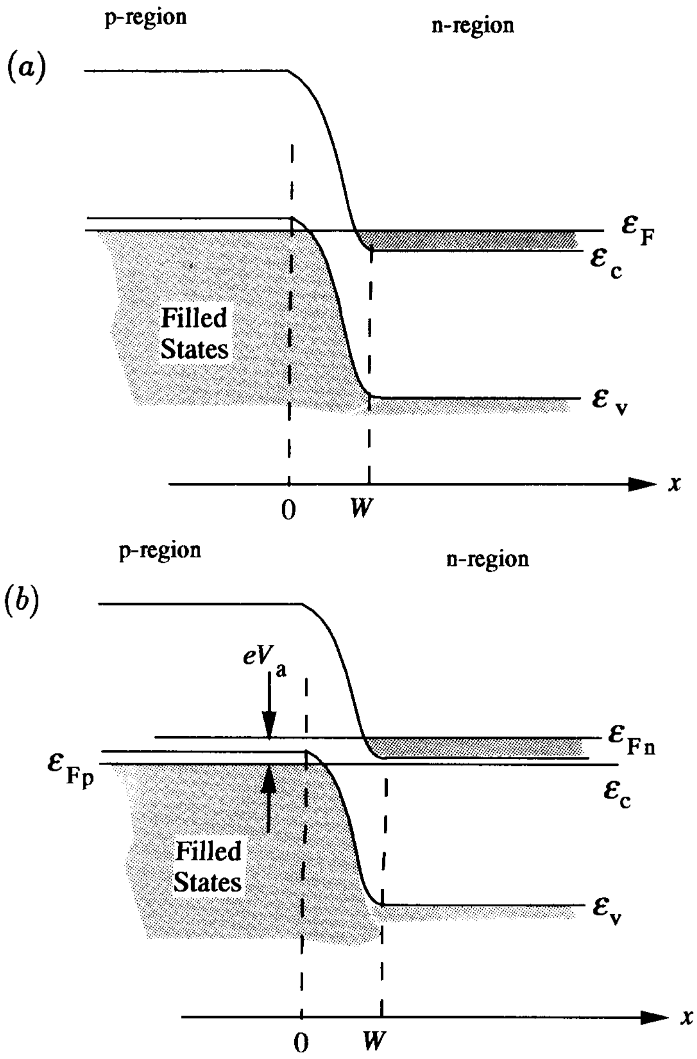


**Figure 3.12.** The band line-up for degenerately doped p–n junctions (*a*), and the possible tunnelling transitions for small forward bias (*b*).

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

$$\gamma(x) \simeq \begin{cases} \sqrt{\dfrac{2m\mathcal{E}_{\mathrm{G}}}{\hbar^2}\left(1 - \dfrac{x}{W} + \dfrac{\mathcal{E}_\perp}{\mathcal{E}_{\mathrm{G}}}\right)} & 0 < x < W \\ 0 & \text{elsewhere} \end{cases} \tag{3.94}$$

where we have factored the energy gap out of the potential term and evaluated the electric field as $\mathcal{E}_{\mathrm{G}}/eW$. 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 (3.74) over the tunnelling region, which produces

$$\begin{aligned} T &\simeq \exp\left[-2\int_0^W \sqrt{\frac{2m\mathcal{E}_{\mathrm{G}}}{\hbar^2}\left(1 - \frac{x}{W} + \frac{\mathcal{E}_\perp}{\mathcal{E}_{\mathrm{G}}}\right)}\,\mathrm{d}x\right] \\ &\simeq \exp\left[-\frac{4W}{3}\sqrt{\frac{2m\mathcal{E}_{\mathrm{G}}}{\hbar^2}}\left(1 + \frac{3\mathcal{E}_\perp}{2\mathcal{E}_{\mathrm{G}}}\right)\right] \end{aligned} \tag{3.95}$$

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 (3.95) 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 (3.95) is probably as good as any other. We can rewrite (3.94) as

$$T \simeq T_0 \exp\left[-\frac{\mathcal{E}_\perp}{\mathcal{E}_0}\right] \tag{3.96}$$

where

$$\mathcal{E}_0 = \frac{eE}{2}\sqrt{\frac{\hbar^2}{2m\mathcal{E}_{\mathrm{G}}}}. \tag{3.97}$$

This can now be used in (3.92) 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

$$\int_0^{\mathcal{E}_{\mathrm{F}} - eV_{\mathrm{a}}} \frac{\mathrm{d}^2 k_\perp}{(2\pi)^2} \exp(-\mathcal{E}_\perp/\mathcal{E}_0) = \frac{m\mathcal{E}_0}{2\pi\hbar^2}[1 - \exp(-(\mathcal{E}_{\mathrm{F}t} - eV_{\mathrm{a}})/\mathcal{E}_0)]. \tag{3.98}$$

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 $eV_\mathrm{a}$. This leads to the result, in the linear-response limit, that

$$\begin{aligned}[f^\mathrm{s}(\mathcal{E}_z+\mathcal{E}_\perp)-f^\mathrm{s}(\mathcal{E}_z+\mathcal{E}_\perp+eV_\mathrm{a})] &\simeq \frac{eV_\mathrm{a}}{k_\mathrm{B}T}f^\mathrm{s}(1-f^\mathrm{s}) \\ &\simeq -eV_\mathrm{a}\frac{\partial f^\mathrm{s}}{\partial\mathcal{E}_z}\simeq eV_\mathrm{a}\delta(\mathcal{E}_z-\mathcal{E}_\mathrm{F}).\end{aligned} \tag{3.99}$$

The last approximation is for strongly degenerate material (or equivalently, very low temperature). Then the integration over $\mathcal{E}_z$ gives just $eV_\mathrm{a}$ times the tunnelling probability $T_0$. We can now put (3.98) and (3.99) in the general equation (3.92) to obtain the total current density

$$J_z=\frac{eT_0}{\pi\hbar}\frac{m\mathcal{E}_{\mathrm{F}t}}{2\pi\hbar^2}\left(1-\frac{eV_\mathrm{a}}{\mathcal{E}_{\mathrm{F}t}}\right)eV_\mathrm{a}. \tag{3.100}$$

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 3.13, 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 3.14. 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

$$[f^\mathrm{s}(\mathcal{E}_z+\mathcal{E}_\perp)-f^\mathrm{s}(\mathcal{E}_z+\mathcal{E}_\perp+eV_\mathrm{a})]\simeq eV_\mathrm{a}\delta(\mathcal{E}_z+\mathcal{E}_\perp-\mathcal{E}_\mathrm{F}). \tag{3.101}$$

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 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

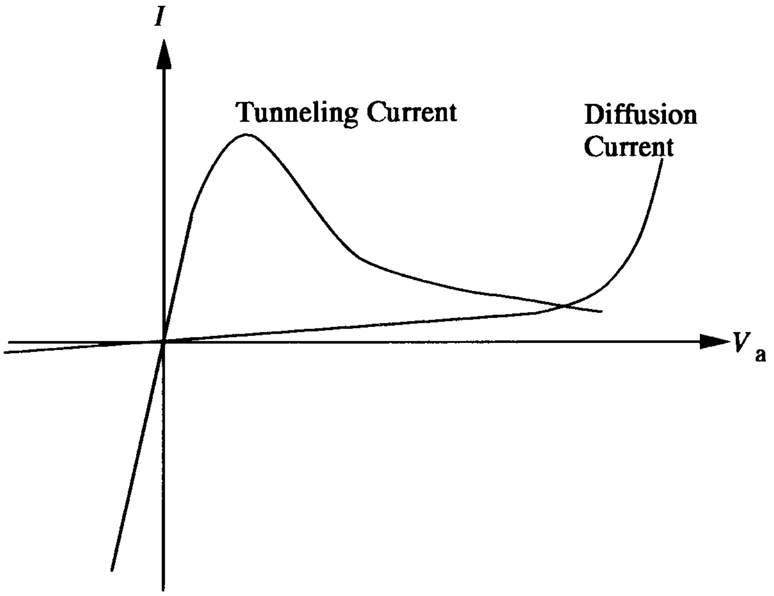


**Figure 3.13.** The contribution of the tunnelling current to the overall current of a tunnel diode.

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)

$$T(\mathcal{E}) \simeq \mathcal{E}_\mathrm{W}\delta(\mathcal{E}_z + eV_\mathrm{a}/2 - \mathcal{E}_0) \tag{3.102}$$

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 (3.92) as

$$\begin{aligned} J_z &= \frac{e^2V_\mathrm{a}}{\pi\hbar}\frac{m\mathcal{E}_\mathrm{W}}{2\pi\hbar^2}\int \mathrm{d}\mathcal{E}_z \int_0^{\mathcal{E}_\mathrm{F}} \delta(\mathcal{E}_z + \mathcal{E}_\perp - \mathcal{E}_\mathrm{F})\delta(\mathcal{E}_z + eV_\mathrm{a}/2 - \mathcal{E}_0)\,\mathrm{d}\mathcal{E}_\perp \\ &= \frac{e^2V_\mathrm{a}}{\pi\hbar}\frac{m\mathcal{E}_\mathrm{W}}{2\pi\hbar^2}\int_0^{\mathcal{E}_\mathrm{F}} \delta(\mathcal{E}_\mathrm{F} - \mathcal{E}_\perp + eV_\mathrm{a}/2 - \mathcal{E}_0)\,\mathrm{d}\mathcal{E}_\perp \\ &= \frac{e^2V_\mathrm{a}}{\pi\hbar}\frac{m\mathcal{E}_\mathrm{W}}{2\pi\hbar^2} \qquad 2(\mathcal{E}_0 - \mathcal{E}_\mathrm{F}) < eV_\mathrm{a} < 2\mathcal{E}_0. \end{aligned} \tag{3.103}$$

Outside 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 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

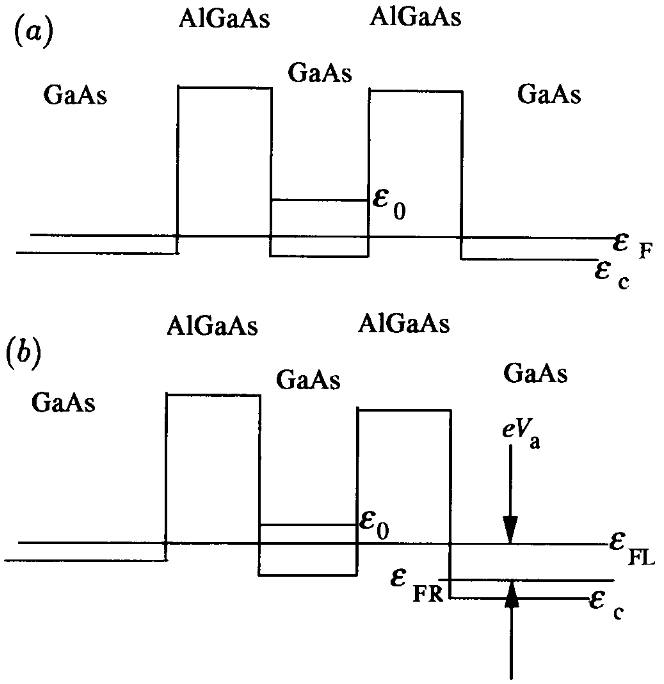


**Figure 3.14.** A typical double-barrier resonant tunnelling diode potential system, grown by heteroepitaxy in the GaAs–AlGaAs system. In (*a*), the basic structure is shown for an n-type GaAs well and cladding. In (*b*), the shape under bias is shown.

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 3.15, while experimentally observed curves are shown in figure 3.16.

### 3.5.4 Resonant interband tunnelling

In the previous section, we dealt with a simple resonant tunnelling diode created by heterostructure growth in the GaAs–AlGaAs system. One advantage of such heterostructure growth is the ability to engineer particular band (and band-gap) alignments by the choice of materials with the only real limitation being the need to incorporate good epitaxial growth during the process. The limitation in the structure of the last section is the relatively high value of the 'valley' current—the current that flows after the peak tunnelling current has decayed. If the devices are to be used in, for example, logic circuits, this valley current is an extraneous source of dissipation in the circuit, and the need to preserve power means that the valley current must be reduced. In most cases, this is cast in a different

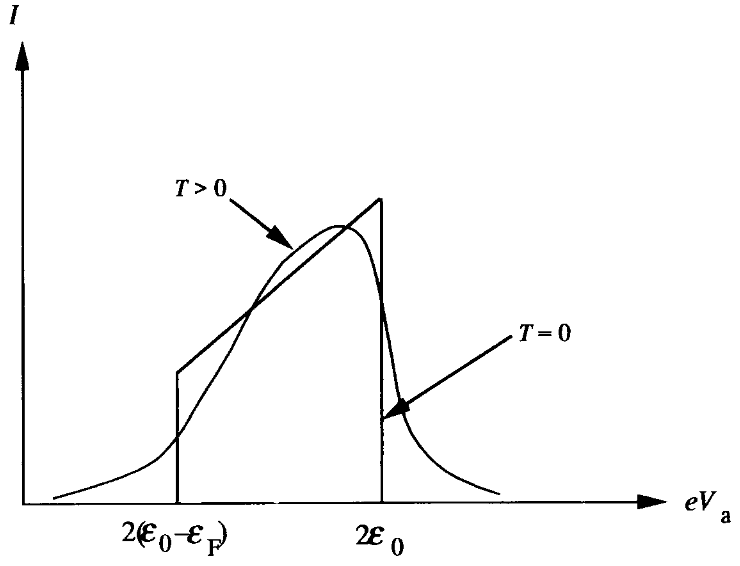


**Figure 3.15.** The theoretical curves for the simple model of (3.103) are shown for zero temperature and finite temperature.

phraseology—the peak-to-valley ratio. The latter is simply the ratio of peak current to valley current, yet the most important aspect is to reduce the valley current to acceptable levels. One way to do this is to increase the barrier energy heights and/or thicknesses to reduce the off-peak tunnelling currents. But, since the barriers are usually not matched under applied bias, this results in reductions of the peak current as well. Another approach is creatively design different structures. One of these is the resonant *interband* tunnelling device.

Consider the band structure alignment shown in figure 3.17, which represents a structure grown in the InAs/AlSb/GaSb system. The basic transport layers are formed in InAs, which is a relatively narrow band-gap system ($E_G \sim 0.4$ eV) with a low effective mass and a high mobility at room temperature. The barriers are formed from AlSb (with an energy gap of 2.2 eV), while the well is formed in GaSb (with an energy gap of 0.67 eV). The advantageous nature of this system is the band lineup, in which the Fermi level of n-type InAs lies in the *valence* band of the GaSb (Söderström *et al* 1989). Hence, the tunnelling transition is from a conduction-band state in the InAs, through a resonant level in the valence band of the GaSb, to another conduction-band state in the final InAs layer. Hence, the resonant transition is an interband one, in which two interband processes are required. With the high barrier in AlSb, the peak current can be maintained by thinning the thickness of these layers. On the other hand, the valley current can be significantly reduced since forward bias will bring the conduction-band

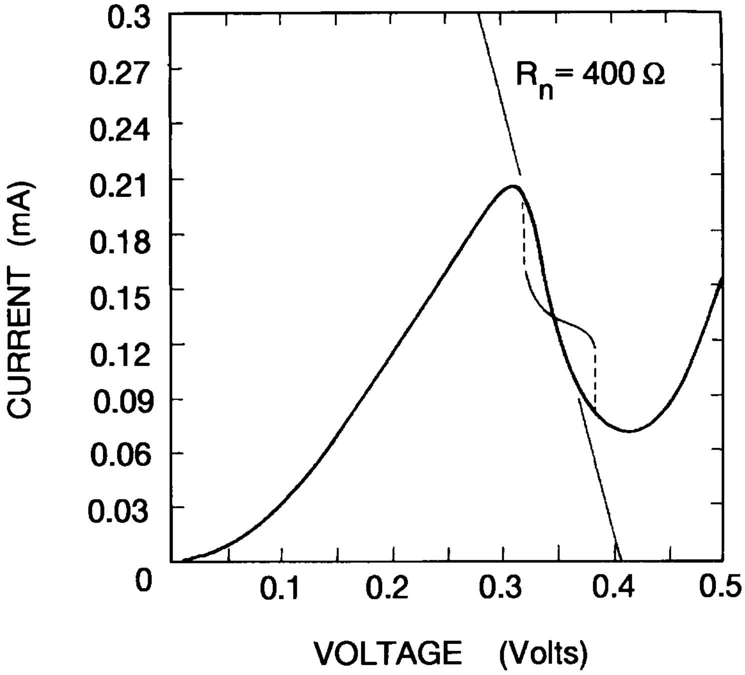


**Figure 3.16.** The experimental curves obtained for a GaAs–AlGaAs structure with 5 nm barriers and a 5 nm well. The extra current above the drop-off is due to higher resonant states and emission over the top of the barrier; both are forms of leakage. (After Sollner *et al* (1983), by permission.)

states of the InAs into alignment with the band gap of GaSb, once the resonant level is surpassed. This transition is now strongly forbidden, which results in a greatly reduced valley current (remaining levels of valley current must arise from thermionic emission over the barriers). The original structures did not have the AlSb barriers, and the tunnelling disappeared if the GaSb layers were too thin (Yu *et al* 1990). Kitabayashi *et al* (1997) recently found that, for 17 monolayers of GaSb (approximately 5.2 nm), peak tunnelling current could be achieved for barriers two monolayers (approximately 0.6 nm) thick. For thinner barriers, the peak current rose, but was attributed to processes other than interband tunnelling. For thicker barriers, both the peak current and the interband tunnelling decreased (although the peak-to-valley ratio could be further increased).

The idea of resonant interband tunnelling can be extended to silicon-based systems through the use of the alloy SiGe (Rommel *et al* 1998). In these structures, the 'resonant levels' actually lie in two induced quantum wells on either side of a SiGe barrier placed between two Si layers. One is an n-type quantum well induced by heavy doping placed in a very thin layer adjacent to the

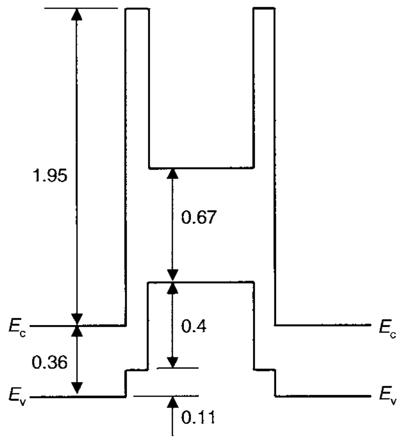


**Figure 3.17.** Band lineup for a GaSb well, AlSb barriers, embedded in InAs. The resonant level will lie in the valence band of the GaSb, while the transport layers are n-type InAs. All energies are shown in eV.

hetero-interface. The silicon transport layer on this side of the barrier is n-type as well. The other quantum well naturally forms in the valence band when strained SiGe is grown on unstrained Si, but this well is further enhanced by implanting a narrow, p-doped region. The transport layer on this side of the barrier is p-type as well. Hence, the structure is a basic Si p–n junction, in which a strained SiGe layer is placed within the depletion region, and resonant levels are induced in dopant-induced quantum wells on either side of the strained layer. Peak-to-valley ratios in the structure are not very good, but the advantage is that this is a resonant interband diode compatible with silicon integrated circuit processing.

### 3.5.5 Self-consistent simulations

When we are dealing with devices such as the resonant-tunnelling diode, it is important to understand that real devices must be solved self-consistently. That is, the actual potential $V(x)$ will change with the applied bias—this change will have a dependence on position and not just the obvious one on the amplitude difference between the two ends of the device. For example, as the voltage is ramped up from zero to a value just beyond the peak current, there is charge in the quantum well (the well is charged as the resonant level is reduced to a position near the cathode Fermi level). However, when approaching the peak current from the high voltage side, the well is empty. Just this simple charging/discharging of the well implies a difference in the detailed positional dependence of the potential. This is an effect in most devices, not merely the tunnelling devices.

The potential is given by a solution to Poisson's equation, which is obtained from electrostatic theory as

$$\frac{\partial^2 V(x,t)}{\partial x^2} = -\frac{\rho(x,t)}{\varepsilon} \tag{3.104}$$

where we have indicated that both the charge density $\rho$ and the potential $V$ are both time and position dependent (we deal only with a one-dimensional problem here). The Schrödinger equation gives us the wave function for a particular potential. This wave function can be used to determine the charge density at a position $x$ through

$$\rho(x) = -e|\psi(x,t)|^2 - \rho_0(x) \tag{3.105}$$

where it is assumed that the wave function represents electrons and that $\rho_0$, the background charge density, is also negative. If the latter is uniform (which is not the usual case), and the wave function is normalized over a distance $L$, (3.105) can be simplified to be

$$\rho(x) = -e\left[|\psi(x,t)|^2 - \frac{1}{L}\right] \tag{3.106}$$

so that

$$\int_0^L \rho(x)\,\mathrm{d}x = -e\left[\int_0^L |\psi(x,t)|^2\,\mathrm{d}x - 1\right]. \tag{3.107}$$

This last form just ensures the normalization of the wave function in the region of interest, and also ensures that the total device is space-charge neutral. This neutrality is a requirement on all devices—the total charge in the device must sum to zero.

The procedure by which a quantum device is solved self-consistently is to set up a recursive loop. The potential determines the wave function through Schrödinger's equation. The wave function determines the charge density, as a function of position, in the device. This charge density then determines the new potential through Poisson's equation. This loop can be iterated until a convergent solution is obtained. This is done at each time step in a time-dependent device. Needless to say, this looping of the equations to reach convergence is a major time consuming process in simulation. The device calculations of the previous sub-sections ignore this need to establish self-consistent solutions. This does not invalidate them; rather, it is necessary to understand that they are at best approximations and estimations of the behaviour of a real device in such circumstances. Such a self-consistent solution is illustrated later in figure 3.20.

## 3.6 The Landauer formula

The general approach that was used to evaluate the current equation (3.93) 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 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 (3.92) over the transverse dimensions, so that it can be rewritten as

$$\begin{aligned} I &= \frac{e}{\pi\hbar}A\int\frac{\mathrm{d}^2k_\perp}{(2\pi)^2}\int \mathrm{d}\mathcal{E}_z\,\frac{T(\mathcal{E}_z)}{1-T(\mathcal{E}_z)}[f^{\mathrm{s}}(\mathcal{E}_z+\mathcal{E}_\perp)-f^{\mathrm{s}}(\mathcal{E}_z+\mathcal{E}_\perp+eV_{\mathrm{a}})] \\ &= \frac{e^2V_{\mathrm{a}}}{\pi\hbar}\frac{mA}{2\pi\hbar^2}\int \mathrm{d}\mathcal{E}_\perp\int \mathrm{d}\mathcal{E}_z\frac{T(\mathcal{E}_z)}{1-T(\mathcal{E}_z)}\delta(\mathcal{E}_z+\mathcal{E}_\perp-\mathcal{E}_{\mathrm{F}}) \\ &= \frac{e^2V_{\mathrm{a}}}{\pi\hbar}\frac{mA}{2\pi\hbar^2}\int_0^{\mathcal{E}_{\mathrm{F}}} \mathrm{d}\mathcal{E}_\perp\frac{T(\mathcal{E}_{\mathrm{F}}-\mathcal{E}_\perp)}{1-T(\mathcal{E}_{\mathrm{F}}-\mathcal{E}_\perp)} \\ &= \frac{e^2V_{\mathrm{a}}}{\pi\hbar}\frac{m\mathcal{E}_{\mathrm{a}}A}{2\pi\hbar^2}\frac{T(\mathcal{E}_{\mathrm{a}})}{1-T(\mathcal{E}_{\mathrm{a}})}. \end{aligned} \tag{3.108}$$

Here, $\mathcal{E}_{\mathrm{a}}$ is an average transverse energy. The second fraction in (3.108) is an interesting quantity, in that it is essentially just $k_{\mathrm{a}}^2A$, where $\boldsymbol{k}_{\mathrm{a}}$ is the wave vector 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 $N_{\mathrm{t}}$, then we can write (3.108) as (we use $I = GV_{\mathrm{a}}$ to write only the conductance $G$)

$$G = \frac{e^2}{\pi\hbar}\sum_{i=1}^{N_{\mathrm{t}}}\frac{T_i}{1-T_i} \tag{3.109}$$

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 (3.108) 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.

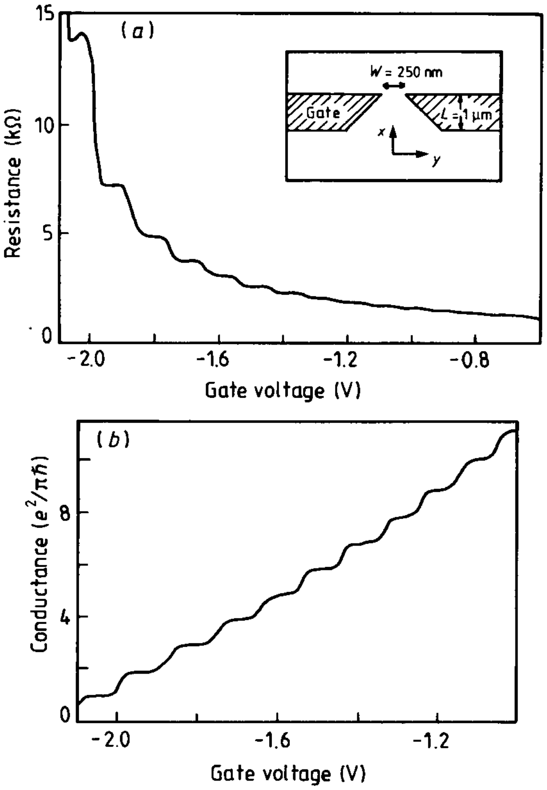


**Figure 3.18.** Quantized resistance (*a*) and conductance (*b*) can be observed in small conducting systems. Here, the number of transverse states in the small opening between the metal gates is varied by changing the bias on the gates (shown in the inset to (*a*)). (After van Wees *et al* (1988), by permission.)

There is an interesting suggestion in (3.108). 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 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 3.18. 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

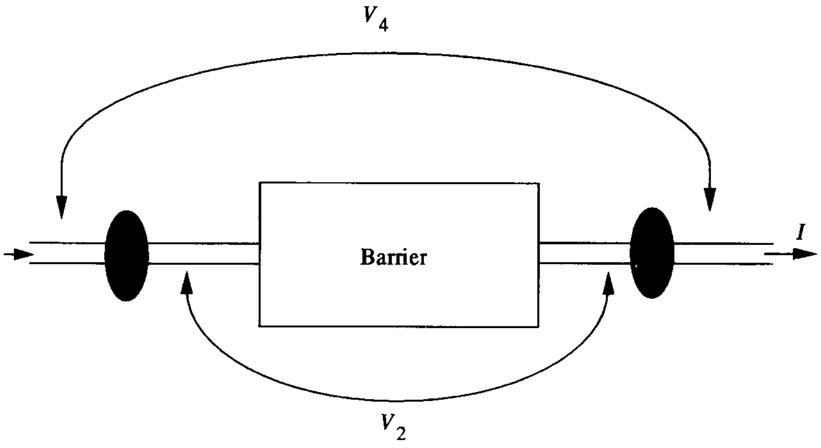


**Figure 3.19.** The two-terminal and four-terminal resistances are defined in terms of the voltages $V_2$ and $V_4$, respectively. They differ by the contact resistances.

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 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 (3.109) can be written as

$$G = \frac{e^2}{\pi\hbar}\frac{T_i}{1-T_i}. \tag{3.110}$$

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 3.19). If we rewrite this equation in terms of resistances, then

$$R_4 = \frac{\pi\hbar}{e^2}\left(\frac{1}{T_i} - 1\right) \tag{3.111}$$

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

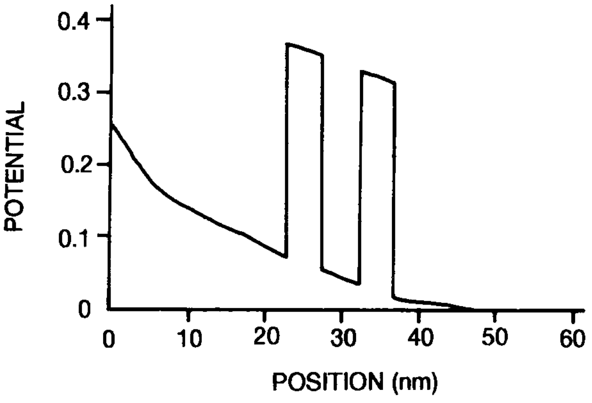


**Figure 3.20.** The potential profile for a resonant tunnelling diode, in which a depletion region (to the left of the barriers) creates a contact resistance to balance the current-carrying preferences of the barriers and the contacts.

difference lies in the fact that the contacts are areas where equilibration occurs. If we recognize that the original form of the current density (3.83) implied a two-terminal definition, we can say that

$$R_2 = \frac{\pi\hbar}{e^2}\frac{1}{T_i} \tag{3.112}$$

and the difference is given by

$$R_2 = R_4 + R_c \qquad R_c = \frac{\pi\hbar}{e^2}. \tag{3.113}$$

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 3.20 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 3.11 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.

## 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 3.21. 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 3.21, 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

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\Psi_1}{\mathrm{d}x^2} - E\Psi_1 = 0 \qquad 0 < x < a \tag{3.114$a$}$$

and

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\Psi_1}{\mathrm{d}x^2} - E\Psi_1 = -V_0\Psi_1 \qquad -b < x < 0. \tag{3.114$b$}$$

Of course, shifts of the $x$-axis by the amount $d$ bring in other regions in which (3.114) 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,

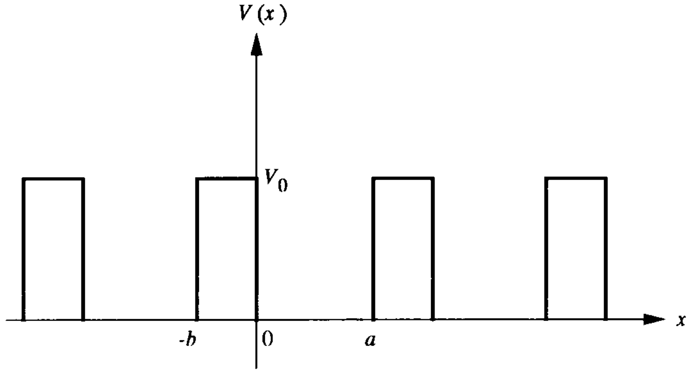


**Figure 3.21.** A simple periodic potential.

or

$$\Psi_1(x) = \mathrm{e}^{\mathrm{i}Kx}u(x) \qquad (3.115)$$

where $u(x)$ has the periodicity of the lattice. A wave of the form (3.115) is termed a Bloch function. If we insert (3.115) into (3.114), the resulting equation is

$$\frac{\mathrm{d}^2u_1}{\mathrm{d}x^2} + 2\mathrm{i}K\frac{\mathrm{d}u_1}{\mathrm{d}x} + (k^2 - K^2)u_1 = 0 \qquad 0 < x < a \qquad (3.116a)$$

and

$$\frac{\mathrm{d}^2u_2}{\mathrm{d}x^2} + 2\mathrm{i}K\frac{\mathrm{d}u_2}{\mathrm{d}x} + (\gamma^2 + K^2)u_2 = 0 \qquad -b < x < 0. \qquad (3.116b)$$

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

$$u_1 = A\mathrm{e}^{-\mathrm{i}(K-k)x} + B\mathrm{e}^{-\mathrm{i}(K+k)x} \qquad 0 < x < a \qquad (3.117a)$$
$$u_2 = C\mathrm{e}^{-(\mathrm{i}K-\gamma)x} + D\mathrm{e}^{-(\mathrm{i}K+\gamma)x} \qquad -b < x < 0. \qquad (3.117b)$$

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 (3.117). 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

$$u_1(0) = u_2(0) \qquad (3.118)$$
$$u_1(a) = u_2(-b) \qquad (3.119)$$

$$\frac{\mathrm{d}u_1(0)}{\mathrm{d}x} = \frac{\mathrm{d}u_2(0)}{\mathrm{d}x} \tag{3.120}$$

$$\frac{\mathrm{d}u_1(a)}{\mathrm{d}x} = \frac{\mathrm{d}u_2(-b)}{\mathrm{d}x}. \tag{3.121}$$

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

$$\begin{vmatrix} 1 & 1 & -1 & -1 \\ \mathrm{e}^{-\mathrm{i}(K-k)a} & \mathrm{e}^{-\mathrm{i}(K+k)a} & -\mathrm{e}^{(\mathrm{i}K-\gamma)b} & -\mathrm{e}^{(\mathrm{i}K+\gamma)b} \\ -\mathrm{i}(K-k) & -\mathrm{i}(K+k) & \mathrm{i}K-\gamma & \mathrm{i}K+\gamma \\ -\mathrm{i}(K-k)\mathrm{e}^{-\mathrm{i}(K-k)a} & -\mathrm{i}(K+k)\mathrm{e}^{-\mathrm{i}(K+k)a} & (\mathrm{i}K-\gamma)\mathrm{e}^{(\mathrm{i}K-\gamma)b} & (\mathrm{i}K+\gamma)\mathrm{e}^{(\mathrm{i}K+\gamma)b} \end{vmatrix}$$
$$= 0. \tag{3.122}$$

Evaluating this determinant leads to

$$\frac{\gamma^2 - k^2}{2k\gamma}\sinh(\gamma b)\sin(ka) + \cosh(\gamma b)\cos(ka) = \cos[K(b+a)]. \tag{3.123}$$

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 $\gamma b$. 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 $\gamma b$ varies as the square root of the potential, this quantity approaches zero, so (3.123) can be rewritten as

$$\frac{\gamma^2 b}{2k}\sin(ka) + \cos(ka) = \cos(Ka). \tag{3.124}$$

The right-hand side of (3.124) 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 (3.124) is satisfied. This range can be found graphically, as shown in figure 3.22 (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 (3.124) 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

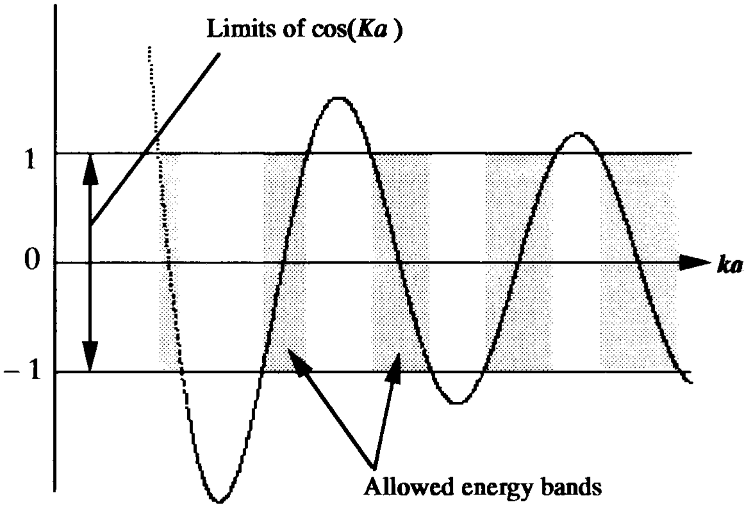


**Figure 3.22.** The allowed and forbidden values of $ka$. The shaded areas represent the allowed range of energy values.

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 (3.124) 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

$$\sin(ka) = 0 \qquad k = \frac{n\pi}{a} \tag{3.125}$$

which produces the bound-state energies of (2.53) (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 3.22 corresponds to one allowed energy band). In figure 3.23, 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 3.23 are termed the *reduced zone scheme* (as opposed to taking $K$ over an infinite range). We note that the momentum $\boldsymbol{K}$ (or more properly $\hbar\boldsymbol{K}$) 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 $\boldsymbol{K}$.

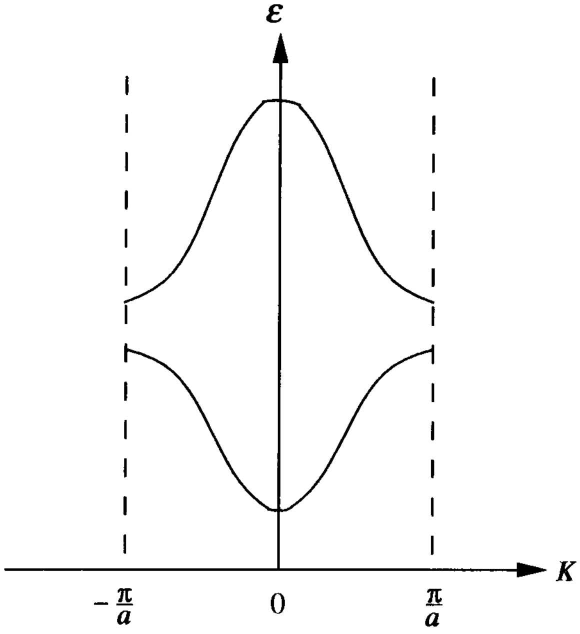


**Figure 3.23.** The energy band structure that results from the solution diagram of figure 3.22.

### 3.7.1 Velocity

The energy bands that were found in the preceding section can be used to represent the concept of conduction and valence bands. The second band in figure 3.23 can be thought of as the valence band, which is given approximately by the formula

$$E = E_2 + \frac{W_2}{2}\cos(Ka) \tag{3.126}$$

where $W_2$ is the *band width*. The next band up, not shown in figure 3.23, would have a form much like the lowest band, and could be written as

$$E = E_3 - \frac{W_3}{2}\cos(Ka). \tag{3.127}$$

Here, $W_3$ is the width of this band. The quantities $E_i$ are the centres of the band, about which the cosinusoidal variation occurs. In this sense, the second band can be thought of as the valence band and the third band as the conduction band with a gap at the centre of the Brillouin zone. The important question we wish to ask at this point is: What is the velocity of an electron in this third band?

The idea of the group velocity has already been developed in (1.47). This is the velocity at which the wave packet, representing the electron, moves through space. Here, the radian frequency $\omega$ is defined by the energy structure of (3.127). We can use (1.47) to evaluate this velocity as

$$\boldsymbol{v}_{\mathrm{g}} = \frac{\partial \omega}{\partial K} = \frac{1}{\hbar}\frac{\partial E}{\partial K} = \frac{W_3 a}{2\hbar}\sin(Ka). \tag{3.128}$$

This is an interesting result. The velocity is zero at $K = 0$ as expected. It then rises with increasing $K$ until reaching a peak at $Ka = \pi/2$, and then *decreases*. As the electron cycles through the entire conduction band (under the influence of an electric field for example), its velocity oscillates and the average value is zero. Indeed, if we apply an electric field to increase $K$, we have

$$K(t) = K(0) + \frac{eE}{\hbar}t \tag{3.129}$$

and the velocity varies as

$$\boldsymbol{v}_{\mathrm{g}} = \frac{W_3 a}{2\hbar}\sin\left(K(0)a + \frac{eEat}{\hbar}\right). \tag{3.130}$$

The velocity oscillates with a frequency

$$\omega_{\mathrm{B}} = \frac{eEa}{\hbar} \tag{3.131}$$

and this is known as the Bloch frequency. Of course, the presence of scattering processes in real materials keeps this from happening, and the electron cannot move very far away from $K(0) = 0$. However, if one could accelerate the electron over the maximum of (3.128), then the velocity would be decreasing and one conceivably could have a negative differential conductance. What is actually happening when the electron undergoes Bloch oscillations is that it is sampling the *entire* Brillouin zone. The Brillouin zone is a full Fourier transform of the lattice, so that a summation over all $k$-states in the band *localizes* the electron in real space at one lattice site. It is thought that the electron vibrates around this localized position.

The problem in real crystals is that the value of $K$ needed to reach $Ka = \pi/2$ is large. If the lattice constant is typically 0.25 nm, then $K \simeq 6 \times 10^{9}\ \mathrm{m}^{-1}$. If we try to accelerate the electron to this value, we require $K = eE\tau/\hbar$. Even if the scattering time is as large as 0.1 ps (at high energy), this requires an electric field of 400 kV cm$^{-1}$. This is an enormous field. On the other hand, if the periodicity could be enhanced to make $a$ much larger, then some interesting effects could occur. This is the goal of superlattices.

### 3.7.2 Superlattices

While the fields required in the previous section were unusually high, Esaki and Tsu (1970) put forward the idea of reducing the size of the Brillouin zone in $k$-space by creating *superlattices*. The idea was to grow, by molecular-beam epitaxy,

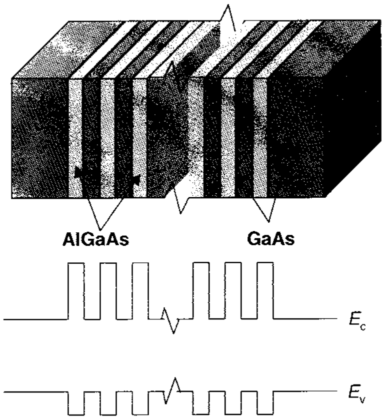


**Figure 3.24.** A superlattice of materials to form a new periodic structure. Here, the upper panel shows the layers of GaAs and AlGaAs. These give the band structure (versus position) indicated in the lower panel.

thin layers of GaAs and $\mathrm{Al_{0.3}Ga_{0.7}As}$ (an alloy of AlAs and GaAs). This would create a new periodicity in the crystal and create mini-bands in the new, reduced Brillouin zone. Such an idea is shown in figure 3.24. If the thicknesses of the thin layers were, for example, 10 nm, then the new value of $K$ at the minizone edge would be $1.6 \times 10^8$ m$^{-1}$. Now, the required field is only 10 kV cm$^{-1}$, a much more reasonable field. Several studies confirmed that negative differential conductance should occur within a simple model (Lebwohl and Tsu 1970, Reich and Ferry 1982). Indeed, it was also confirmed that the electron shows the oscillatory Bloch behaviour in the electron correlation function (Reich *et al* 1983). In fact, experimental studies did show negative differential conductance (Chang *et al* 1974, Tsu *et al* 1975), but this is now thought to not be due to Bloch oscillations. Rather, it is more likely that the second bound state in quantum well $i + 1$ lines up with the first bound state in well $i$, and that resonant tunnelling from one well to the next produces the negative differential conductivity.

The presence of the interwell tunnelling does not, by itself, rule out Bloch oscillations. This may be seen from figure 3.25, where the process is illustrated. In panel (*a*), it may be seen how the electron tunnels into a well at the second energy state, relaxes to the first energy state, and then tunnels out of the well, repeating the process in the next well. The question that is important relates to how the electron relaxes from the second state to the lower state. If the process is a direct transition (route '1' in panel (*b*)) and emits an optical photon at $\hbar\omega = eV = eEa = \hbar\omega_{\mathrm{B}}$, then we can say that Bloch oscillation and radiation have occurred. However, the more likely route is via the elastic scattering event labelled '2' in panel (*b*), where the transverse energy of the states is plotted as a function of the transverse momentum. This can be induced by impurities

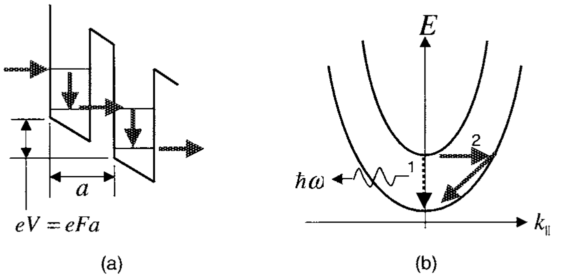


**Figure 3.25.** (*a*) The band alignment along the direction normal to the superlattice layers for a bias voltage $V$ that brings the second level in one well into alignment with the first level of the preceding well. This allows *sequential* tunnelling and relaxation (within the well) to produce a current flow. (*b*) The transverse energies of the two states are plotted as a function of the momentum in the planes of the superlattice layers. Two forms of relaxation within the well can occur. The process labelled '1' is a direct process in which a photon is emitted at the Bloch frequency. The inelastic process labelled '2' is a scattering process from the upper energy state to the lower one.

for example (to be discussed in chapter 7). Subsequent decay to the bottom of the lower subband occurs via additional scattering events. Negative differential conductance will still occur, as the current peaks when the levels are in alignment, and then decreases for further increases in the bias voltage which moves the levels out of alignment.

The problem, which is apparent by comparing figures 3.25 and 3.21 is that the miniband is being broken up by the field. Normally, the miniband should form around each of the bound states in the quantum well, providing a dispersion such as shown in figure 3.23. This represents a coherent, long-range wave function that extends through all the quantum wells. Under a sufficiently high electric field, however, this coherence is broken up and the discrete levels shown in figure 3.25 are re-established. If an AC electric field is added to the DC electric field in (3.130), this level structure can be probed. For example, an optical transition from the first *hole* bound state in well $i$ to the lowest *electron* bound state is denoted as $E_{11}$. On the other hand, if the transition is from the first hole band in well $i$ to the lowest electron band in well $i + 1$, then the transition energy is $E_{11} - \hbar\omega_{\mathrm{B}}$. In general, the transition from the first hole level in well $i$ to the lowest electron level in well $i \pm n$ is

$$\hbar\omega_{\mathrm{optical}} = E_{11} \pm n\hbar\omega_{\mathrm{B}}. \tag{3.132}$$

This set of levels forms what is called a *Stark ladder* of energy levels. This Stark ladder was measured in photoluminescence by a group at IBM, confirming the coupling of external signals to the Bloch oscillation capabilities within the superlattice by following the change in the levels with applied voltage (Agulló-Rueda *et al* 1989).

## 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 $\mathrm{SiO}_2$ found in MOS structures). The tunnelling through the capacitor oxide is an example of a very 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 an MOS 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 $\mathrm{SiO}_2$) 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 lowers the energy stored in the capacitor by the amount

$$\delta\mathcal{E} = \frac{e^2}{2C}. \tag{3.133}$$

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

$$\delta V = \frac{e}{C}. \tag{3.134}$$

What this means is that the tunnelling current cannot occur until a voltage equivalent to (3.134) 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 ($k_\mathrm{B}T/e$). On the other hand, suppose that the capacitance is defined with a lateral dimension of only 50 nm. Then, the area is $2.5 \times 10^{-15}$ $\mathrm{m}^2$, and our capacitor discussed above has a capacitance of $2.8 \times 10^{-17}$ F, and the required voltage of (3.134) is 5.7 mV. These capacitors are easily made, and the effects easily measured at low temperatures. In figure 3.26, we show measurements by Fulton and Dolan (1987) on such structures. The retardation of the tunnelling current until a voltage according

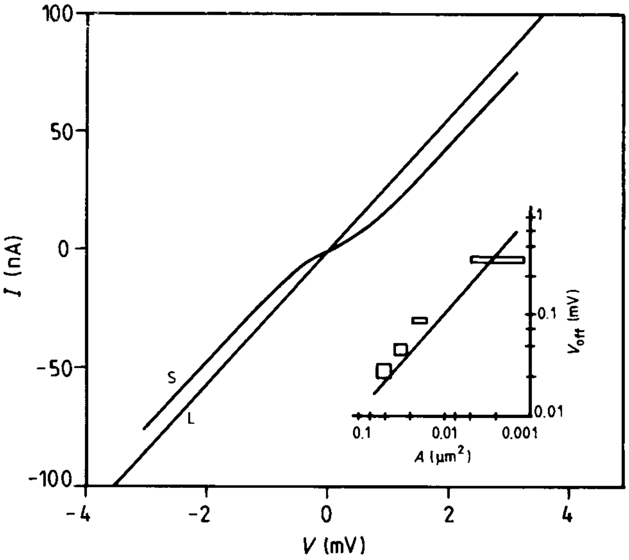


**Figure 3.26.** Single-electron tunnelling currents in small capacitors. The voltage offset is due to the Coulomb blockade. (After Fulton and Dolan (1987), by permission.)

to (3.134) 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 (3.134).

### 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

$$Q(t) = \int_0^t I\,\mathrm{d}t = It. \tag{3.135}$$

When the voltage reaches the value given by (3.134), 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

$$T = \frac{Q}{I} = \frac{e}{I} = \frac{2\pi}{\omega_{\mathrm{B}}} \tag{3.136}$$

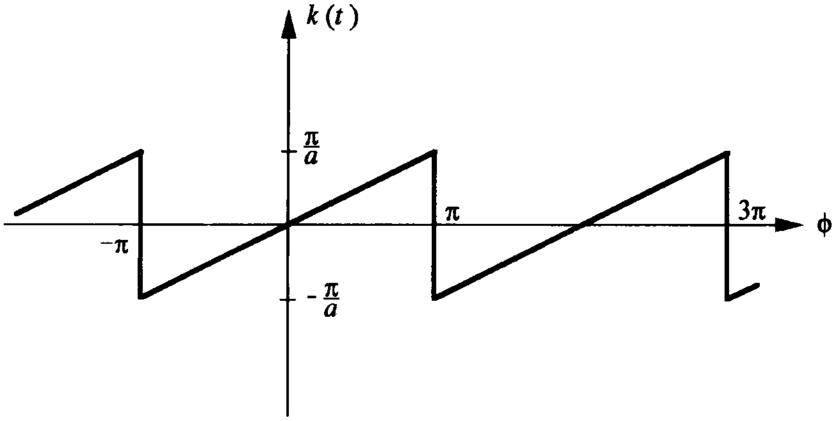


**Figure 3.27.** The variation in wave vector (or charge) in a periodic potential under the action of a constant electric field. The phase is $\phi = \omega_{\mathrm{B}}t = (eFat)/\hbar$.

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 (3.135) 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 3.23, for example. If we apply a constant electric field to the solid represented by this band structure, then the momentum responds according to (3.129). The meaning of (3.129) is that the magnitude of the wave vector $k$ increases linearly with electric field, and when it reaches the zone boundary at $\pi/a$ 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 3.27, where $\phi = (eFat)/\hbar = \omega_{\mathrm{B}}t$ is defined to be the phase, and $\omega_{\mathrm{B}}$ 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.

We have used the same symbol and description as Bloch oscillations for both the results of (3.131) in a superlattice and (3.136) in the single-electron tunnelling description. That this is correct can be illustrated by utilizing the dualism that exists in electromagnetics and electrical circuits. The dual of voltage is current. The dual of flux $(h/e)$ is charge $(e)$. Hence, we can use these dual relationships in (3.131) as

$$\omega_{\mathrm{B}} = \frac{eEa}{\hbar} = 2\pi\frac{V}{h/e} \to 2\pi\frac{I}{e} \tag{3.137}$$

which is just a rearranged version of (3.136). Thus, the two processes are in fact just duals of one another and the Bloch oscillation describes the same physics in each case—cycling of a particle through a periodic potential. In the case of single-electron tunnelling, we have not yet identified the existence of this periodic potential, but it must exist. Then the vertical axis in figure 3.27 becomes the voltage across the capacitor, which charges until a level is reached to provide the energy necessary for an electron to tunnel through the oxide.

### 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 (3.135), and figure 3.27, by the expression for the charge on the capacitor as

$$Q(t) = \frac{e}{2}\sin(\omega_{\mathrm{B}}t) \tag{3.138}$$

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

$$\phi = \omega_{\mathrm{B}}t. \tag{3.139}$$

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 3.23 for the band structure. We now take the *momentum* coordinate to correspond to

$$-\mathrm{i}e\frac{\partial}{\partial\phi} = Q. \tag{3.140}$$

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 (3.138) 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 (1.28), but for the 'correct' answer, we need to scale the phase by the factor $\hbar/e$. Then,

$$[Q, (\hbar/e)\phi] = -\mathrm{i}\hbar. \tag{3.141}$$

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

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = F = -\frac{\partial V}{\partial x}. \tag{3.142}$$

Thus, the potential energy is just

$$V(\phi) \sim -\frac{\hbar\omega_\mathrm{B}}{2}\int \cos(\phi)\,\mathrm{d}\phi \sim \frac{\hbar\omega_\mathrm{B}}{2}[1 - \sin(\phi)] \tag{3.143}$$

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 3.21. It is possible to expand the potential in figure 3.21 in a Fourier series, and it is obvious that (3.143) is just the lowest-order term in that expansion. It is also possible to expand the charge behaviour of figure 3.27 in a Fourier series and (3.138) is the lowest term in that expansion. Thus, the potential of (3.143) and the charge of (3.138) both correspond to the simplest periodic potential, which is just the lowest Fourier term of any actual potential. The constant term in (3.143) 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 3.21. For reference, the value of phase $\phi = \pi/2$ corresponds to $x = 0$ in figure 3.21. 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 (3.142), which leads to the adjusted potential

$$V(\phi) = \frac{\hbar\omega_\mathrm{B}}{2}[1 \pm \sin(\phi)] \tag{3.144}$$

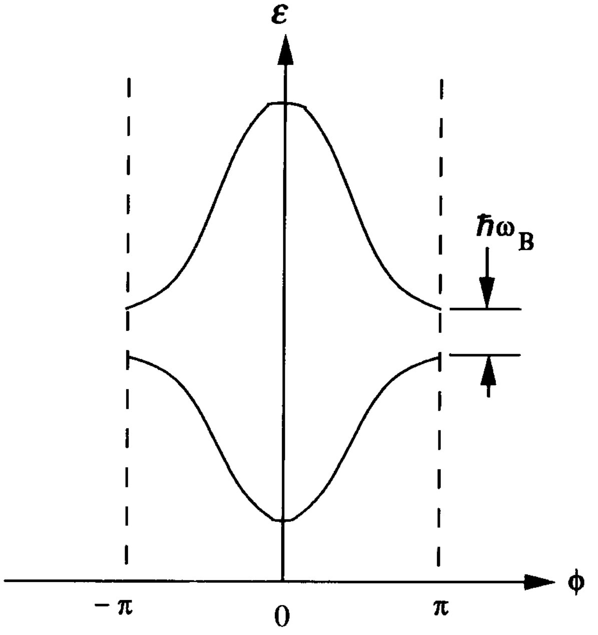


**Figure 3.28.** The energy band spectrum for the single-tunnelling capacitor.

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

$$H = \frac{Q^2}{2C} \pm \frac{\hbar\omega_\mathrm{B}}{2}\sin(\phi). \tag{3.145}$$

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 (3.138) to eliminate the phase, as

$$\mathcal{E} = \frac{Q^2}{2C} \pm \frac{\hbar\omega_\mathrm{B}}{2}\frac{2Q}{e}. \tag{3.146}$$

This energy is shown in figure 3.28. 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 (3.146). 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 3.27, gives rise to physical behaviour equivalent to that of a periodic potential. Thus, we can use the equivalent band structure of figure 3.28 to investigate other physical effects, all of which have their origin in the strange periodic behaviour of the charge on the capacitor.

### 3.8.3 The double-barrier quantum dot

To understand the Coulomb blockade somewhat more quantitatively, we consider the double-barrier structure shown in figure 3.14 as it may be created with extremely small lateral dimensions so that the capacitance of each barrier satisfies

$$\delta E = \frac{e^2}{2C} \gg k_{\mathrm{B}}T \tag{3.147}$$

as required for Coulomb blockade. The quantum well region, since it is laterally confined to a small dimension, may be termed a *quantum dot*. In general, each of the two capacitors (barriers) will have some leakage, but the equivalent circuit is as shown in figure 3.29($a$). The leakage is indicated as the shunt resistance through each capacitor. This leakage, or tunnelling resistance, represents the current flow when the electrons tunnel through the barrier. In this sense, these resistances are different from ordinary resistances. In the latter, charge and current flow is essentially continuous. Here, however, the current flow only occurs when the electrons are tunnelling, and this is for a very short time. The charge on the two capacitors is related to the two voltages, and is given by

$$Q_1 = C_1 V_1 \qquad Q_2 = C_2 V_2. \tag{3.148}$$

The net charge on the quantum dot island is given by

$$Q_{\mathrm{dot}} = Q_2 - Q_1 = -ne \tag{3.149}$$

where $n$ is the net number of excess electrons on the island. The sum of the junction voltages is the applied voltage $V_{\mathrm{a}}$, so that combining the two equations above, the voltage drops across the two junctions are just

$$V_1 = \frac{1}{C_{\mathrm{eq}}}(C_2 V_{\mathrm{a}} + ne) \qquad C_{\mathrm{eq}} = C_1 + C_2 \qquad V_2 = \frac{1}{C_{\mathrm{eq}}}(C_1 V_{\mathrm{a}} - ne). \tag{3.150}$$

The electrostatic energy can be written as

$$E_{\mathrm{s}} = \frac{Q_1^2}{2C_1} + \frac{Q_2^2}{2C_2} = \frac{1}{2C_{\mathrm{eq}}}(C_1 C_2 V_{\mathrm{a}}^2 + Q_{\mathrm{dot}}^2). \tag{3.151}$$

In addition, we must consider the work done by the voltage source in charging the two capacitors. This is an integral over the power delivered to the

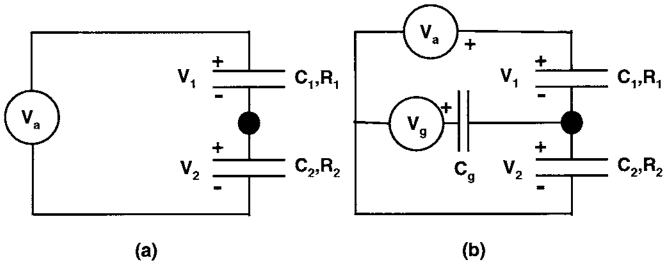


**Figure 3.29.** (*a*) Equivalent circuit for a quantum dot connected through two tunnelling capacitors. The indicated resistance is the tunnelling resistance, as discussed in the text. (*b*) The circuit when a gate bias is applied to directly vary the potential at the dot.

tunnel junctions during the charging process, or

$$E_{\mathrm{a}} = \int \mathrm{d}t\, V_{\mathrm{a}} I(t) = V_{\mathrm{a}} \Delta Q \tag{3.152}$$

where $\Delta Q$ is the total charge transferred from the voltage source to the capacitors. This includes the integer number of electrons on the quantum dot as well as the continuous polarization charge that builds up in response to the voltages on the capacitors (resulting in an electric field in the island). A change in the charge on the island due to one electron tunnelling through $C_2$ changes the charge on the island to $Q' = Q + e$, and $n' = n - 1$. From (3.150), the voltage change on $C_1$ results in $V_1' = V_1 - e/C_{\mathrm{eq}}$. Therefore a polarization charge flows in from the voltage source $\Delta Q = -eC_1/C_{\mathrm{eq}}$ to compensate. The total work done to pass $n_2$ charges through $C_2$ is then

$$E_{\mathrm{a}}(n_2) = -n_2 e V_{\mathrm{a}} C_1 / C_{\mathrm{eq}}. \tag{3.153}$$

A similar argument is used to generate the total work done to pass $n_1$ charges through $C_1$, which is

$$E_{\mathrm{a}}(n_1) = -n_1 e V_{\mathrm{a}} C_2 / C_{\mathrm{eq}}. \tag{3.154}$$

Combining these two results with (3.151), the total energy of the complete circuit, including that provided by the voltage source, is given by

$$E(n_1, n_2) = E_{\mathrm{s}} - E_{\mathrm{a}} = \frac{1}{2C_{\mathrm{eq}}}(C_1 C_2 V_{\mathrm{a}}^2 + Q_{\mathrm{dot}}^2) + \frac{eV_{\mathrm{a}}}{C_{\mathrm{eq}}}(C_1 n_2 + C_2 n_1). \tag{3.155}$$

With this description of the total energy in terms of the charge on each of the capacitors (given by the factors $n_i$), we can now look at the *change* in the energy when a particle tunnels through either capacitor. At low temperature, the tunnelling transition must take the system from a state of higher energy to a state of lower energy. The change in energy for a particle tunnelling through $C_2$ is

$$\begin{aligned}\Delta E_2^{\pm} &= E(n_1, n_2) - E(n_1, n_2 \pm 1) \\ &= \left[\frac{Q_{\mathrm{dot}}^2}{2C_{\mathrm{eq}}} - \frac{(Q_{\mathrm{dot}} \mp e)^2}{2C_{\mathrm{eq}}}\right] \mp \frac{eV_{\mathrm{a}}C_1}{C_{\mathrm{eq}}} \\ &= \frac{e}{C_{\mathrm{eq}}}\left[-\frac{e}{2} \pm (Q_{\mathrm{dot}} - V_{\mathrm{a}}C_1)\right].\end{aligned} \tag{3.156}$$

The value of the charge on the dot $Q_{\mathrm{dot}}$ is *prior* to the tunnelling process. Similarly, the change in energy for a particle tunnelling through $C_1$ is given by

$$\begin{aligned}\Delta E_1^{\pm} &= E(n_1, n_2) - E(n_1 \pm 1, n_2) \\ &= \left[\frac{Q_{\mathrm{dot}}^2}{2C_{\mathrm{eq}}} - \frac{(Q_{\mathrm{dot}} \pm e)^2}{2C_{\mathrm{eq}}}\right] \mp \frac{eV_{\mathrm{a}}C_1}{C_{\mathrm{eq}}} \\ &= \frac{e}{C_{\mathrm{eq}}}\left[-\frac{e}{2} \mp (Q_{\mathrm{dot}} + V_{\mathrm{a}}C_1)\right].\end{aligned} \tag{3.157}$$

According to this discussion, only those transitions are allowed for which $\Delta E_i > 0$; e.g., the initial state is at a higher energy than the final state.

If we consider a system in which the dot is initially uncharged, $Q_{\mathrm{dot}} = 0$, then (3.156) and (3.157) reduce to

$$\Delta E_{1,2}^{\pm} = -\frac{e^2}{2C_{\mathrm{eq}}} \pm \frac{eV_{\mathrm{a}}C_{2,1}}{C_{\mathrm{eq}}} > 0. \tag{3.158}$$

Initially, at low voltage, the leading term on the right-hand side makes $\Delta E < 0$. Hence, no tunnelling can occur until the voltage reaches a threshold that depends upon the lesser of the two capacitors. For the case in which $C_1 = C_2 = C$, the requirement becomes $|V_{\mathrm{a}}| > e/C_{\mathrm{eq}}$. Tunnelling is prohibited and no current flows below this threshold voltage. This region of *Coulomb blockade* is a direct result of the additional Coulomb energy which is required for an electron to tunnel through one of the capacitors.

Now, consider the situation when we make an additional contact, through another capacitor, to the quantum dot as shown in figure 3.29(*b*). A new voltage source, $V_{\mathrm{g}}$, is coupled to the quantum dot through an *ideal* capacitor $C_{\mathrm{g}}$ (no tunnelling is possible through this capacitor). This additional voltage source modifies the charge balance on the quantum dot, so that

$$Q_{\mathrm{g}} = C_{\mathrm{g}}(V_{\mathrm{g}} - V_2). \tag{3.159}$$

The charge on the quantum dot now becomes

$$Q_{\mathrm{dot}} = Q_2 - Q_1 - Q_{\mathrm{g}} = -ne. \tag{3.160}$$

The voltages across the two junctions is also modified, and we can write

$$\begin{aligned}V_1 &= \frac{1}{C_{\mathrm{T}}}((C_{\mathrm{g}} + C_2)V_{\mathrm{a}} - C_{\mathrm{g}}V_{\mathrm{g}} + ne) \qquad C_{\mathrm{T}} = C_{\mathrm{g}} + C_{\mathrm{eq}} \\ V_2 &= \frac{1}{C_{\mathrm{T}}}(C_1V_{\mathrm{a}} + C_{\mathrm{g}}V_{\mathrm{g}} - ne).\end{aligned} \tag{3.161}$$

The electrostatic energy (3.151) now must be modified to include the gate capacitance as

$$E_{\mathrm{s}} = \frac{1}{2C_{\mathrm{T}}}(C_1 C_2 V_{\mathrm{a}}^2 + C_{\mathrm{g}} C_2 V_{\mathrm{g}}^2 + C_{\mathrm{g}} C_1 (V_{\mathrm{a}} - V_{\mathrm{g}})^2 + Q_{\mathrm{dot}}^2). \tag{3.162}$$

The work done by the voltage sources during the tunnelling now includes the work done by the gate voltage and the additional charge flowing onto the gate capacitor. Equations (3.153) and (3.154) now become

$$E_{\mathrm{a}}(n_2) = -n_2 \lfloor eV_{\mathrm{a}} C_1/C_{\mathrm{T}} + eV_{\mathrm{g}} C_{\mathrm{g}}/C_{\mathrm{T}} \rfloor \tag{3.163}$$

and

$$E_{\mathrm{a}}(n_1) = -n_1 \lfloor eV_{\mathrm{a}} C_2/C_{\mathrm{T}} + e(V_{\mathrm{a}} - V_{\mathrm{g}}) C_{\mathrm{g}}/C_{\mathrm{T}} \rfloor. \tag{3.164}$$

The total energy for the charge state characterized by $n_1$ and $n_2$ is now given by

$$\begin{aligned} E(n_1, n_2) = \frac{1}{2C_{\mathrm{T}}}&[(C_1 C_2 V_{\mathrm{a}}^2 + Q_{\mathrm{dot}}^2) + C_2 C_{\mathrm{g}} V_{\mathrm{g}}^2 + C_{\mathrm{g}} C_1 (V_{\mathrm{a}} - V_{\mathrm{g}})^2 \\ &+ 2eV_{\mathrm{a}}(C_1 n_2 + (C_2 + C_{\mathrm{g}}) n_1) - 2n_1 C_{\mathrm{g}} V_{\mathrm{g}}]. \end{aligned} \tag{3.165}$$

For the tunnelling of an electron across $C_1$, the energy change is now given by

$$\begin{aligned} \Delta E_1^{\pm} &= E(n_1, n_2) - E(n_1 \pm 1, n_2) \\ &= \left[\frac{Q_{\mathrm{dot}}^2}{2C_{\mathrm{T}}} - \frac{(Q_{\mathrm{dot}} \pm e)^2}{2C_{\mathrm{T}}}\right] \mp \frac{e[V_{\mathrm{a}}(C_2 + C_{\mathrm{g}}) - C_{\mathrm{g}} V_{\mathrm{g}}]}{C_{\mathrm{T}}} \\ &= \frac{e}{C_{\mathrm{T}}}\left[-\frac{e}{2} \mp (Q_{\mathrm{dot}} + V_{\mathrm{a}}(C_1 + C_{\mathrm{g}}) - C_{\mathrm{g}} V_{\mathrm{g}})\right]. \end{aligned} \tag{3.166}$$

Similarly, the change in energy for an electron tunnelling through $C_2$ is

$$\begin{aligned} \Delta E_2^{\pm} &= E(n_1, n_2) - E(n_1, n_2 \pm 1) \\ &= \left[\frac{Q_{\mathrm{dot}}^2}{2C_{\mathrm{T}}} - \frac{(Q_{\mathrm{dot}} \mp e)^2}{2C_{\mathrm{T}}}\right] \mp \frac{e(V_{\mathrm{a}} C_1 + V_{\mathrm{g}} C_{\mathrm{g}})}{C_{\mathrm{T}}} \\ &= \frac{e}{C_{\mathrm{T}}}\left[-\frac{e}{2} \pm (Q_{\mathrm{dot}} - V_{\mathrm{a}} C_1 - C_{\mathrm{g}} V_{\mathrm{g}})\right]. \end{aligned} \tag{3.167}$$

When we compare these results with those of (3.156) and (3.157), it is apparent that the gate voltage allows us to change the effective charge on the quantum dot, and therefore to shift the region of Coulomb blockade with $V_{\mathrm{g}}$. As before, the condition for tunnelling at low temperature is that the change in energy must be negative and the tunnelling must take the system to a lower energy state. We now have two conditions that exist for forward and backward tunnelling as

$$\begin{aligned} -\frac{e}{2} \mp [ne + (C_{\mathrm{g}} + C_2)V_{\mathrm{a}} - C_{\mathrm{g}} V_{\mathrm{g}}] > 0 \\ -\frac{e}{2} \pm [ne - C_1 V_{\mathrm{a}} - C_{\mathrm{g}} V_{\mathrm{g}}] > 0. \end{aligned} \tag{3.168}$$

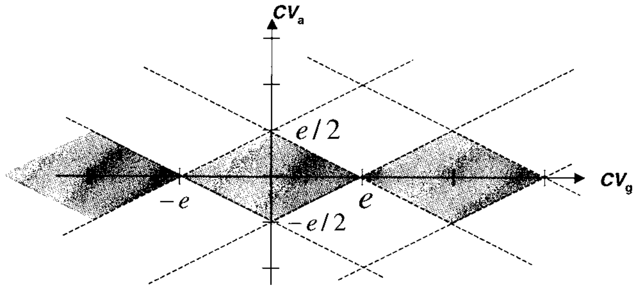


**Figure 3.30.** A stability diagram for the single-electron quantum dot. The parameters are discussed in the text, and it is assumed that the temperature is $T = 0$. The shaded regions are where the dot electron number is stable and Coulomb blockade exists.

The four equations (3.168) may be used to generate a stability plot in the $(V_a, V_g)$ plane, which shows stable regions corresponding to each value of $n$, and for which no tunnelling can occur. This diagram is shown in figure 3.30 for the case in which $C_g = C_2 = C$ and $C_1 = 2C$ $(C_T = 4C)$. The lines represent the boundaries given by (3.168). The trapezoidal shaded areas correspond to regions where no solution satisfies (3.168) and hence where Coulomb blockade exists. Each of the regions corresponds to a different number of electrons on the quantum dot, which is stable in the sense that this charge state does not change easily. The gate voltage allows us to 'tune' between different stable regimes, essentially adding or subtracting one electron at a time to the dot region. For a given gate bias, the range of $V_a$ over which Coulomb blockade occurs is given by the vertical extent of the shaded region. The width of this blockaded region approaches zero as the gate charge approaches half-integer values of a single electron charge, and here tunnelling occurs easily. It is at these gate voltages that tunnelling current peaks can be observed as the gate bias is varied.

In figure 3.30, the superlattice behaviour of the charge in terms of either the gate voltage or the charging voltage $V_a$ is clear. This gives a distance between the current peaks for tunnelling through the structure of $\Delta V_g = e/C = e/C_g$. Between these peaks the number of electrons on the quantum dot remains constant. We will return to the quantum dot in chapter 8, where we begin to worry about the quantum levels within the dot itself. To this point, we have ignored the fact that the quantum dot may be so small that the energy levels within the dot are quantized. Yet, this can occur, and the behaviour of this section is modified. As remarked, we return to this in chapter 8, where we deal with spectroscopy of just these quantum levels.

## References

Aguÿló-Rueda F, Mendez E E and Hong J M 1989 *Phys. Rev.* B **40** 1357

Bohm D 1951 *Quantum Theory* (Englewood Cliffs, NJ: Prentice-Hall)

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

Chang L L, Esaki L and Tsu R 1974 *Appl. Phys. Lett.* **24** 593

Esaki L and Tsu R 1970 *IBM J. Res. Dev.* **14** 61

Ferry D K and Goodnick S M 1997 *Transport in Nanostructures* (Cambridge: Cambridge University Press)

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

Grabert H and Devoret M H (ed) 1992 *Single Charge Tunneling, Coulomb Blockade Phenomena in Nanostructures (ASI Series B 294)* (New York: Plenum)

Kitabayashi H, Waho T and Yamamoto M 1997 *Appl. Phys. Lett.* **71** 512

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

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

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

Lebwohl P A and Tsu R 1970 *J. Appl. Phys.* **41** 2664

Price P 1999 *Microelectron. J.* **30** 925

Reich R K and Ferry D K 1982 *Phys. Lett.* A **91** 31

Reich R K, Grondin R O and Ferry D K 1983 *Phys. Rev.* B **27** 3483

Rommel S L *et al* 1998 *Appl. Phys. Lett.* **73** 2191

Söderström J R, Chow D H and McGill T C 1989 *Appl. Phys. Lett.* **55** 1094

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

Tsu R, Chang L L, Sai-Halasz G A and Esaki L 1975 *Phys. Rev. Lett.* **34** 1509

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

Yacoby A, Heiblum M, Umansky V and Shtrikman H 1994 *Phys. Rev. Lett.* **73** 3149

Yacoby A, Heiblum M, Mahalu D and Shtrikman H 1995 *Phys. Rev. Lett.* **74** 4047

Yu E T, Collins D A, Ting D Z-Y, Chow D H and McGill T C 1990 *Appl. Phys. Lett.* **57** 2675

## Problems

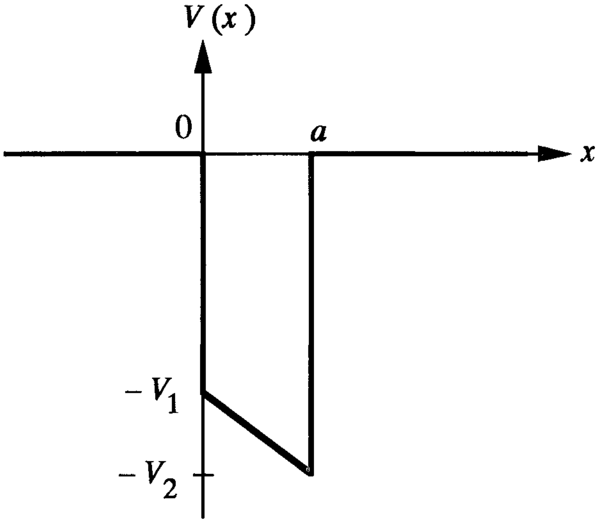
1. For a potential barrier with $V(x) = 0$ for $x > |a/2|$, and $V(x) = 0.3$ eV for $x < |a/2|$, plot the tunnelling probability for $\mathcal{E}$ in the range 0–0.5 eV. Take the value $a = 5$ nm and use the effective mass of GaAs, $m^* = 6.0 \times 10^{-32}$ kg.

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 $T_{\min}/T_{\max}$).

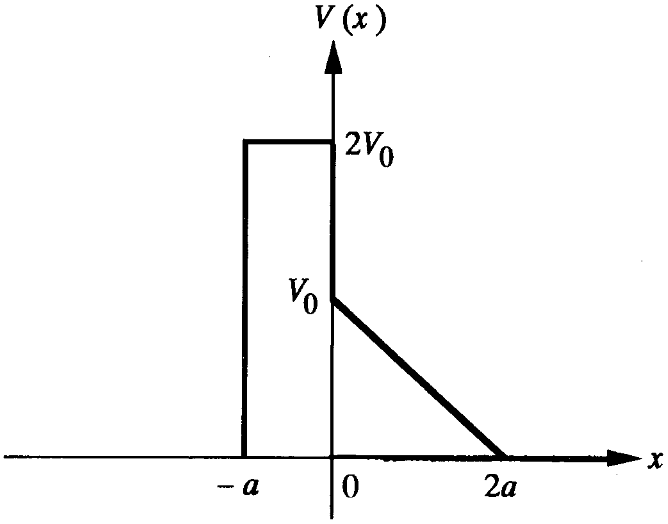
5. Let us consider a trapezoidal potential well, such as that shown in the figure below. Using the WKB method, find the bound states within the well. If $V_1 = 0.3$ eV, $V_2 = 0.4$ eV, and $a = 5$ nm, what are the bound-state energies?



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

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

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 (3.124), 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.

11. Consider a single rectangular barrier in which the barrier height is 2.5 eV. A free electron of energy 0.5 eV is incident from the left. If the barrier is 0.2 nm thick, what are the tunnelling and reflection coefficients?

12. In an infinite potential of width 15 nm, an electron is initialized with the wave function $\psi(x) = x(1 - x/a)\mathrm{e}^{-x}$, where $a$ is the width of the well. Develop a time-dependent solution of the Schrödinger equation to show how this wave function evolves in time.

13. Solve the two-dimensional Schrödinger equation for a two-dimensional infinite potential well with $a_x = 5\ \mathrm{nm}$ and $a_y = 7\ \mathrm{nm}$. Determine the seven lowest energy levels.

14. Consider a potential barrier which is described by $V(x) = -Ax^2$. Using the WKB method, compute the tunnelling coefficient as a function of energy for $E < 0$.

# Chapter 4

# The harmonic oscillator

One of the most commonly used quantum mechanical systems is that of the simple *harmonic oscillator*. Classically, the motion of the system exhibits harmonic, or sinusoidal oscillations. Typical examples are a mass on a spring and a simple linear pendulum, the latter of which will be explored here. Quantum mechanically, the motion described by the Schrödinger equation is more complex. Although quite a simple system in principle, the harmonic oscillator assumes almost overwhelming importance due to the fact that almost any interaction potential may be expanded in a series with the low-order terms cast into a form that resembles this system. The sinusoidal motion of classical mechanics is the simplest system that produces oscillatory behaviour, and therefore is found in almost an infinity of physical systems. Thus, the properties of the harmonic oscillator become important for their use in describing quite diverse physical systems. In this chapter, we will develop the general mathematical solution for the wave functions of the harmonic oscillator, then develop a simple operator algebra that allows us to obtain these wave functions and properties in a much more usable and simple manner. We then turn to two classic examples of systems in which we use the results for the harmonic oscillator to explain the properties: the simple $LC$-circuit and vibrations of atoms in a crystalline lattice.

The simplest example of sinusoidal behaviour in classical physics is that of a mass on a linear spring, in which the mass responds to a force arising from the extension or compression of the spring (figure 4.1). The differential equation describing this is just

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = F = -Cx \tag{4.1}$$

where $m$ is the mass and $C$ is the spring constant. Instead of solving this equation classically, we instead write the total energy, using the square of the momentum $p = m\,\mathrm{d}x/\mathrm{d}t$ and the potential energy $V = Cx^2/2$ that arises from integrating the right-hand side, as

$$\mathcal{E} = \frac{p^2}{2m} + \frac{m\omega^2x^2}{2} \tag{4.2}$$

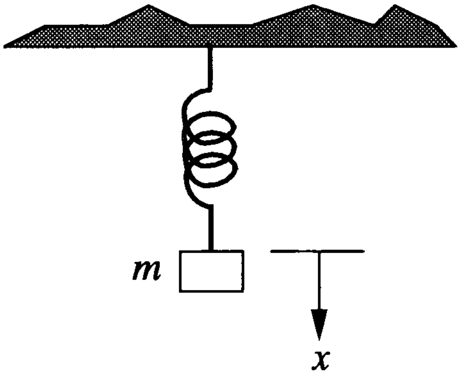


**Figure 4.1.** A simple mass on a spring.

where we have introduced the oscillator frequency $\omega$ through $C = m\omega^2$. This is the energy function that is used in the Schrödinger equation to find the quantum mechanical motion of the simple harmonic oscillator. We will explore the transition from classical dynamics to quantum dynamics in the next chapter.

The simple pendulum is one of the first problems one attacks in introductory physics. Quite simply, a mass $M$ is suspended by a rigid rod of length $l$ from a fixed point, about which it can rotate. The angle that the rod makes with the vertical is given by $\theta$. As the rod is pushed away from the vertical, gravity provides a restoring force of value $Mg$, which is directed directly downward (figure 4.2), and which leads to a force that tends to reduce the angular deflection of the rod, given by $-Mg\sin\theta$. We may then write the differential equation for the angular deflection $\theta$ as

$$M\frac{\mathrm{d}}{\mathrm{d}t}\left(l\frac{\mathrm{d}\theta}{\mathrm{d}t}\right) = -Mg\sin\theta. \tag{4.3}$$

The factor in the parentheses on the left-hand side is just the angular velocity of the pendulum, so the left-hand side is the time derivative of the angular momentum, and the right-hand side is the restoring force on the angle. We linearize this equation by expanding the sine function for small angles, so the resulting equation is just

$$M\frac{\mathrm{d}}{\mathrm{d}t}\left(l\frac{\mathrm{d}\theta}{\mathrm{d}t}\right) + Mg\theta = 0. \tag{4.4}$$

We can easily develop the Hamiltonian for this, by remarking that this latter quantity is the sum of the kinetic and potential energies, just as for the mass on a spring discussed in the previous paragraph. The kinetic energy is obtained from the momentum as $p^2/2m$, and this is evaluated using the angular momentum in (4.3) and the mass of the pendulum (neglecting the mass of the rod itself).

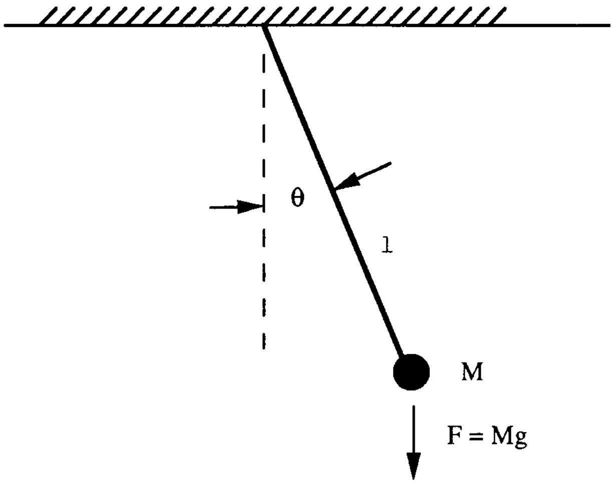


**Figure 4.2.** The simple pendulum.

The potential energy is obtained by noting that the right-hand side of (4.3) is the force, which is obtained from the spatial derivative of the potential energy. In the linearized version of (4.3), this leads to

$$F = -Mg\theta = -\frac{\partial V}{l\partial\theta} \tag{4.5}$$

and

$$V(\theta) = \tfrac{1}{2}Mgl\theta^2. \tag{4.6}$$

This leads to the Hamiltonian taking the form

$$H = \frac{1}{2M}\left(Ml\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 + \frac{1}{2}Mgl\theta^2. \tag{4.7}$$

To make (4.7) into a general Hamiltonian such as (4.2), which can be applied to many systems, we will make a few changes of variables. First, we take the 'position' angle into a position notation by making the change $l\theta \to x$. Next, we replace the gravity by a general frequency through making the change $g/l \to \omega^2$. Finally, we introduce quantization of the system by defining the momentum as a differential operator through

$$Ml\frac{\mathrm{d}\theta}{\mathrm{d}t} \to -\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}x}. \tag{4.8}$$

This now leads us to the Hamiltonian of (4.2) and subsequently to the Schrödinger equation, with the additional substitution $M \to m$ made in order to be consistent with previous chapters. The Schrödinger equation then becomes

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2}+\frac{m\omega^2}{2}x^2\Psi=\mathcal{E}\Psi(x). \tag{4.9}$$

It is this formulation of the Schrödinger equation, as well as the linearized-angle equation (4.7), that is found to occur in a great many applications. The recognition of this commonality is very important, since if we can recognize a known set of solutions to a new problem, it becomes quite easy to interpret the expected results for that problem. The harmonic oscillator is one of the most frequently studied problems for just this reason.

Another version of the harmonic oscillator that is more familiar to electrical engineers is the resonant $LC$-circuit. We consider the circuit of figure 4.3. The familiar equations for current and voltage in the inductor and capacitor (directions are indicated in the figure, so this will change some signs) are

$$I=-C\frac{\mathrm{d}V}{\mathrm{d}t}\qquad V=L\frac{\mathrm{d}I}{\mathrm{d}t}. \tag{4.10}$$

On the other hand, the energy stored in the circuit is given by

$$E=\frac{1}{2}LI^2+\frac{1}{2}CV^2=\frac{1}{2L}\phi^2+\frac{L}{2}\omega^2Q^2 \tag{4.11}$$

where we have introduced the *flux linkages* $\phi=LI$, the charge $Q=CV$, and the resonant frequency $\omega=1/\sqrt{LC}$. If we now relate $L$ to the mass $m$, $Q$ to the position $x$, and $\phi$ to the momentum $p$, (4.11) is exactly (4.2). In this case, these relationships are *analogues* to one another. Indeed, we may now rewrite (4.10) as

$$\frac{\mathrm{d}V}{\mathrm{d}t}=\frac{\mathrm{d}}{\mathrm{d}t}\left(L\frac{\mathrm{d}I}{\mathrm{d}t}\right)=\frac{\mathrm{d}^2\phi}{\mathrm{d}t^2}=-\frac{I}{C} \tag{4.12}$$

or

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}t^2}+\omega^2\phi=0. \tag{4.13}$$

This last form should be compared to (4.4), where we relate $\omega^2=g/l$. A similar form can be developed for $Q$ as well. Since we connected $Q$ with position, we will be able to develop a form of the Schrödinger equation for this variable, just as any other quantum variable. We return to this in section 4.5.

## 4.1 Hermite polynomials

The traditional approach to solving the Schrödinger equation for the harmonic oscillator is through the use of a set of orthonormal polynomials known, for the particular equation that we shall obtain, as the Hermite polynomials. Equation (4.9) is a particular case of the general Sturm–Liouville problem, whose

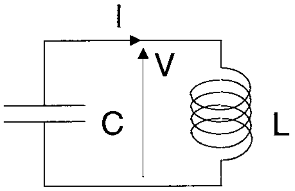


**Figure 4.3.** A resonant circuit composed of an inductor $L$ and a capacitor $C$. The voltage $V$ and current $I$ are discussed in the text.

particular solutions are known to be the Hermite polynomials. Here, the approach is simply to walk through the traditional solution method. In this section and the next, we will show how these polynomials provide the proper solutions for the wave functions, and give the basic quantization of the energy levels. Then, in section 4.4, we will pursue a more elegant, but mathematically simpler, method of solution in terms of a pair of operators and their algebra defined by the commutator of position and momentum.

To begin with, we want to make again a change of variables in (4.9), to normalize the position and bury most of the various constants in this variable change. To this end, we introduce a reduced position variable:

$$\xi = \sqrt{\frac{m\omega}{\hbar}}x. \tag{4.14}$$

With this variable change, and a rearrangement of the terms, we can rewrite (4.9) as

$$\frac{\mathrm{d}^2\Psi(\xi)}{\mathrm{d}\xi^2} + \left[\frac{2\mathcal{E}}{\hbar\omega} - \xi^2\right]\Psi(\xi) = 0. \tag{4.15}$$

At this point, the substitution (4.14) has merely simplified the equation by removing the plethora of constants. The usual approach, which is, of course, based upon experience, is for the practitioner (the professor in the classroom) to arrive magically at another change of variables, which miraculously works out to give the desired solutions. The 'magic' is obtained by consideration of the WKB approach treated in the last chapter. We know that the wave functions must ultimately decay for large values of $|x|$, since any given energy level must eventually be less than the potential height as $|x|$ is increased. We know from the WKB principles that the wave function must decay as $\exp[-\int^{\xi}\sqrt{V}\,\mathrm{d}\xi]$, which then leads to $\mathrm{e}^{-\xi^2/2}$ behaviour. Thus, we introduce the new function

$$\Psi(\xi) \rightarrow \mathrm{e}^{-\xi^2/2}\psi(\xi). \tag{4.16}$$

The introduction of (4.16) into (4.15) leads to the new differential equation

$$\frac{\mathrm{d}^2\psi(\xi)}{\mathrm{d}\xi^2} - 2\xi\frac{\mathrm{d}\psi(\xi)}{\mathrm{d}\xi} + \left[\frac{2\mathcal{E}}{\hbar\omega} - 1\right]\psi(\xi) = 0. \tag{4.17}$$

At this point, no more magic is possible. We have forced the overall wave function to have the decay properties expected from WKB considerations. Now, the resulting equation, (4.17), does not have a recognizable form (for the novice), and the hard work must begin.

It is possible that the solutions to (4.17) are in terms of some special functions discovered by a mathematician long ago, but unless we are experienced this is not an obvious result. Therefore, we take a tried and true brute-force approach. We shall assume a polynomial solution, with terms of the form $\xi^n$. In fact, we do know some properties of the solutions. First, we have a second-order differential equation in position, and two boundary conditions (vanishing of the wave function) for large positive and negative $\xi$. Thus, we expect to have two independent solutions. We also know that the potential is an even function, so we expect the wave functions to be of either even or odd parity; for example $\psi(\xi) = \pm\psi(-\xi)$. In looking for two independent solutions, we will keep this parity in mind, since it is quite likely that one solution will have one parity, while the second will have the opposite parity. We also know that the reduced wave functions $\psi(\xi)$ have these same properties, plus they must diverge less rapidly than the exponential for the overall wave functions to vanish for large $\xi$. It turns out that this is an enormous collection of information, which will prove quite sufficient for solving the problem completely!

Now, with the above considerations fully in mind, we make the substitution into (4.17) of a power series in the form

$$\psi(\xi) = \sum_n a_n \xi^n. \tag{4.18}$$

This leads to the equation

$$\sum_n \left\{ n(n-1)a_n\xi^{n-2} - 2na_n\xi^n + \left[\frac{2\mathcal{E}}{\hbar\omega} - 1\right] a_n\xi^n \right\} = 0. \tag{4.19}$$

We make the change $n \to n+2$ in the first term of the series, so that each term is of the same order in $\xi$. For the series to vanish, as required by the differential equation, each coefficient of each power of $\xi$ must also vanish. This leads to an iterative relation

$$a_{n+2} = -\frac{2\mathcal{E}/\hbar w - 1 - 2n}{(n+2)(n+1)} a_n. \tag{4.20}$$

This result is excellent for our preconceived expectations. First, only even terms or odd terms (in the exponents of the power series (4.18)) are coupled. Thus, the even series and the odd series are the two independent solutions. Secondly, $a_0$ and $a_1$ are the two constants needed for the second-order differential equation and each of these two governs only one of the two independent solutions. Thus, the needs for parity and two independent solutions are both satisfied. Finally, we can ensure that the solutions do not diverge for large values of the variable $\xi$ only if we terminate the series at some finite value of $n$. (If we do not terminate the series, (4.20) tells us that the reduced wave function will diverge as $\mathrm{e}^{\xi^2}$ for large $|\xi|$, which diverges faster than the exponential factor introduced in (4.16). Thus, we must terminate the series at some finite order.) Which value should we choose? It doesn't matter. We recall that in a Fourier series expansion, we get an entire family of solutions with different values of $n$. The same is true here. Each choice of $n$ gives one member of the entire family of solutions. Thus, in order to terminate the solutions for some value of $n$, we set the numerator of the right-hand side of (4.20) to zero, which leads to

$$\mathcal{E}_n = (n + \tfrac{1}{2})\hbar\omega \qquad n = 0, 1, 2, 3, \ldots. \tag{4.21}$$

*This equation is perhaps the most important one to come from considering the harmonic oscillator*. It clearly tells us that the modes of vibration of the oscillator are quantized. Each allowed mode has a particular amplitude given by the appropriate wave function for that value of $n$, and the energy stored in that mode is given by (4.21). If we want to give the oscillator more energy, it must be by at least one quantum of value $\hbar\omega$ and this lifts the oscillator into a higher-lying state with a wave function of higher index $n$. Similarly, if the oscillator is to lose energy, it must do so in units of $\hbar\omega$, and consequently drops into states with wave functions of lower index $n$. It is found that the problems that can be put into the harmonic oscillator form are those that involve particles that have integer spin, and are thus termed *bosons* (for the appearance of Bose–Einstein statistics). Examples are *photons* (light particles of integer spin) and *phonons* (lattice vibrations of zero spin). Each quantum unit of excitation for these particles comes in units of $\hbar\omega$, and these are said to be single photons, or phonons, or some other type of boson.

The actual solutions to (4.17) are polynomials known as the Hermite polynomials. The two constants $a_0$ and $a_1$ are taken to have an actual value set by normalization considerations. For this purpose, we modify (4.20) to

$$a_i = -\frac{2n - 2(i-2)}{i(i-1)} a_{i-2} \qquad i \le n. \tag{4.22}$$

The initial sets of these polynomials can be found from (4.21) and (4.22) as (we assign an overall sign that makes the highest-power term positive)

$$\begin{aligned} H_0(\xi) &= 1 \\ H_2(\xi) &= 4\xi^2 - 2 \\ H_4(\xi) &= 16\xi^4 - 48\xi^2 + 12 \end{aligned} \tag{4.23}$$

for the even-symmetry functions, and

$$\begin{aligned} H_1(\xi) &= 2\xi \\ H_3(\xi) &= 8\xi^3 - 12\xi \\ H_5(\xi) &= 32\xi^5 - 160\xi^3 + 120\xi \end{aligned} \tag{4.24}$$

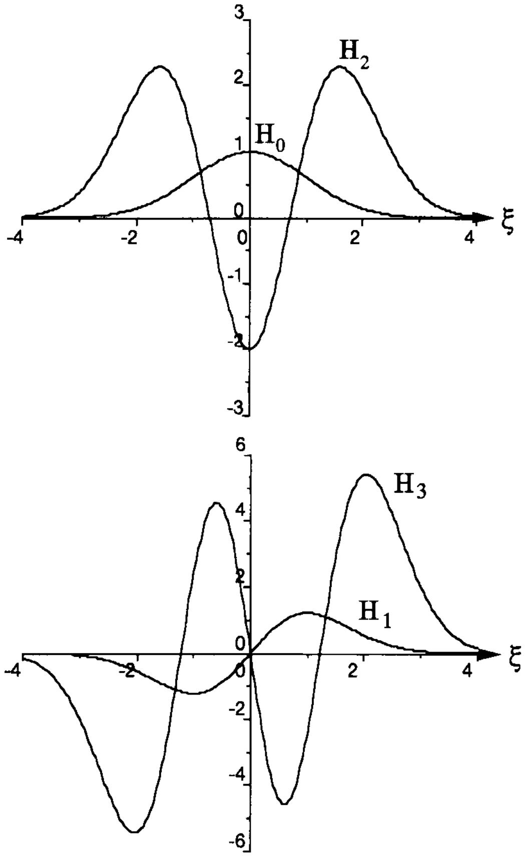


**Figure 4.4.** The lowest four weighted Hermite polynomials.

for the lowest odd-symmetry functions. The coefficients of the leading terms are set by a condition, discussed below, relating to the *generating function*, but do not provide normalization. The actual normalization will be described below with a methodology that does not require us to integrate each and every function. We plot four of these in figure 4.4.

## 4.2 The generating function

At this point, we diverge from the main line of discussion in order to develop some general properties of the Hermite polynomials, properties such as the normalization and orthonormality properties. This is most efficiently done by using what is known as a generating function, which we define in the following manner:

$$F(s,\xi) = \sum_{n=0}^{\infty} \frac{s^n}{n!} H_n(\xi). \qquad (4.25)$$

The generating function is determined as a power series in $s$, with the coefficients given by the Hermite polynomials of the appropriate order. Our problem now is to determine what the proper form of $F(s,\xi)$ should be.

To begin the task of identifying the function $F(s,\xi)$, we shall take the derivative with respect to the argument of the Hermite polynomials. This leads to

$$\frac{\partial F}{\partial \xi} = \sum_{n=0}^{\infty} \frac{s^n}{n!} \frac{\partial H_n(\xi)}{\partial \xi}. \qquad (4.26)$$

Now we must identify the last partial derivative. If we look carefully at the six lowest-order Hermite polynomials given at the end of the last section, it may be seen that for these

$$\frac{\partial H_n(\xi)}{\partial \xi} = 2nH_{n-1}(\xi). \qquad (4.27)$$

This is a general result that can be established from the properly normalized power series representations for the Hermite polynomials. The proof is left to the problems. Using (4.27), we find that

$$\frac{\partial F}{\partial \xi} = 2sF(s,\xi) \qquad (4.28)$$

and

$$F(s,\xi) = F(s,0)\mathrm{e}^{2s\xi}. \qquad (4.29)$$

The coefficient may be obtained from (4.25) as

$$F(s,0) = \sum_{n=0}^{\infty} \frac{s^n}{n!} H_n(0). \qquad (4.30)$$

Again, by looking at the first six Hermite polynomials in the preceding section, we observe immediately that for $n$ odd, all of them vanish at 0, which is required by their anti-symmetric behaviour. For the even values, we note that we can cast the three even ones in the form $(-1)^{n/2}(n!)/(n/2)!$, which leads to the values $1, -2, 12, -120, \ldots$ for $n = 0, 2, 4, 6, \ldots$. We insert this directly into (4.30), and then make the change of index $k = n/2$, so that we achieve

$$F(s,0) = \sum_{k=0}^{\infty} \frac{s^{2k}}{k!}(-1)^k = \mathrm{e}^{-s^2}. \qquad (4.31)$$

Finally, we can write the generating function as

$$F(s,\xi) = \mathrm{e}^{-s^2}\mathrm{e}^{2s\xi} = \mathrm{e}^{\xi^2-(s-\xi)^2}. \qquad (4.32)$$

This result is based upon our rather free manner in which the coefficient of the lowest-order term was obtained. Thus, the relations that we are using are all based upon this first assumption, and we must eventually show this to be the proper case.

By the use of the generating function, we can find a general relationship for the Hermite polynomial. To begin, we note that the generating function (4.25) has been defined so that

$$H_n(\xi) = \left[\frac{\mathrm{d}^n}{\mathrm{d}s^n}F(s,\xi)\right]_{s=0}. \tag{4.33}$$

Using (4.31), this leads to

$$\begin{aligned} H_n(\xi) &= \left[\frac{\mathrm{d}^n}{\mathrm{d}s^n}\mathrm{e}^{\xi^2-(s-\xi)^2}\right]_{s=0} \\ &= \left[\mathrm{e}^{\xi^2}\frac{\mathrm{d}^n}{\mathrm{d}s^n}\mathrm{e}^{-(s-\xi)^2}\right]_{s=0} \\ &= \left[\mathrm{e}^{\xi^2}\frac{\mathrm{d}^n}{\mathrm{d}\xi^n}\mathrm{e}^{-(s-\xi)^2}(-1)^n\right]_{s=0} \\ &= (-1)^n\mathrm{e}^{\xi^2}\frac{\mathrm{d}^n}{\mathrm{d}\xi^n}\mathrm{e}^{-\xi^2}. \end{aligned} \tag{4.34}$$

While one might jump to the use of this to verify the initial values used in the definitions of the Hermite polynomials in this last section, recall that these definitions were built into the generating function, and should naturally result from its use. However, it is this simple formula for the Hermite polynomials that leads to the coefficients found in (4.23) and (4.24). Thus, it is actually (4.34) that gives us the lower-ordered Hermite polynomials listed in these latter equations.

The orthonormality of the Hermite polynomials can be examined through the use of combinations of the generating functions. Consider, for example, the integral

$$\begin{aligned} \mathcal{I} &= \int_{-\infty}^{\infty} F(s,\xi)F(t,\xi)\mathrm{e}^{-\xi^2}\,\mathrm{d}\xi \\ &= \int_{-\infty}^{\infty} \mathrm{e}^{\xi^2-(s-\xi)^2}\mathrm{e}^{\xi^2-(t-\xi)^2}\mathrm{e}^{-\xi^2}\,\mathrm{d}\xi \\ &= \mathrm{e}^{2st}\int_{-\infty}^{\infty} \mathrm{e}^{-(s+t-\xi)^2}\,\mathrm{d}\xi \\ &= \sqrt{\pi}\mathrm{e}^{2st} = \sqrt{\pi}\sum_n \frac{2^n s^n t^n}{n!}. \end{aligned} \tag{4.35}$$

Now, on the other hand, instead of the generating functions, we use the series expansions in terms of Hermite polynomials, as

$$\mathcal{I} = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty}\frac{s^n}{n!}H_n(\xi)\sum_{k=0}^{\infty}\frac{t^k}{k!}H_k(\xi)\mathrm{e}^{-\xi^2}\,\mathrm{d}\xi$$

$$= \sum_{n=0}^{\infty} \frac{s^n}{n!} \sum_{k=0}^{\infty} \frac{t^k}{k!} \int_{-\infty}^{\infty} H_n(\xi) H_k(\xi) \mathrm{e}^{-\xi^2} \, \mathrm{d}\xi. \tag{4.36}$$

The two equations (4.35) and (4.36) must be exactly equal term by term, because we began with two equivalent formulations. Thus, terms in (4.36) for which $n \neq k$ must vanish in the integral of (4.36), which provides the orthogonality of the Hermite polynomials. When $n = k$, equating the proper terms in (4.35) and (4.36) leads to

$$\int_{-\infty}^{\infty} H_n(\xi) H_k(\xi) \mathrm{e}^{-\xi^2} \, \mathrm{d}\xi = \sqrt{\pi} 2^n n! \delta_{nk} \tag{4.37}$$

where $\delta_{nk}$ is the Kronecker delta function, and is unity for $n = k$ and zero otherwise. Now, the truth is found; the values of the lowest-order terms for the Hermite polynomials in the previous section, and that we used to get the generating functions, do not normalize the wave functions. In fact, they arise from the convenient relation (4.34), obtained from the generating function, but the orthonormal forms of $\psi(\xi)$ used, for example, in (4.18) must be rewritten as

$$\psi_n(\xi) = \frac{1}{\sqrt{2^n n!}} \left( \frac{1}{\pi} \right)^{1/4} H_n(\xi) \tag{4.38}$$

and the full wave function is

$$\Psi_n(x) = \frac{1}{\sqrt{2^n n!}} \left( \frac{m\omega}{\hbar \pi} \right)^{1/4} \mathrm{e}^{-m\omega x^2/2\hbar} H_n \left( \sqrt{\frac{m\omega}{\hbar}} x \right). \tag{4.39}$$

## 4.3 Motion of the wave packet

In the classical harmonic oscillator, which is our simple linear pendulum, the position of the pendulum oscillates back and forth across the point $\theta = 0$, which is the point $x = 0$ in the present coordinates. We want to examine this effect on the quantized harmonic oscillator, and the quantized wave functions describing this oscillator. We can write the total wave function, at $t = 0$, as

$$\Psi(x, 0) = \sum_n c_n \Psi_n(x) \tag{4.40}$$

where the expansion wave functions are our normalized, and generalized, Hermite polynomials described by (4.39). Just as in (2.100), the expansion coefficients $c_n$ are given by

$$c_n = \int_{-\infty}^{\infty} \Psi_n^*(x) \Psi(x, 0) \, \mathrm{d}x. \tag{4.41}$$

We also need to recall that the fact that the probability must sum to unity requires that

$$\sum_n |c_n|^2 = 1. \tag{4.42}$$

We use (4.21) and (2.93) to write the time-varying total wave function as

$$\Psi(x,t) = \mathrm{e}^{-\mathrm{i}\omega t/2} \sum_n c_n \Psi_n(x) \mathrm{e}^{-\mathrm{i}n\omega t} \tag{4.43}$$

where we recall that $\omega$ is defined by the force (we replaced the gravitation constant by the square of the frequency). Thus, this frequency is not a variable, but defines the force constants of the problem.

If we calculate the average energy in the harmonic oscillator, we do so by the use of (1.17), or

$$\langle H \rangle = (\Psi, H\Psi) = \sum_n |c_n|^2 \mathcal{E}_n \tag{4.44}$$

so the measured energy is a weighted average of the energy values of the quantized levels of the harmonic oscillator. The probability of any particular energy eigenvalue $\mathcal{E}_n$ is given by $|c_n|^2$. Again, we may use (1.17) to compute the average position $x$ as

$$\begin{aligned}\langle x \rangle &= \sum_n c_n^* \sum_k c_k \mathrm{e}^{\mathrm{i}(n-k)\omega t} \int_{-\infty}^{\infty} \Psi_n^*(x) x \Psi_k(x) \,\mathrm{d}x \\ &= \sum_n c_n^* \sum_k c_k \mathrm{e}^{\mathrm{i}(n-k)\omega t} X_{nk}\end{aligned} \tag{4.45}$$

where the last line defines the *matrix element* $X_{nk}$ between the two different generalized Hermite polynomials. We need to compute this quantity to determine the various values of the average position as a function of time. Now, the integral may be expressed as

$$\begin{aligned}X_{nk} &= \frac{1}{\sqrt{\pi 2^{n+k} n! k!}} \int_{-\infty}^{\infty} \mathrm{e}^{-m\omega x^2/\hbar} H_n(x) H_k(x) x \,\mathrm{d}x \\ &= \sqrt{\frac{1}{2^{n+k}\pi n! k!} \frac{\hbar}{m\omega}} \int_{-\infty}^{\infty} \mathrm{e}^{-\xi^2} H_n(\xi) H_k(\xi) \xi \,\mathrm{d}\xi\end{aligned} \tag{4.46}$$

where we have re-introduced the reduced space variables. The approach to evaluate this expression follows that utilizing the generating function that led to (4.35). Consider the integral

$$\begin{aligned}\mathcal{I} &= \int_{-\infty}^{\infty} F(s,\xi) F(t,\xi) \mathrm{e}^{2\lambda\xi - \xi^2} \,\mathrm{d}\xi \\ &= \int_{-\infty}^{\infty} \mathrm{e}^{\xi^2-(s-\xi)^2} \mathrm{e}^{\xi^2-(t-\xi)^2} \mathrm{e}^{2\lambda\xi-\xi^2} \,\mathrm{d}\xi \\ &= \mathrm{e}^{2st+\lambda^2+2\lambda(t+s)} \int_{-\infty}^{\infty} \mathrm{e}^{-(s+t+\lambda-\xi)^2} \,\mathrm{d}\xi \\ &= \sqrt{\pi} \mathrm{e}^{2st+\lambda^2+2\lambda(t+s)} \\ &= \sqrt{\pi} \sum_n \frac{2^n s^n t^n}{n!} \sum_k \frac{\lambda^{2k}}{k!} \sum_j \frac{(2)^j (t+s)^j \lambda^j}{j!}.\end{aligned} \tag{4.47}$$

Now, by the same token, we can utilize the opposite side of the integral, obtained from (4.36), as

$$\begin{aligned}\mathcal{I} &= \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{s^n}{n!} H_n(\xi) \sum_{r=0}^{\infty} \frac{t^r}{r!} H_r(\xi) \mathrm{e}^{2\lambda\xi-\xi^2}\, \mathrm{d}\xi \\ &= \sum_{n=0}^{\infty} \frac{s^n}{n!} \sum_{r=0}^{\infty} \frac{t^r}{r!} \int_{-\infty}^{\infty} H_n(\xi) H_r(\xi) \sum_{j} \frac{2^j \lambda^j \xi^j}{j!} \mathrm{e}^{-\xi^2}\, \mathrm{d}\xi. \end{aligned} \tag{4.48}$$

The term that we are interested in has $j = 1$ in both (4.47) and (4.48). Moreover, we must set $k = 0$ in (4.47). This leaves us with terms like $s^n t^n (t + s)$ in (4.47), while in (4.48) the terms are like $s^n t^r$. In comparing these two terms, we see that $r = n \pm 1$. From this relationship, we can write the equivalence (for $\lambda = 1$) as

$$2\sqrt{\pi} \frac{(2)^n s^n t^n (t + s)}{n!} = 2 \frac{s^n}{n!} \frac{t^r}{r!} X_{nr} \tag{4.49}$$

or

$$X_{nr} = 2^n \sqrt{\pi} r! \delta_{r,n+1} + 2^{n-1} n \sqrt{\pi} r! \delta_{r,n-1} \tag{4.50}$$

where, for the second term, one replaces $n$ by $n-1$ on the left-hand side of (4.49). Caution must be exercised at this point, because the Hermite polynomials are not completely orthogonalized in these expansions, but this is corrected when we use this result in (4.46), which leads to

$$X_{nr} = \sqrt{\frac{\hbar}{m\omega}} \left[ \sqrt{\frac{n+1}{2}} \delta_{r,n+1} + \sqrt{\frac{n}{2}} \delta_{r,n-1} \right]. \tag{4.51}$$

The matrix element couples one level only to that level just above or below it. Thus, the position increases in time by moving from one energy level to the next higher (or next lower) excitation level. Finally, the average of the position is given by

$$\begin{aligned}\langle x \rangle &= \sum_n c_n^* \sum_k c_k \mathrm{e}^{\mathrm{i}(n-k)\omega t} X_{nk} \\ &= \sqrt{\frac{\hbar}{m\omega}} \sum_n c_n^* \sum_k c_k \mathrm{e}^{\mathrm{i}(n-k)\omega t} \left[ \sqrt{\frac{n+1}{2}} \delta_{k,n+1} + \sqrt{\frac{n}{2}} \delta_{k,n-1} \right] \\ &= \sqrt{\frac{\hbar}{2m\omega}} \sum_{n=1}^{\infty} \sqrt{n} (c_n^* c_{n-1} \mathrm{e}^{\mathrm{i}\omega t} + c_{n-1}^* c_n \mathrm{e}^{-\mathrm{i}\omega t}). \end{aligned} \tag{4.52}$$

The basic time dependence is clearly determined by the force parameters through the frequency and mass.

Computing the average of the momentum is a somewhat more direct process. For this, however, we will use a symmetrized version:

$$\langle p \rangle = (\langle p \rangle + \langle p^* \rangle)/2 \tag{4.53}$$

in order to ensure that the result is a real quantity. As for the case of the position, we expect that this will result in a series:

$$\langle p \rangle = \sum_n c_n^* \sum_k c_k \mathrm{e}^{\mathrm{i}(n-k)\omega t} P_{nk} \tag{4.54}$$

where

$$\begin{aligned} P_{nk} &= \frac{1}{\sqrt{\pi 2^{n+k} n! k!}} \int_{-\infty}^{\infty} \mathrm{e}^{-m\omega x^2/\hbar} H_n(x) p H_k(x)\,\mathrm{d}x \\ &= \sqrt{\frac{1}{2^{n+k} n! k!}} \sqrt{\frac{m\omega}{\hbar}} \int_{-\infty}^{\infty} \mathrm{e}^{-\xi^2} H_n(\xi) p H_k(\xi)\,\mathrm{d}\xi \end{aligned} \tag{4.55}$$

and the momentum operator is in the reduced coordinates. Now, using the symmetrized form of the momentum leads us to the representation of the integral as

$$\begin{aligned} 2\mathcal{I} &= (\psi, (p + p^*)\psi) = (\psi, p\psi) + (p\psi, \psi) \\ &= \mathrm{i}\hbar \int_{-\infty}^{\infty} \mathrm{e}^{-\xi^2} \left[ \frac{\mathrm{d}H_n(\xi)}{\mathrm{d}\xi} H_k(\xi) - H_n(\xi) \frac{\mathrm{d}H_k(\xi)}{\mathrm{d}\xi} \right] \mathrm{d}\xi. \end{aligned} \tag{4.56}$$

The derivative has been expressed earlier in (4.25), so the integral becomes

$$\mathcal{I} = \frac{\mathrm{i}\hbar}{2} \sqrt{\pi} 2^n n! [\delta_{n-1,k} - 2k\delta_{n+1,k}] \tag{4.57}$$

which, when properly normalized through (4.52), becomes

$$P_{nk} = \mathrm{i}\sqrt{\hbar m\omega} \left[ \sqrt{\frac{n}{2}} \delta_{n-1,k} - \sqrt{\frac{n+1}{2}} \delta_{n+1,k} \right]. \tag{4.58}$$

This now leads to the expectation of the momentum as

$$\langle p \rangle = \mathrm{i}\sqrt{\frac{\hbar m\omega}{2}} \sum_{n=1}^{\infty} \sqrt{n} [c_n^* c_{n-1} \mathrm{e}^{\mathrm{i}\omega t} - c_{n-1}^* c_n \mathrm{e}^{-\mathrm{i}\omega t}]. \tag{4.59}$$

It is reassuring that if we develop $\langle p \rangle = m\,\mathrm{d}\langle x \rangle/\mathrm{d}t$ using (4.52), the same answer is obtained for the expectation value of the momentum, as expected from the treatment of section 2.8.1 and equation (2.96). We also note that the momentum increases or decreases on moving from one energy level to the next higher, or lower, respectively, energy level.

## 4.4 A simpler approach with operators

The Hamiltonian developed for the harmonic oscillator is doubly quadratic; that is, it contains one term quadratic in the momentum and one term quadratic in the position. We want to rearrange this Hamiltonian, by introducing a set of operators that combine the position and momentum in a way that leads to a simpler approach to the problem. The rationale is based upon the fact that we have discovered that both the position and momentum expectation values vary by the 'jump' of the state from one eigenfunction to another by moving up or down in the energy levels. We suppose that a proper combination of the individual operators will provide a new set of operators that corresponds only to upward movement, or only to downward movement, among the energy levels. This will be the case, and as a result the entire approach becomes somewhat simpler. Because of the various coefficients in the Hamiltonian, the operators will be defined by combinations of these coefficients as well; the guiding principle will be that the Hamiltonian should be a simple product of the resulting operators.

With the above concepts in mind, we begin by defining the pair of adjoint operators in terms of the position and momentum operators as

$$a = \sqrt{\frac{m\omega}{2\hbar}}\left(x + \mathrm{i}\frac{p}{m\omega}\right) \tag{4.60$a$}$$

$$a^{+} = \sqrt{\frac{m\omega}{2\hbar}}\left(x - \mathrm{i}\frac{p}{m\omega}\right). \tag{4.60$b$}$$

These operators may be combined in the manner

$$\begin{aligned} a^{+}a &= \frac{m\omega}{2\hbar}\left(x - \mathrm{i}\frac{p}{m\omega}\right)\left(x + \mathrm{i}\frac{p}{m\omega}\right) \\ &= \frac{1}{\hbar\omega}\left(\frac{p^2}{2m} + \frac{m\omega^2}{2}x^2\right) - \frac{1}{2} \end{aligned} \tag{4.61}$$

or, upon comparing with the Hamiltonian,

$$H = \hbar\omega(a^{+}a + \tfrac{1}{2}). \tag{4.62}$$

By comparing with the energy in (4.21), we *assume* that the product is the number of particles in a level, so this product is called the *number operator*. We will show below that this product of operators (unusually) produces a non-operator, called a $c$-number (for constant number), that gives the quantum number of the energy level corresponding to a particular state, when operating on the Hamiltonian eigenstates.

The momentum and position operators satisfy a basic commutator relation since they are non-commuting operators. We expect that the above operators are also non-commuting since they arise from combinations of these non-commuting operators. The reverse product to (4.61) is given by

$$\begin{aligned} aa^{+} &= \frac{m\omega}{2\hbar}\left(x + \mathrm{i}\frac{p}{m\omega}\right)\left(x - \mathrm{i}\frac{p}{m\omega}\right) \\ &= \frac{1}{\hbar\omega}\left(\frac{p^2}{2m} + \frac{m\omega^2}{2}x^2\right) + \frac{1}{2} \end{aligned} \tag{4.63}$$

and so

$$aa^+ - a^+a = 1. \tag{4.64}$$

This establishes the desired commutator relation. We note that both terms in (4.64) are $c$-numbers, from which the Hamiltonian can be defined.

The operator pair $a^+a$ is a $c$-number, so if we multiply a function by this quantity, when the function is one of the eigenfunctions of the harmonic oscillator Hamiltonian, we expect to obtain the corresponding eigenvalue, that is

$$a^+a\Psi_n = \lambda_n\Psi_n \tag{4.65}$$

and similarly for the opposite choice of ordering, where $\lambda_n$ is the eigenvalue still to be determined (but that we expect to be $n$). Certainly, $\lambda_n$ is positive, since

$$(\Psi_n, a^+a\Psi_n) = (a\Psi_n, a\Psi_n) \geq 0. \tag{4.66}$$

Suppose that we operate with this operator on a state that has been arrived at from the product $a^+\Psi_n$. Then,

$$(a^+a)a^+\Psi_n = a^+(a^+a + 1)\Psi_n = a^+(\lambda_n + 1)\Psi_n = (\lambda_n + 1)a^+\Psi_n. \tag{4.67}$$

Thus, if the wave function (and state) $\Psi_n$ has the eigenvalue $\lambda_n$, then operating on this state with the operator $a^+$ produces a new state with the eigenvalue $(\lambda_n + 1)$. In essence, we have 'kicked' the energy of the initial state *upward* by one unit (which may be observed by reference to the Hamiltonian), and produced a new wave function corresponding to the level of higher energy. By the same token, if we operate with $a$, we find

$$(a^+a)a\Psi_n = (aa^+ - 1)a\Psi_n = (\lambda_n - 1)a\Psi_n. \tag{4.68}$$

Thus, operating with the operator $a$ produces a new state with the eigenvalue $(\lambda_n - 1)$. In essence, we have 'kicked' the energy of the initial state *downward* by one unit, and produced a new wave function corresponding to the level of lower energy. For these reasons, we normally refer to $a^+$ as a *raising* operator (or *creation* operator since it creates one additional unit of energy in the system) and $a$ as a *lowering* operator (or *annihilation* operator, since it destroys one unit of energy).

Now, there is always a lowest-energy state, the *ground state*. In the harmonic oscillator, it corresponds to $n = 0$, the state in which the Hermite polynomial is just $H_0$. Thus, if we try to lower the energy, and kick the state down by one unit, by operating with the lowering operator, the result must be zero:

$$a\Psi_0 = 0. \tag{4.69}$$

This means that $\lambda_0 = 0$, and by the repeated use of (4.67), we find that

$$\lambda_n = n \tag{4.70}$$

as expected. We can also use (4.69) to determine the eigenfunctions, since

$$a\Psi_0 = \sqrt{\frac{m\omega}{2\hbar}}\left(x + \mathrm{i}\frac{p}{m\omega}\right)\Psi_0 = 0. \tag{4.71}$$

Rearranging the terms leads to the first-order differential equation (already an improvement over the differential equation for the Hermite polynomials)

$$-\frac{\hbar}{m\omega}\frac{\mathrm{d}\Psi_0}{\mathrm{d}x} = x\Psi_0 \tag{4.72}$$

and

$$\Psi_0(x) = C_0 \exp\left(-\frac{m\omega x^2}{2\hbar}\right) \tag{4.73}$$

which may be recognized as $H_0$ with its exponential weighting function. By direct normalization, we find that

$$C_0 = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4}. \tag{4.74}$$

From (4.64), we find that we can use (4.73) and (4.74) to show that

$$\Psi_n = (a^+)^n\Psi_0 \sim \left(\frac{m\omega}{2\hbar}\right)^{n/2}\left(x - \mathrm{i}\frac{p}{m\omega}\right)^n\Psi_0. \tag{4.75}$$

Thus, by simply solving the one first-order differential equation, we can now use the simple differential operators to find the wave function for any energy level in the harmonic oscillator!

We now want to turn to the expectation values of the operators themselves. From the basic properties expressed in (4.64) and (4.65), we know that for unnormalized wave functions such that $(\Psi_n, \Psi_n) = C_n^2$, we can write (we take the coefficients as real)

$$(\Psi_k, a^+\Psi_n) = (\Psi_k, \Psi_{n+1}) = \alpha_n\delta_{n+1,k}C_n^2 \tag{4.76a}$$

$$(\Psi_k, a\Psi_n) = (\Psi_k, \Psi_{n-1}) = \beta_n\,\delta_{n-1,\,k}C_n^2. \tag{4.76b}$$

We still have to determine the constants $\alpha_n$ and $\beta_n$, which in fact are related to the normalization of the eigenfunctions (the two eigenfunctions in the expectation value have different normalization constants). For this, we use

$$(a\Psi_n, a\Psi_n) = (\Psi_n, a^+a\Psi_n) = nC_n^2 = (\Psi_{n-1}, \Psi_{n-1}) = C_{n-1}^2 \tag{4.77}$$

and

$$C_{n-1} = \sqrt{n}C_n. \tag{4.78}$$

We could as easily have found this value by extending the arguments that led to (4.74) and the earlier discussion of $\lambda_n$, but it is instructive to repeat the work and reinforce this point of understanding. Thus, we find that $\alpha_n = C_{n+1}/C_n = \sqrt{n+1}$, and the normalized relationship becomes

$$(\Psi_k, a^+\Psi_n) = (\Psi_k, \Psi_{n+1}) = \sqrt{n+1}\delta_{n+1,k}. \tag{4.79$a$}$$

Similarly, $\beta_n = C_{n-1}/C_n = \sqrt{n}$, and the normalized relationship becomes

$$(\Psi_k, a\Psi_n) = (\Psi_k, \Psi_{n-1}) = \sqrt{n}\delta_{n-1,k}. \tag{4.79$b$}$$

These now lead to

$$\psi_n = \frac{1}{\sqrt{(n+1)!}}(a^+)^n\psi_0. \tag{4.79$c$}$$

In these, $n$ begins with 1 and works upward, just as was done in (4.52) and (4.59). We can now use (4.61) to find the expectation of the momentum and position as

$$x_{nk} = \sqrt{\frac{\hbar}{2m\omega}}(a + a^+)_{nk} \tag{4.80$a$}$$

$$p_{nk} = -\mathrm{i}\sqrt{\frac{\hbar m\omega}{2}}(a - a^+)_{nk}. \tag{4.80$b$}$$

To complete the form shown in the previous equations, we perform the double sum over the two sets of wave functions, including the temporal variation from the energy levels. It must be concluded that taking this approach is considerably simpler and easier to accomplish than taking that of the previous section, and it is for this reason that special operators like the raising and lowering operators have come to be used so extensively.

We note from the form of $(4.79c)$ that there can be $n$ particles in a given mode; that is we may excite the harmonic oscillator to contain many units of the quantized energy. This means that these particles cannot be subject to the Pauli exclusion principle. If they were subject to the Pauli exclusion principle, there could be no more than two particles (of opposite spin) in this harmonic oscillator. Indeed, the product $a^+a$ is called the number operator, in which its action on the wave function produces the number $n$ of that wave function, which yields the number of excited levels we have. Particles which do not obey the Pauli exclusion principle, and which can be described by these harmonic oscillator states, are called *bosons*. This is in contrast to particles which satisfy the Pauli exclusion principle and are called *fermions*, as they obey Fermi–Dirac statistics as well. We will treat the creation and annihilation operators for fermions in section 9.4.

## 4.5 Quantizing the $LC$-circuit

The $LC$-circuit is one of the most pervasive systems found in electrical engineering. It represents not only a range of filters, but also the resonant cavity of microwave circuits. In this section, we want to show that the quantum mechanics of the $LC$-circuit is essentially that of the linear harmonic oscillator. To begin, we consider the energy stored in the $LC$-circuit at any instant of time (we take a parallel $LC$-circuit, but this is not an important assumption and does not appear anywhere in the treatment). We can write the energy stored as

$$\mathcal{E} = \tfrac{1}{2}CV^2 + \tfrac{1}{2}LI^2. \tag{4.81}$$

The first term on the right-hand side is the energy stored in the capacitor, while the second term is that stored in the inductor. In fact, we can proceed immediately to proclaim that this is the Hamiltonian that we should use in the Schrödinger equation. However, before we can do that, it is still necessary to identify the conjugate operators and their relationship to each other.

One obvious possibility is to take the charge $Q$ $(=CV)$ as one operator and then the current $I = \mathrm{d}Q/\mathrm{d}t$ as the other (in analogy between the position and the velocity). However, this doesn't quite work out. The analogy would have the current as the velocity (of the charge) and not the momentum. We can be guided by our understanding that the two fundamental quantities in circuits are the charge and the *flux*. It is these two quantities that are invariant in special relativity theory and it is these two that are subject to important conservation laws in circuit theory—the conservation of charge and the conservation of flux linkages. Moreover, the flux linkage in the present context is just the flux in the inductor, given by $\phi = LI$. In section 3.8, we took the charge as the 'momentum', but here it is better to take the charge as the 'coordinate' variable, and take the flux as the 'momentum' variable. If this is done, we can rewrite (4.81) as

$$\mathcal{E} = \frac{\phi^2}{2L} + \frac{Q^2}{2C}. \tag{4.82}$$

By introducing the resonant frequency through $\omega^2 = 1/LC$, this can be rewritten as

$$\mathcal{E} = \frac{\phi^2}{2L} + L\omega^2\frac{Q^2}{2}. \tag{4.83}$$

Thus, if we make the connection $L \to m$, this is the same equation as (4.13). Hence, we find that the resonant circuit is just a simple harmonic oscillator.

Finally, let us look at the commutator of charge and flux, when these are interpreted as operators. By making the analogy between the flux and the momentum coordinate, and between the charge and position coordinate, we have from (1.27) the relationship

$$\phi = -\mathrm{i}\hbar\frac{\partial}{\partial Q} \tag{4.84}$$

which satisfies the commutator relationship

$$[\phi, Q] = -\mathrm{i}\hbar. \tag{4.85}$$

In this interpretation, the charge and flux are non-commuting variables, with the flux interpreted as the variation with respect to the charge. On the other hand, this relationship vanishes in the classical limit as $\hbar \to 0$. In this latter limit, charge and flux are independent quantities, but in the quantum case they are not independent and cannot be simultaneously measured.

The interpretation of the $LC$-circuit as a linear harmonic oscillator means that the energy in the resonator is a fixed quantity that is given in terms of the resonant frequency by (4.21). Energy storage can be increased or decreased only in units of the energy quantum $\hbar\omega = \hbar/\sqrt{LC}$. Thus, the photons that enter or leave an electromagnetic cavity (circuit) have well defined energies in terms of the resonant frequency of the cavity. This is obvious in the measurement of the frequency of the oscillating cavity, where we measure the energy of the photons emanating from the cavity. However, it is often useful to consider the resonant circuit quantum mechanically rather than classically. We have the full understanding of this in terms of the harmonic oscillator results of the previous sections.

With this discussion, it is now clear that electromagnetic waves satisfy the resonant $LC$-circuit, which itself is a linear harmonic oscillator. We previously established that the particles which satisfied the harmonic oscillator were *bosons*. Thus, it is clear that photons, the particles which are associated with electromagnetic waves, are bosons. The photons actually have spin, but it is quantized into values of $\pm 1$, corresponding to the circular polarization of the plane-wave of electromagnetics. Thus, we may say that bosons have *integer* values of spin, while fermions have half-integer values of spin.

This approach also opens the door to a treatment of almost any oscillatory behaviour to be treated by a quantum approach. As here, one only needs to identify the proper *generalized* coordinates, which are then subjected to an uncertainty principle. It is this step that introduces the quantum nature of the subsequent solutions in terms of harmonic oscillator coordinates. Generalized coordinates have long been treated in classical mechanics, and their use naturally flows over to quantum mechanics.

## 4.6 The vibrating lattice

In a solid, the motion of the atoms is much like the motion of free particles, with the important exception that the atoms are forced *on average* to retain positions within the solid that specify a particular crystal structure (we can also treat non-crystalline solids by similar methods, but our approach here will be limited to the crystalline solids, where the atoms are equally spaced on a *lattice*). In any real solid, the lattice is a three-dimensional structure. However, when the motion is along one of the principal axes of the structure, it is usually possible to treat the atomic motion as a one-dimensional system. Although this is a simple model, its applicability can be extended to the real crystal if each atom represents the typical motion of an entire plane of atoms normal to the wave motion.

The total Hamiltonian for the atomic motion is given by the momentum of the individual atoms plus the potential that keeps the atoms in the lattice positions (on average). We can write this Hamiltonian as

$$H = \sum_i \frac{P_i^2}{2M_i} + \sum_{i \neq j} V(R_i - R_j). \tag{4.86}$$

In general, the atoms will vibrate around their equilibrium position, just like little harmonic oscillators. However, all of these oscillators are coupled through the potential term in (4.86). We have studied only the single oscillator. How will we treat the coupled oscillators? The answer to this is that we do a Fourier transform into the dominant Fourier modes of the overall vibration, and this will result in us obtaining a set of uncoupled oscillators, one for each mode of vibration of the entire set of coupled oscillators. Each of these modes will then be quantized, and the quantum unit of amplitude of each of the modes is termed a *phonon*.

The Hamiltonian in (4.86) describes the motion of the atoms about the equilibrium (or rest) positions. We expect this motion to be small, so shall use a Taylor series expansion about these equilibrium positions, and

$$H = \sum_i \frac{P_i^2}{2M_i} + \frac{1}{2} \sum_{i \neq j} \delta r_i \delta r_j \frac{\partial^2 V(R_i - R_j)}{\partial R_i \partial R_j} + \cdots. \tag{4.87}$$

The zero-order term in the potential is just an offset in energy and will be ignored, while the first partial derivative term must be zero if we are expanding about the equilibrium position. We also recognize that

$$P_i = M_i \frac{\mathrm{d}\delta r_i}{\mathrm{d}t}. \tag{4.88}$$

Now, we introduce the appropriate Fourier series for the displacement as

$$\delta r_i = \frac{1}{\sqrt{N}} \sum_q u_q \mathrm{e}^{\mathrm{i}qR_i} \tag{4.89}$$

where we assume that there is a lattice constant $a$ that is the equilibrium distance between atoms and that, from some arbitrary reference point, $R_{i0} = n_i a$. Here, $n_i$ is an integer specifying just how far along the chain of atoms the $i$th atom resides in equilibrium. The last term in the Hamiltonian is the most complicated one. This is given by the Fourier terms

$$\begin{aligned} \mathcal{V}_{qq'} &= \frac{1}{N} \sum_{i \neq j} u_q \mathrm{e}^{\mathrm{i}qR_i} u_{q'} \mathrm{e}^{\mathrm{i}q'R_j} \frac{\partial^2 V(R_i - R_j)}{\partial R_i \partial R_j} \\ &= \frac{1}{N} \sum_{i \neq j} u_q \mathrm{e}^{\mathrm{i}q(R_i - R_j)} u_{q'} \mathrm{e}^{\mathrm{i}(q'+q)R_j} \frac{\partial^2 V(R_i - R_j)}{\partial R_i \partial R_j}. \end{aligned} \tag{4.90}$$

We assume that the partial derivative can be treated as constant (but may vary with the particular mode). Each of the position vectors is expanded in terms of its equilibrium position and its absolute position. The first exponential depends only upon the relative positions, while the second exponential mainly treats the equilibrium positions (the deviations are small). Thus, we bring this latter summation out to make a sum only over the equilibrium positions:

$$\sum_j \mathrm{e}^{\mathrm{i}(q'+q)R_j} = \sum_{n_j} \mathrm{e}^{\mathrm{i}(q'+q)n_j a} = N\delta_{q,-q'}. \tag{4.91}$$

To achieve the last term, we assert that the summation over $n_j$ is a summation over all the Fourier modes, which by the principle of closure of a complete set must vanish unless the coefficient itself vanishes. This leads to the Kronecker delta. The number $N$, which also appears in (4.87), is the number of atoms in the chain (and hence the number of Fourier modes). The remaining part of the potential is just (summing over $q'$, with $\mathcal{V}_{qq'} = \mathcal{V}_q$)

$$\mathcal{V}_q = \sum_{i\neq j} u_q u_{-q} \mathrm{e}^{\mathrm{i}q(\delta r_i - \delta r_j)} \frac{\partial^2 V(R_i - R_j)}{\partial \delta r_i \partial \delta r_j}. \tag{4.92}$$

The mode amplitudes can be brought out of the summation, and the remainder is defined to be the spring constant for the particular mode $C_q$. The same summation over the lattice modes can be made in the kinetic energy term as well, and this allows us to write the final result as

$$H = \sum_q \left[\frac{1}{2M}P_q P_{-q} + \frac{C_q}{2}u_q u_{-q}\right] = \sum_q \left[\frac{1}{2M}P_q P_q^+ + \frac{C_q}{2}u_q u_q^+\right] \tag{4.93}$$

where

$$P_q = -\mathrm{i}\hbar\frac{\mathrm{d}u_q}{\mathrm{d}t} \tag{4.94}$$

is the mode momentum, and we have used general properties of the Fourier coefficients. By writing $C_q = M\omega_q^2$, we see that we have now obtained a set of uncoupled harmonic oscillators. Each mode is characterized by a wave vector $q$ and frequency $\omega_q$, and we can compute the wave functions as if that mode were totally isolated with a mode energy described by (4.21). Thus, this particular mode may contain a great many phonons as described by the energy level number for the mode. The total energy of the lattice vibration is now just the sum over the energy (and hence the number of phonons) in each mode and a sum over the various modes. Quantization of the vibrations occurs by having the position $u_q$ and the momentum $P_q$ be non-commuting operators subject to a commutator relationship $[P_q, u_q] = -\mathrm{i}\hbar$. We can at this point also introduce the creation and annihilation operators for each mode, which serve to create or destroy one unit of amplitude for that particular mode. Thus, the number of units of amplitude (or energy) in the mode of wave vector $\boldsymbol{q}$ is said to the number of phonons in that particular mode.

Because each mode of the lattice vibrations satisfies a harmonic oscillator equation, these *phonon* modes are *bosons*. The atoms which constitute the vibrations are such that the phonon mode has zero spin, which counts as an integer. In terms of the parameters of (4.93), the creation and annihilation operators are written as

$$\begin{aligned} a_q^+ &= \sqrt{\frac{C}{2\hbar\omega_q}}\left(u_q - \mathrm{i}\frac{\omega_q}{C}P_q\right) \\ a_q &= \sqrt{\frac{C}{2\hbar\omega_q}}\left(u_q + \mathrm{i}\frac{\omega_q}{C}P_q\right). \end{aligned} \tag{4.95}$$

In terms of these operators, the amplitude of the $q$th Fourier mode is given by

$$u_q = \sqrt{\frac{\hbar\omega_q}{2C}}(a_q + a_q^+) \tag{4.96}$$

and the energy in this mode is

$$E_q = \hbar\omega_q(N_q + \tfrac{1}{2}). \tag{4.97}$$

The number of phonons in mode $q$ is given by the *Bose–Einstein* distribution

$$N_q = \frac{1}{\exp\left(\frac{\hbar\omega_q}{k_{\mathrm{B}}T}\right) - 1} \tag{4.98}$$

which differs from the Fermi–Dirac distribution by the sign in the denominator and the lack of a Fermi energy. All bosons satisfy this distribution but, as with the Fermi–Dirac distribution, the limiting case in which the factor of unity can be ignored gives the classical Maxwell–Boltzmann distribution.

Finally, we need to address the question of just how the wave vector $\boldsymbol{q}$ and the frequency $\omega_q$ are related to each other. For this, we return to real space, but will use the fact that the Fourier modes require the motion to exist as a set of waves with propagation according to $\mathrm{e}^{\mathrm{i}(qR-\omega_q t)}$. Thus, if we write the equation of motion of any particular atom at position $R_i$, we obtain from (4.86)

$$\begin{aligned} M\frac{\mathrm{d}^2\delta r_i}{\mathrm{d}t^2} = F_i &= -\sum_{j\neq i}\frac{\partial V(R_i - R_j)}{\partial \delta r_i} \\ &\simeq C(R_{i+1} - R_i) - C(R_i - R_{i-1}) \end{aligned} \tag{4.99}$$

where the approximation is that only the nearest neighbours are important, and we have treated the potential as a linear spring. Because of the lattice behaviour of the waves, we can rewrite the last term as

$$-M\omega_q^2\delta r_i = C\delta r_i[(\mathrm{e}^{\mathrm{i}qa} - 1) + (\mathrm{e}^{-\mathrm{i}qa} - 1)] \tag{4.100}$$

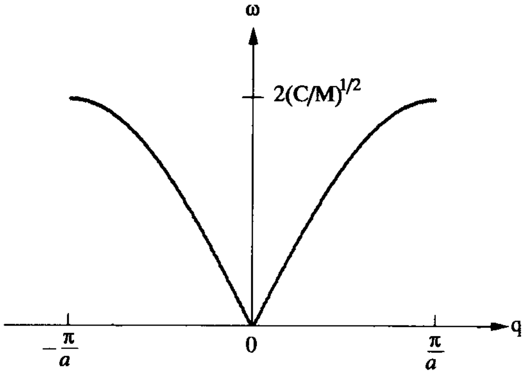


**Figure 4.5.** The relationship between the frequency and wave vector of a particular mode of lattice vibration.

and

$$\omega_q^2 = \frac{2C}{M}[1 - \cos(qa)] = \frac{4C}{M}\sin^2\left(\frac{qa}{2}\right). \tag{4.101}$$

Now, any range of $q$ that spans a total value of $2\pi/a$ will give all allowed values of the frequency that can be expected (which is of course positive). We generally choose this range to be $-\pi/a < q < \pi/a$, in keeping with the band structure arguments concerning periodic potentials in the last chapter. We show these frequencies in figure 4.5. Real lattices can be more complicated, containing for example, two different types of atom, or having the spacing between alternate atoms differing. In each case, the behaviour is more complicated, and motion of the two atoms together (low frequency) plus relative motion of the two atoms (high frequency) can occur. This is beyond the level we treat here, where we want merely to introduce the periodic potential as an example of a harmonic oscillator.

## 4.7 Motion in a quantizing magnetic field

The last example of a harmonic oscillator we want to consider is the motion of an electron orbiting around magnetic field lines. Classically, the electron is pulled into a circular orbit such that the magnetic field is directed normal to the plane of the orbit, and the motion arises from the Lorentz force acting on the electron:

$$\boldsymbol{F} = -e(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{4.102}$$

The charge on the electron has been taken to be its proper negative value. Clearly, for motion in the $\boldsymbol{a}_\phi$-direction (in cylindrical coordinates) and a magnetic field in the $\boldsymbol{a}_z$-direction, the force is an inward centripetal force which causes the motion to be a closed circular orbit (ignoring the role of the electric field). Here, we want to examine the quantization of this orbit, and will deal only with the electric field that is induced by the magnetic field $\boldsymbol{B}$, ignoring any explicitly applied field.

The approach we follow is exactly the one we used in section 1.3. The accelerative force on the wave momentum of the electron may be expressed as in (1.7) as

$$\frac{\hbar \mathrm{d}\boldsymbol{k}}{\mathrm{d}t} = -e\boldsymbol{E}. \tag{4.103}$$

In fact, we should use the entire Lorentz force in (4.103), but we want to introduce the magnetic field through the electric field that it produces. The approach is the one used in (1.8), but we will be somewhat more advanced here. The magnetic field produces the electric field through Faraday's law, which may be expressed as

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}. \tag{4.104}$$

We now introduce the concept of the *vector potential* $\boldsymbol{A}$, from which the magnetic field may be determined as

$$\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A} \tag{4.105}$$

so (4.104) and (4.105) may be combined as

$$\boldsymbol{E} = -\frac{\partial \boldsymbol{A}}{\partial t}. \tag{4.106}$$

Here, the vector curl operation has been dropped, which means that (4.105) is satisfied up to a quantity for which the curl vanishes. This quantity is just $-\boldsymbol{\nabla}\phi$, where $\phi$ is the scalar potential with which we usually define the electric field in quasi-static situations. Here, however, we want to define the electric field from the magnetic field alone, and hence from the vector potential that gives rise to the magnetic field. This suggests that if we now use (4.106) in (4.103), the proper total momentum to use in the Hamiltonian is just the combined quantity

$$\boldsymbol{p} \to (\hbar\boldsymbol{k} + e\boldsymbol{A}). \tag{4.107}$$

While the substitution (4.107) has been derived and developed by a great many workers, it is usually referred to as the Peierls substitution in solid-state physics.

We are now interested in using this value for the momentum in the Hamiltonian (4.2), except that we do not include any external potential. Nevertheless, the magnetic field will provide such a potential. To proceed, we must decide upon the vector potential $\boldsymbol{A}$. Our interest is in motion perpendicular to the magnetic field, which we take to be constant and oriented in the positive $z$-direction. Then, a vector potential defined as $\boldsymbol{A} = Bx\boldsymbol{a}_y$ will provide this magnetic field. We note that this definition is only *sufficient*, as many other definitions of the vector potential could as easily describe a uniform field in the $z$-direction. However, this particular definition, called the *Landau gauge*, will suit our purposes sufficiently well. Further, we will only treat the motion in the plane normal to the magnetic field.

In the absence of any external potential, other than the vector potential giving rise to the magnetic field, Schrödinger's equation can be written as

$$H\Psi = \left[-\frac{\hbar^2}{2m}\left(\boldsymbol{\nabla} - \frac{e\boldsymbol{A}}{\mathrm{i}\hbar}\right)^2\right]\Psi = \mathcal{E}\Psi \tag{4.108}$$

where the wave momentum has been replaced by the normal gradient operator for the momentum. Expanding the momentum operator bracket leads to the differential equation

$$-\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x^2} - \frac{2eBx}{\mathrm{i}\hbar}\frac{\partial}{\partial y} + \frac{\partial^2}{\partial y^2} - \left(\frac{eBx}{\hbar}\right)^2\right]\Psi = \mathcal{E}\Psi. \tag{4.109}$$

The motion in the $y$-direction is essentially free motion, at least within this formulation of the Schrödinger equation, as the only deviation from the normal differential operators is the $x$-dependence arising from the magnetic field. Thus, we will take the wave function to have the general form

$$\Psi(x,y) = \mathrm{e}^{-\mathrm{i}ky}\psi(x). \tag{4.110}$$

With this substitution, (4.109) becomes

$$-\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x^2} + \frac{2eBkx}{\hbar} - k^2 - \left(\frac{eBx}{\hbar}\right)^2\right]\psi = \mathcal{E}\psi. \tag{4.111}$$

Our goal is to put this into a form reminiscent of (4.9), the equation for the Hermite polynomials and the standard harmonic oscillator. To this end, we introduce the cyclotron frequency $\omega_\mathrm{c} = eB/m$, and rewrite (4.111) as

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + \frac{m}{2}\omega_\mathrm{c}^2(x - x_0)^2\psi = \mathcal{E}\psi \tag{4.112}$$

where

$$x_0 = \frac{\hbar k}{eB} \tag{4.113}$$

and we recall that $k$ is the $y$-component of the wave motion. Comparison of this with (4.9) shows that the $x$-motion satisfies a harmonic oscillator equation, which yields Hermite polynomials as solutions. Thus, the energy levels are given as

$$\mathcal{E}_n = \hbar\omega_\mathrm{c}(n + \tfrac{1}{2}) \tag{4.114}$$

and the corresponding wave function is just

$$\psi_n(x) = \frac{1}{\sqrt{2^n n!}}\left(\frac{m\omega_\mathrm{c}}{\hbar\pi}\right)^{1/4}\mathrm{e}^{-m\omega_\mathrm{c}(x-x_0)^2/2\hbar}H_n\left(\sqrt{\frac{m\omega_\mathrm{c}}{\hbar}}(x - x_0)\right). \tag{4.115}$$

These solutions are, of course, just the normal harmonic oscillator wave functions shifted by the offset (4.113). Here, both the scaling factor and the offset depend upon the magnetic field, just as the cyclotron frequency does. Noting the argument of the Hermite polynomial in (4.115) gives a natural scaling length described by $(\hbar/eB)^{1/2}$, which is termed the *magnetic length*. To lowest order, the Hermite polynomial has its peak at the position $\sqrt{2n+1}$ times this basic length, so this is the natural cyclotron radius of the harmonic motion for a particular energy level.

### 4.7.1 Connection with semi-classical orbits

The form of the wave function (4.115) and the energy (4.114) do not give us a clear view of the orbiting nature of the electron in the magnetic field. True, it is a harmonic oscillator in the $x$-direction, but what about its other motion? As remarked earlier, the electron orbits around the magnetic field, and this motion is a solution of (4.102), which we can rewrite as

$$m\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -e\boldsymbol{v} \times \boldsymbol{B}. \tag{4.116}$$

The magnetic field is in the $z$-direction, according to the Landau gauge previously adopted, so that the motion in the $(x, y)$-plane is given by

$$\begin{aligned} m\frac{\mathrm{d}\boldsymbol{v}_x}{\mathrm{d}t} &= -e\boldsymbol{v}_y B \\ m\frac{\mathrm{d}\boldsymbol{v}_y}{\mathrm{d}t} &= e\boldsymbol{v}_x B. \end{aligned} \tag{4.117}$$

These two equations may then be solved, for an arbitrary initial condition, as

$$\begin{aligned} \boldsymbol{v}_x &= \boldsymbol{v}_0 \cos(\omega_\mathrm{c} t) \\ \boldsymbol{v}_y &= \boldsymbol{v}_0 \sin(\omega_\mathrm{c} t). \end{aligned} \tag{4.118}$$

Similarly, we can now find the position as

$$\begin{aligned} x &= \frac{v_0}{\omega_\mathrm{c}} \sin(\omega_\mathrm{c} t) \\ y &= -\frac{v_0}{\omega_\mathrm{c}} \cos(\omega_\mathrm{c} t). \end{aligned} \tag{4.119}$$

We can now combine (4.118), (4.119), and (4.114) to give

$$\begin{aligned} E = \tfrac{1}{2} m\boldsymbol{v}_0^2 = \hbar\omega_\mathrm{c}(n + \tfrac{1}{2}) \qquad \boldsymbol{v}_0 = \sqrt{(2n+1)\frac{\hbar eB}{m^2}} \\ r^2 = x^2 + y^2 = \frac{v_0^2}{\omega_\mathrm{c}^2} = \sqrt{(2n+1)\frac{\hbar}{eB}}. \end{aligned} \tag{4.120}$$

The last line gives us the result hypothesized at the end of the previous section, $\sqrt{\hbar/eB}$ is a magnetic length that gives us the radius of the lowest energy level and provides a natural scaling of the size of the cyclotron orbit around the magnetic field. The quantization of the orbit radius arises directly from the quantization of the energy levels of the harmonic oscillator. These quantized energy levels are termed *Landau levels*.

One of the important aspects of the quantization in a magnetic field is the fact that the continuum of allowed momentum states is broken up by this quantization. The allowed states are coalesced into the Landau levels, and each spin-degenerate Landau level can hold

$$n_{\mathrm{s}} = \frac{1}{\pi r_0^2} = \frac{eB}{\pi\hbar} \tag{4.121}$$

electrons. In some magneto-transport experiments, this quantization can be seen as oscillatory magnetoresistance, in which the periodicity in $1/B$ arises from (4.121) as

$$\Delta\left(\frac{1}{B}\right) = \frac{e}{\pi\hbar n_{\mathrm{s}}} \tag{4.122}$$

for spin-degenerate Landau levels. This oscillatory magnetoresistance is termed the Shubnikov–de Haas effect. We return to it again in the following discussion.

### 4.7.2 Adding lateral confinement

Let us now consider what happens when the magnetic field is added to a simple harmonic oscillator potential in the $x$-direction—that is, we confine the electron in the lateral $x$-direction and add an additional magnetic field in the $z$-direction. Hence, we combine the Hamiltonians of (4.9) and (4.111) to seek a solution of

$$-\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x^2} + \frac{2eBx}{\hbar}k - k^2 - \left(\frac{eB}{\hbar}x\right)^2\right]\psi + \frac{1}{2}m\omega^2 x^2\omega = E\psi. \tag{4.123}$$

It is clear that this can now be re-arranged to give

$$\begin{aligned}\left(E - \frac{\hbar^2 k^2}{2m}\right)\psi = & -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} - \hbar\omega_{\mathrm{c}}kx\psi + \frac{1}{2}m(\omega^2 + \omega_{\mathrm{c}}^2)x^2\psi \\ = & -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + \frac{1}{2}m\Omega^2(x - x_0)^2\psi - \frac{\hbar^2 k^2}{2m}\left(\frac{2\omega_{\mathrm{c}}}{\Omega}\right)^2\end{aligned} \tag{4.124}$$

where

$$\Omega^2 = \omega^2 + \omega_{\mathrm{c}}^2 \qquad x_0 = \frac{\hbar k\omega_{\mathrm{c}}}{m\Omega^2}. \tag{4.125}$$

The motion remains that of a harmonic oscillator, but now the shift of the wave function and the energy levels are hybrids of the confinement harmonic oscillator and the magnetic harmonic oscillator. Each energy level of the magnetic field, the Landau levels, is raised by the confinement to a higher value. In essence,

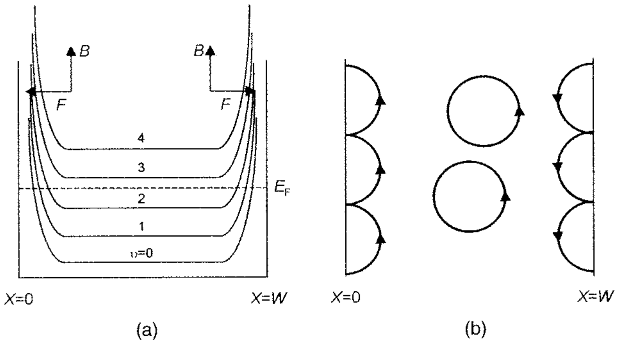


**Figure 4.6.** (*a*) The bending of the Landau levels at the edges of a confined sample. The electric field is shown for comparison at the two edges. (*b*) The confined and bouncing orbits for the situation of (*a*). The magnetic field is out of the page.

this coupling of the two harmonic oscillators leads to enhanced confinement, and stronger confinement always costs energy—the result is that the energy levels lie at higher values.

This becomes more important if the harmonic oscillator is nonlinear. That is, if the value of the parameter $\omega$ varies with distance from the centre of the harmonic oscillator, then the energy levels are further increased as they are found farther from the centre. This reaches the extreme in hard-wall boundary conditions at certain points on the $x$-axis. This is shown in figure 4.6(*a*). Here, the Landau levels rapidly increase in energy as they approach the hard walls due to the extra confinement, just as in (4.125). The importance of this is the fact that an electron whose orbit is within the orbit radius of the edge will strike the wall and create a bouncing orbit that moves along the wall, as shown in figure 4.6(*b*). Instead of being trapped in a Landau orbit, these bouncing orbits can carry current along the walls of the confining region, and are called *edge states*.

When the Fermi level lies in a Landau level, there are many states available for the electron to gain small amounts of energy from the applied field and therefore contribute to the conduction process. On the other hand, when the Fermi level is in the energy gap between two Landau levels, the upper Landau levels are empty and the lower Landau levels are full. Thus there are no available states for the electron to be accelerated into, and the conductivity drops to zero in two dimensions. In three dimensions it can be scattered into the direction parallel to the field (the $z$-direction), and this conductivity provides a positive background on which the oscillations ride. Figure 4.7 shows a typical measurement of the longitudinal resistance. The 'zeros' of the longitudinal resistance correspond to the magnetic field for which there are full Landau levels. ($R_{xx}$ has zeros as does the longitudinal conductance, since in the presence of the magnetic field, $R_{xx} = G_{xx}/(G_{xx}^2 + G_{xy}^2)$. Although the longitudinal conductance vanishes, the

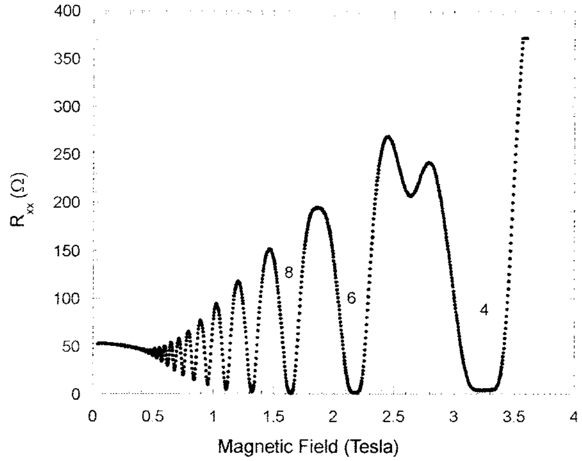


**Figure 4.7.** The longitudinal resistance for a quasi-two-dimensional electron gas in a high magnetic field. The oscillations are the Shubnikov–de Haas oscillations, and correspond to the sequential emptying of Landau levels. (Data courtesy of D P Pivin Jr, Arizona State University.)

transverse conductance does not, and this means that the longitudinal resistance also vanishes.) However, the index that is shown (4, 6, 8, etc) is that for spin-resolved levels, rather than spin-degenerate Landau levels. Hence, for the case of the zero at index 4 ($B = 3.25$ T, termed the $\nu = 4$ level), the $n = 0$ and $n = 1$ Landau levels (both doubly spin degenerate) are full. The zero corresponds to the transition of the Fermi energy between the 5/2 and 3/2 (in units of $\hbar\omega_c$) levels in figure 4.6. These measurements are for a quasi-two-dimensional electron gas at the interface of a GaAlAs/GaAs heterostructure. From (4.121), we can determine the areal density to be approximately $3.3 \times 10^{11}$ cm$^{-2}$.

### 4.7.3 The quantum Hall effect

The zeros of the conductivity that occur when the Fermi energy passes from one Landau level to the next-lowest level are quite enigmatic. They carry some interesting by-products. A full derivation of the quantum Hall effect is well beyond the level at which we are discussing the topic here. However, we can use a consistency argument to illustrate the quantization exactly, as well as to describe the effect we wish to observe. When the Fermi level lies between the Landau levels, the lower Landau levels are completely full. We may then say that

$$E_{\mathrm{F}} \approx \nu\hbar\omega_{\mathrm{c}} \tag{4.126}$$

where $\nu$ is an integer giving the number of *spin-resolved* Landau levels. The presence of the edge states means that some carriers have moved to the edge

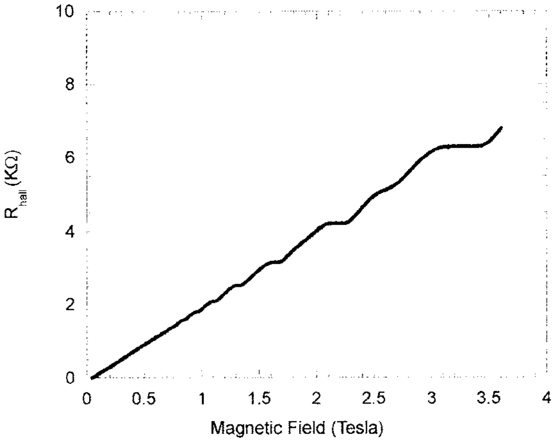


**Figure 4.8.** The Hall resistance of quantized Landau levels. The normalization is chosen so that the index value is of spin-resolved Landau levels. The first spin splitting is just being resolved at $\nu = 5$ ($\sim$2.6 T). (Data courtesy of D P Pivin Jr, Arizona State University.)

of the sample, and this arises from including the electric field in the second of equations (4.117) as

$$m\frac{\mathrm{d}\boldsymbol{v}_y}{\mathrm{d}t} = -eE_y + e\boldsymbol{v}_x B. \tag{4.127}$$

In a stable situation, the lateral acceleration must vanish at the edge, so that a transverse field must exist, given by

$$E_y = \boldsymbol{v}_x B = -\frac{J_x}{ne}B. \tag{4.128}$$

Using (4.121) for the density in each Landau level, we have (inserting a factor of two to raise the spin degeneracy)

$$n_\mathrm{s} = \nu\frac{eB}{2\pi\hbar}. \tag{4.129}$$

The density is constant in the material, so using (4.129) in (4.128) to define the *Hall resistance* gives

$$R_\mathrm{H} = -\frac{E_y}{J_x} = \frac{h}{\nu e^2}. \tag{4.130}$$

The quantity $h/e^2 = 25.81$ k$\Omega$ is a ratio of fundamental constants. Thus the conductance (reciprocal of the resistance) increases stepwise as the Fermi level passes from one Landau level to the next-lower level. Between the Landau levels, when the Fermi energy is in the localized state region, the Hall resistance is constant (to better than 1 part in $10^7$) at the quantized value given by (4.130) since the lower Landau levels are completely full. In figure 4.8, the variation of the Hall resistance as a function of magnetic field for a typical sample is shown. These measurements were made in the same sample and geometry of figure 4.7, so that they can be easily compared.

The magnetic field could, of course, be swept to higher values in both figures 4.7 and 4.8. When this is done, new features appear, and these are not explained by the above theory. Klaus von Klitzing received the Nobel prize for the discovery of the quantum Hall effect (von Klitzing *et al* 1980). In fact, in high quality samples, once the Fermi energy is in the lowest Landau level, one begins to see fractional filling and plateaus, in which the resistance is a fraction of $h/e^2$ (Tsui *et al* 1982). This *fractional quantum Hall effect* is theorized to arise from the condensation of the interacting electron system into a new many-body state characteristic of an incompressible fluid (Laughlin 1983). Tsui, Störmer, and Laughlin shared the Nobel prize for this discovery. However, the properties of this many-body ground state are clearly beyond the present level.

## References

Ando T, Fowler A B and Stern F 1982 *Rev. Mod. Phys.* **54** 437

Born M and Huang K 1954 *Dynamical Theory of Crystal Lattices* (London: Oxford University Press)

Ferry D K 1991 *Semiconductors* (New York: Macmillan)

Ferry D K and Goodnick S M 1997 *Transport in Nanostructures* (Cambridge: Cambridge University Press)

Hall J M 1967 *Phys. Rev.* **161** 756

Kroemer H 1994 *Quantum Mechanics* (Englewood Cliffs, NJ: Prentice-Hall)

Landau L 1930 *Z. Phys.* **64** 629

Landau L and Lifshitz E M 1958 *Quantum Mechanics* (London: Pergamon)

Laughlin R B 1983 *Phys. Rev. Lett.* **50** 1395

Merzbacher E 1970 *Quantum Mechanics* (New York: Wiley)

Schiff L I 1955 *Quantum Mechanics* 2nd edn (New York: McGraw-Hill)

Smith R A 1969 *Wave Mechanics of Crystalline Solids* (London: Chapman and Hall)

Tsui D, Störmer H L and Gossard A C 1982 *Phys. Rev. Lett.* **48** 1559

von Klitzing K, Dorda G and Pepper M 1980 *Phys. Rev. Lett.* **45** 494

von Neumann J 1955 *Mathematical Foundations of Quantum Mechanics* (Princeton, NJ: Princeton University Press)

Ziman J 1964 *Principles of the Theory of Solids* (Cambridge: Cambridge University Press)

## Problems

1. Using the WKB formula for the bound states of a potential well (3.71), compute the energy levels of a harmonic oscillator, whose potential is described by (4.2).

2. Consider a simple, classical harmonic oscillator, whose amplitude of oscillation is $\xi$. Show that this amplitude is related to the energy of the oscillator through the formula

$$\xi = \sqrt{\frac{2\mathcal{E}}{m\omega^2}}\sin(\omega t + \beta)$$

where $\beta$ is an arbitrary constant, which is unknown. If we consider that the probability of finding the particle in the small increment $\mathrm{d}\xi$ is just the fraction of time spent in this interval during each period of oscillation, show that the probability of finding the mass in a small region $\mathrm{d}\xi$ about $\xi$ is 0 for $|\xi| \geq \sqrt{2\mathcal{E}/m\omega^2}$ and $[\pi\sqrt{2\mathcal{E}/m\omega^2 - \xi^2}]^{-1}$ for $|\xi| \leq \sqrt{2\mathcal{E}/m\omega^2}$.

3. According to the previous problem, there is no chance classically for a particle of energy $5\hbar\omega/2$ to be in the region $|\xi| > \sqrt{5\hbar/m\omega}$. What is the probability for a particle of this energy to be in this region quantum mechanically?

4. Using a series representation for the general Hermite polynomial of order $n$, which has been properly normalized, verify that (4.27) is valid. The series can be developed from the coefficients in (4.22).

5. Using the generating function for the Hermite polynomials, compute $\langle p^2\rangle$ and $\langle x^2\rangle$. Can these results be obtained from a knowledge of the expectations $\langle p\rangle$ and $\langle x\rangle$? What is the uncertainty in position and momentum?

6. Using the creation and annihilation operators, compute $\langle p^2\rangle$ and $\langle x^2\rangle$. How does the complexity of this approach compare with that in the previous problem?

7. Determine the expectation values for the kinetic and potential energies in a harmonic oscillator.

8. Consider an electromagnetic resonator with a resonant frequency of $10^{10}$ Hz. What is the energy separation of the oscillator levels? How does this compare with the thermal energy fluctuation?

9. If the velocity of sound $\mathrm{d}\omega/\mathrm{d}q$ of a set of lattice vibrations is $10^5$ cm s$^{-1}$, and the lattice constant is $a = 0.25$ nm, what is the phonon frequency at the zone edge $q = \pi/a$?

10. If one begins with the so-called symmetric gauge, where $\boldsymbol{A} = (eB/2)(-y\boldsymbol{a}_x + x\boldsymbol{a}_y)$, show that a more complicated harmonic oscillator solution results. Show that this still gives energy levels according to (4.113), and find the appropriate form of the wave functions that are the solutions.

11. Consider an electron moving in a $z$-directed magnetic field and constrained in a quadratic potential $\frac{1}{2}m\omega_0^2x^2$. In the simplest case, in which the system is homogeneous in the $y$-direction, determine the energy levels of the electron.

12. A semiconductor sample measures 1 cm by 0.5 cm and is 0.1 cm thick. For an applied electric field of 1 V $\mathrm{cm}^{-1}$, 5 mA of current flows. If a 0.5 T magnetic field is applied normal to the broad surface, a Hall voltage of 5 mV is developed. Determine the Hall mobility and the carrier density.

13. Consider a free-electron 'gas' with an areal density of $2 \times 10^{12}\ \mathrm{cm}^{-2}$. What is the periodicity (in units of $1/B$) expected for the Shubnikov–de Haas oscillation?

# Chapter 5

# Basis functions, operators, and quantum dynamics

In the past few chapters, we have developed the Schrödinger equation and applied it to the solution of a number of quantum mechanical problems, mainly to develop experience with the results of simple and common systems. The results from these examples can be summarized relatively simply. In general, quantization enters the problem through the non-commuting nature of conjugate operators such as position and momentum. This has led to the Schrödinger equation itself as the primary equation of motion for the wave function solution to the problem. In essence, the system (e.g. the electron) is treated as a wave, rather than as a particle, and the wave equation of interest is the Schrödinger equation. When boundary conditions are applied, either through potential barriers, or through the form of the potential (as in the harmonic oscillator), the time-independent Schrödinger equation yields solutions that are often special functions. Examples of this are the sines and cosines in the rectangular-barrier case, the Airy functions in the triangular-well case, and the Hermite polynomials in the harmonic oscillator case.

It is generally true that any time we examine a bounded system (even a classical system), the allowed energy levels take on discrete values, each of which corresponds to a single one of the family of possible members of the special functions. Thus, in the rectangular-well case, each energy level corresponds to one of the sinusoidal harmonics; in the harmonic oscillator, each energy level corresponds to one of the Hermite polynomials. In the set of levels, there is always a lowest energy level, called the ground state. The higher energy levels are referred to as the excited states, even when some are occupied in thermal equilibrium.

It should not be surprising that the set of all possible solutions to a given problem formulation, for example the set of all sines and cosines for the rectangular barriers, can be shown to form a complete set, and thus can serve as a *basis* for a many-dimensional *linear-vector-space* representation of the problem. In the preceding chapters, we have not employed this terminology, but rather treated the expansion in terms of these functions in a manner that had the time variation appearing directly within the function itself. This is the normal Schrödinger *picture*, in which the so-called basis functions arise from solving the Schrödinger equation subject to the boundary conditions and these basis functions contain any time variation. In this picture, the operators are not time varying, but their projection onto any of the basis functions is time varying. Here, the basis functions can be thought of as the unit vectors of a coordinate system, and the amplitudes such as those of (2.87) can be thought of as amplitudes along each axis of this coordinate system. In the *Schrödinger picture*, the coordinate system rotates around the direction of the vector operator as a function of time. But this is a relative view—it could as easily be that the coordinate system is fixed and the vector operator rotates in the coordinate frame in such a way that the projection on any axis still varies with time in the prescribed manner. This picture, or view, is referred to as the *Heisenberg picture* of quantum mechanics (Heisenberg 1925).

We can think about this connection between the Heisenberg picture and the Schrödinger picture in a relatively simple manner. We have found that the wave function $\Psi(x,t)$ evolves under the action of the Hamiltonian operator via the time-dependent Schrödinger equation (2.8)

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = H\Psi = \left(-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)\right)\Psi. \tag{5.1}$$

In a sense, this wave function evolves as described in (2.93), or

$$\Psi(x,t) = \mathrm{e}^{-\mathrm{i}Ht/\hbar}\Psi(x,0). \tag{5.2}$$

If we consider any arbitrary complete set of functions $\{\phi_n(x)\}$, such as the set of basis states for an infinite potential well lying in the range $0 < x < a$

$$\phi_n(x) = \sqrt{\frac{2}{a}}\sin\left(\frac{n\pi x}{a}\right) \tag{5.3}$$

we can expand the wave function as

$$\Psi(x,t) = \sum_n c_n(t)\phi_n(x). \tag{5.4}$$

The basis functions $\phi_n$ create a infinite set of coordinates, and the coefficients $c_n(t)$ are the projections of the wave function onto each of these coordinates. Hence, the basis states create an infinite-dimensional coordinate system, which is the Hilbert space corresponding to these functions.

The rotation of the total wave function (5.2) in this new coordinate space, described by the basis states, can be found by inserting (5.4) into (5.2). This gives

$$\begin{aligned}\Psi(x,t) &= \mathrm{e}^{-\mathrm{i}Ht/\hbar}\sum_n c_n(0)\phi_n(x) = \sum_n c_n(0)\mathrm{e}^{-\mathrm{i}Ht/\hbar}\phi_n(x)\\ &= \sum_n c_n(0)\mathrm{e}^{-\mathrm{i}E_n t/\hbar}\phi_n(x) = \sum_n c_n(0)\mathrm{e}^{-\mathrm{i}\omega_n t}\phi_n(x).\end{aligned} \tag{5.5}$$

In this last form, the frequency $\omega_n = E_n/\hbar$ has been introduced, and $E_n$ is the energy corresponding to each of the time-independent basis functions. The result is that

$$c_n(t) = c_n(0)\mathrm{e}^{-\mathrm{i}\omega_n t}. \tag{5.6}$$

The difference between the Heisenberg picture and the Schrödinger picture is one of just where to assign the exponential factor. If we take the view that the basis states are fixed and time independent, as indicated in (5.4), then the coordinate axes are fixed and the wave function rotates in this fixed coordinate system with the time-varying projections $c_n(t)$. This is the Heisenberg picture. On the other hand, if we attach the time variation to the basis states, and allow them to vary with time as

$$\phi_n(x,t) = \mathrm{e}^{-\mathrm{i}\omega_n t}\phi_n(x,0) \tag{5.7}$$

and take the various $c_n$ as time independent (and assign their projections at $t = 0$), then the coordinates 'rotate' around the wave function. This is the Schrödinger picture. In fact, we have used the latter in the previous chapters, although we will increasingly use the former in subsequent chapters. The difference is only one of detail.

In this chapter, we want to achieve a number of things, not the least of which is a review of many of the concepts introduced in chapter 1, but at the same time put in some of the details and mathematical rigour omitted in the earlier treatment. However, we want to go further. For example, in sections 3.8 and 4.5, we introduced pseudo-position and pseudo-momentum operators in order to quantize some systems. In essence, this was the introduction of generalized conjugate operators, and we want to examine here the rules that allow us to do this. Finally, we want to formalize the connections between the Schrödinger and Heisenberg pictures, and the operator equations that govern the appropriate dynamics. For example, in the Schrödinger picture, we solved the Schrödinger equation for the wave functions themselves. In the Heisenberg picture, however, the basis functions will be specified by the boundary conditions and coordinates, and the time evolution of the operators themselves must be sought. This will allow us to introduce some general postulates of quantum mechanics that relate to the connection between classical dynamics and the evolution of the operators describing quantum dynamics.

## 5.1 Position and momentum representation

Already, in section 1.4, we have talked about the position and the momentum of a wave function, or of a wave packet. In this earlier discussion, it was pointed out that one normally works with a wave function that is a function of position and time only; or one works with the Fourier-transformed quantity, which is a function of momentum and time alone. This led to the recognition that we could write down the expectation values of the position and momentum as

$$\langle x\rangle = (\Psi, x\Psi) = \int \Psi^*(x,t)x\Psi(x,t)\,\mathrm{d}x \tag{5.8}$$

$$\begin{aligned}\langle p\rangle = (\Psi, p\Psi) &= \int \Psi^*(x,t)\left[-\mathrm{i}\hbar\frac{\partial}{\partial x}\Psi(x,t)\right]\mathrm{d}x \\ &= \int \Phi^*(p,t)p\Phi(p,t)\,\mathrm{d}p.\end{aligned} \tag{5.9}$$

Here, we have used both the differential form of the momentum operator and its parameter form in the momentum representation. This symbolic relationship between the momentum as a simple operator in the momentum representation, and as a differential operator in the position representation, carries through equally to its conjugate variable—the position. Let us take (5.8) and introduce the Fourier transforms of the wave functions from (1.26), as

$$\begin{aligned}\langle x\rangle &= \int \Psi^*(x,t)x\Psi(x,t)\,\mathrm{d}x \\ &= \frac{1}{2\pi}\int \mathrm{d}x \int \phi^*(k',t)\mathrm{e}^{-\mathrm{i}k'x}\,\mathrm{d}k'x\int \phi(k,t)\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k \\ &= -\frac{1}{2\mathrm{i}\pi}\int \mathrm{d}k'\int \phi^*(k',t)\,\mathrm{d}k\frac{\partial\phi(k,t)}{\partial k}\int \mathrm{e}^{\mathrm{i}kx-\mathrm{i}k'x}\,\mathrm{d}x \\ &= \mathrm{i}\hbar\int \phi^*(k',t)\frac{\partial}{\partial p}\phi(k,t)\,\mathrm{d}k\end{aligned} \tag{5.10}$$

from which we may immediately recognize the position operator *in the momentum representation* as

$$x = \mathrm{i}\hbar\frac{\partial}{\partial p}. \tag{5.11}$$

Thus, we have a full symmetry between the two operators—position and momentum—which form the conjugate pair. In the position representation, the momentum becomes a differential operator. In the momentum representation, the position becomes a differential operator. This behaviour carries over to any function of the position and momentum. Any normal function can be expanded as a Taylor series, and term by term the relationships used in (5.10) and (1.29) can be invoked to lead to the relationship that in the position representation

$$F(x,p) \to F\left(x, -\mathrm{i}\hbar\frac{\partial}{\partial x}\right) \tag{5.12}$$

and in the momentum representation

$$F(x,p) \to F\left(\mathrm{i}\hbar\frac{\partial}{\partial p}, p\right). \tag{5.13}$$

The appropriate averages can be developed by using these functional relationships.

Let us now finish this short review by considering what the Schrödinger equation looks like in the momentum representation. We begin by forming the partial derivative, with respect to time, of the inverse Fourier transform of (1.26), as

$$\begin{aligned}
\mathrm{i}\hbar\frac{\partial\phi}{\partial t} &= \frac{\mathrm{i}\hbar}{\sqrt{2\pi}}\int\frac{\partial\Psi}{\partial t}\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}}\int\mathrm{e}^{-\mathrm{i}kx}\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}+V(x)\right]\Psi(x,t)\,\mathrm{d}x \\
&= \frac{1}{2\pi}\int\mathrm{e}^{-\mathrm{i}kx}\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}+V(x)\right]\mathrm{d}x\int\phi(k',t)\mathrm{e}^{\mathrm{i}k'x}\,\mathrm{d}k' \\
&= \left[\frac{p^2}{2m}+V\left(\mathrm{i}\hbar\frac{\partial}{\partial p}\right)\right]\phi(k,t). \qquad (5.14)
\end{aligned}$$

This latter form is the Schrödinger equation in the momentum representation. It is clear that we have used the operator relationships as developed previously, but brute force in the third line of (5.14) produces the same results if a Taylor series expansion for the potential energy is introduced.

The equivalent relationships hold for any pair of conjugate operators. We can choose a representation that diagonalizes one of the pair. Then the other appears as a differential operator in that representation. This is the essential point. In the position representation, we can find the position from the wave function, but the momentum appears as a differential operator. On the other hand, in the momentum representation, the latter can be an eigenvalue, but the position will appear as a differential operator. Thus, we are confronted with a pair of conjugate operators that are non-commuting operators. We will see that this non-commutativity is one of the fundamental postulates of quantum mechanics. In general, we can find a representation in which one of the operators is diagonal, but the second appears as a differential operator. This result is in essence the introduction of the quantum effects that arise from the non-commutativity of the pair of operators. In the next section, we will show how the uncertainty relationship arises from these properties.

## 5.2 Some operator properties

The previous section reinforced the understanding of conjugate operators such as position and momentum. However, it is not always clear that the desired operator is easily developed in terms of a representation in which the various terms in the expansion for the wave function yield simple eigenvalues. In fact, we generally do not know just which basis (the set of basis functions with which to define a generalized coordinate system) to choose to evaluate the expectation value of one or a group of operators. In a few cases, the Schrödinger equation appears in a sufficiently simple form that we can recognize the appropriate special functions. In general, however, there are very few cases in which the basis functions are easily determined from the Schrödinger equation. Thus, we need to understand the dynamics that occurs for an operator in a *general basis set*. This is the aim of this, and the next, section. We first want to show that it is easy to change from the Schrödinger to the Heisenberg representation, and then to look at the properties of Hermitian operators and commutators. We begin with the first of these tasks.

### 5.2.1 Time-varying expectations

In (5.8) and (5.9), we developed the expectation value of the position and the momentum operators in their respective representations. We can do this quite generally for any operator, and the expectation value of such a general operator $A$ is given by the generalization of the above as

$$\langle A\rangle = (\Psi, A\Psi) = \int \Psi^*(x,t)A\Psi(x,t)\,\mathrm{d}x. \tag{5.15}$$

We do not write the dependence of the general operator $A$ on position and momentum, but in general it is a function of these generalized conjugate variables. Thus, $A = A(p,x)$ is some function of the position and of the momentum. An example of this is the quadratic harmonic oscillator potential energy, in which we can write $A = m\omega^2x^2/2$, so the expectation value (5.15) is the average of the potential energy.

How does the average defined in (5.15) change with time in a system in which the wave function $\Psi(x,t)$ evolves in time? We explored this briefly in section 2.8.1, where we discussed the Ehrenfest theorem, but there we were specifically interested in the relationship with the accelerative force. Here, we want to develop a general approach for an operator. We will, however, assume that the operator $A$ is *not a specific function of time itself* (time will not be one of the explicit variables upon which it depends). Thus, we can develop the time variation of (5.15) as

$$\begin{aligned}\frac{\mathrm{d}\langle A\rangle}{\mathrm{d}t} &= \int \frac{\partial\Psi^*(x,t)}{\partial t}A\Psi(x,t)\,\mathrm{d}x + \int \Psi^*(x,t)A\frac{\partial\Psi(x,t)}{\partial t}\,\mathrm{d}x\\ &= \frac{\mathrm{i}}{\hbar}\int H^*\Psi^*(x,t)A\Psi(x,t)\,\mathrm{d}x - \frac{\mathrm{i}}{\hbar}\int \Psi^*(x,t)AH\Psi(x,t)\,\mathrm{d}x\\ &= \frac{\mathrm{i}}{\hbar}\int \Psi^*(x,t)[HA - AH]\Psi(x,t)\,\mathrm{d}x\\ &= \frac{\mathrm{i}}{\hbar}\langle[H,A]\rangle. \end{aligned}\tag{5.16}$$

This result is very important, and is the basis of the Heisenberg approach to quantum mechanics. Previously, we needed to solve directly for the wave function and then develop the expectation value of specific operators. Here, however, *we may choose to construct the wave function with any convenient basis set*. The time variation of the expectation values of the operators is entirely determined by the commutator relation between the specific operator and the total Hamiltonian. *Thus, the expectation value of any operator that commutes with the Hamiltonian remains constant in time*. This is because in the conservative (non-dissipative systems) that we have treated so far, the Hamiltonian is not a function of time and the total energy is conserved. Therefore, any operator that commutes with the Hamiltonian is similarly conserved, and can also be simultaneously measured.

In treating conjugate operators, such as position and momentum, it was not generally possible (with a potential term present) to write the Hamiltonian as a function of only position or of only momentum. In general, the Hamiltonian contains both position (in the potential energy term) and momentum (in the kinetic energy term). Thus, we cannot choose a basis in which both commute with the Hamiltonian, so we cannot measure both simultaneously. Hence, non-commuting operators generally appear in a manner such that they do not commute with the Hamiltonian.

Equation (5.16) leads us to the general view that we can choose a basis set of functions, in terms of which to expand the total wave function, that is *convenient*. Thus, we would choose a set that is the natural set of expansion functions for the coordinate system and boundary conditions present. The choice is only important in that we must be able to evaluate the integrals inherent in the averages expressed in (5.16), and as expansions of the wave functions in for example (2.93), to which we return in the next section. So far, we have disregarded the time variation in the wave functions, although (5.16) begins to transfer this to the operator itself. Let us continue this process.

The time-varying Schrödinger equation allows us to write directly the time variation through a simple exponential term, as was done in section 2.3. Thus, we may write the total wave function in the general form

$$\Psi(x,t) = \mathrm{e}^{-\mathrm{i}\omega t}\Psi(x) \qquad \omega = H/\hbar. \tag{5.17}$$

Now, we know that the energy arises in the argument of the exponential as the eigenvalue for the Hamiltonian (which must be summed over in most cases), so when we introduce (5.17) into the averages in (5.16), we will retain use of the operator form; for example, we will introduce the formal solution of the time-varying Schrödinger equation. Following (5.16), this leads to

$$\begin{aligned}\frac{\mathrm{d}\langle A\rangle}{\mathrm{d}t} &= \frac{\mathrm{i}}{\hbar}\int \Psi^*(x,t)[HA - AH]\Psi(x,t)\,\mathrm{d}x \\ &= \frac{\mathrm{i}}{\hbar}\int \left[\exp\left[-\frac{\mathrm{i}}{\hbar}Ht\right]\Psi(x)\right]^*[HA - AH]\exp\left[-\frac{\mathrm{i}}{\hbar}Ht\right]\Psi(x)\,\mathrm{d}x \\ &= \frac{\mathrm{i}}{\hbar}\int \Psi^*(x)\mathrm{e}^{\mathrm{i}Ht/\hbar}[HA - AH]\mathrm{e}^{-\mathrm{i}Ht/\hbar}\Psi(x)\,\mathrm{d}x.\end{aligned} \tag{5.18}$$

Now, consider the general operator $C$ defined in terms of two operators $A$ and $B$ by the relationship

$$C = \mathrm{e}^{At} B \mathrm{e}^{-At}. \tag{5.19}$$

The time derivative of this operator is given by

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \mathrm{e}^{At} AB \mathrm{e}^{-At} - \mathrm{e}^{At} BA \mathrm{e}^{-At} = \mathrm{e}^{At}[A, B]\mathrm{e}^{-At}. \tag{5.20}$$

We note in passing that, since $A$ and $B$ might well be non-commuting operators, *the order of the terms in (5.20) is quite important and must be preserved.* However, if we relate $C$ and $B$ both to the operator $A$ in (5.18), and the factor $A$ with $\mathrm{i}H/\hbar$, we can rewrite (5.18) as

$$\frac{\mathrm{d}\langle A\rangle}{\mathrm{d}t} = \int \Psi^*(x) \frac{\mathrm{d}A(t)}{\mathrm{d}t} \Psi(x)\, \mathrm{d}x \tag{5.21}$$

provided that $A$ *is not an explicit function of time*, and the so-called *Heisenberg representation* of $A$ is given by

$$A_{\mathrm{H}}(t) = \mathrm{e}^{\mathrm{i}Ht/\hbar} A \mathrm{e}^{-\mathrm{i}Ht/\hbar}. \tag{5.22}$$

The subscript 'H' is often used to denote the Heisenberg representation of an operator, which varies with time only through the effects that arise from the non-commuting of this operator with the Hamiltonian. However, this subscript is just as often not explicitly given.

From (5.21), we can work backwards toward our starting equation (5.15) by dropping the time derivatives, so we can write the expectation of an operator in the Heisenberg picture as

$$\langle A\rangle = \int \Psi^*(x) A_{\mathrm{H}}(t) \Psi(x)\, \mathrm{d}x. \tag{5.23}$$

This is the basis for the Heisenberg picture (or representation) of quantum mechanics. Here, the wave function is expanded in a convenient basis set of functions, which corresponds to a fixed coordinate system. Then, instead of the wave function evolving or varying with time, the operators themselves evolve in time (if they do not commute with the Hamiltonian). The averages are now computed using the arbitrary, fixed basis functions and the 'time-varying' Heisenberg representation of the operators. In essence, if we expand the wave function in terms of the basis set, the above average is a matrix multiplication, which we will deal with below.

### 5.2.2 Hermitian operators

We recall from our earlier discussions in section 2.2 that $A^+$ is the Hermitian adjoint of the operator $A$. If $A$ were represented by a matrix, then the Hermitian adjoint would correspond to the transpose complex conjugate of that matrix. Here, we want to emphasize that all measurable quantities correspond to operators in which the Hermitian conjugate is equal to the operator itself; for example, the operator is said to be a Hermitian operator. We will discuss some of the properties of these operators. We first note that the average value of the operator $A$ is given by (5.15). If the expectation value is a measurable quantity, then this average value must be a real quantity. With this in mind, we rewrite (5.15) as

$$\begin{aligned}\langle A\rangle = \langle A\rangle^* &= \left(\int \Psi^*(x,t)A\Psi(x,t)\,\mathrm{d}x\right)^* \\ &= \int (A\Psi(x,t))^*\Psi(x,t)\,\mathrm{d}x \\ &= \int \Psi^*(x,t)A^+\Psi(x,t)\,\mathrm{d}x. \end{aligned} \tag{5.24}$$

By comparison with (5.15), it is now obvious that for the expectation value to be a real quantity we must have $A = A^+$. As mentioned, operators for which this relation holds are said to be Hermitian operators. The step from the first line to the second has used the fact that $(AB)^* = (B^*A^*)$, and $(AB)^+ = (B^+A^+)$. These are general properties of the complex conjugate and adjoint operations. We can demonstrate this with the momentum differential operator using the second line of (5.24):

$$\begin{aligned}\langle \boldsymbol{p}\rangle &= \int (-\mathrm{i}\hbar\boldsymbol{\nabla}\Psi(x,t))^*\Psi(x,t)\,\mathrm{d}x \\ &= \mathrm{i}\hbar\int (\boldsymbol{\nabla}\Psi^*(x,t))\Psi(x,t)\,\mathrm{d}x \\ &= \mathrm{i}\hbar\int \boldsymbol{\nabla}(\Psi^*(x,t)\Psi(x,t))\,\mathrm{d}x - \mathrm{i}\hbar\int \Psi^*(x,t)(\boldsymbol{\nabla}\Psi(x,t))\,\mathrm{d}x \\ &= \int \Psi^*(x,t)(-\mathrm{i}\hbar\boldsymbol{\nabla}\Psi(x,t))\,\mathrm{d}x \end{aligned} \tag{5.25}$$

which is the normal result. Thus, Hermitian operators correspond to dynamical variables for which the expectation values are real quantities.

Let us now consider the situation in which the wave function is described in such a manner that the operator $A$ produces the eigenvalue $a$, where $a$ is a $c$-number (not an operator, but a scalar constant). Thus, we have the operator relation

$$A\Psi(x,t) = a\Psi(x,t). \tag{5.26}$$

Then, the average value of this operator is given again by (5.15) as

$$\begin{aligned}\langle A\rangle &= \int \Psi^*(x,t)A\Psi(x,t)\,\mathrm{d}x \\ &= \int \Psi^*(x,t)a\Psi(x,t)\,\mathrm{d}x = a. \end{aligned} \tag{5.27}$$

By the same token, we can compute the expectation value of the square of the operator

$$\begin{aligned}\langle A^2\rangle &= \int \Psi^*(x,t)A^2\Psi(x,t)\,\mathrm{d}x \\ &= \int \Psi^*(x,t)Aa\Psi(x,t)\,\mathrm{d}x = a^2. \qquad (5.28)\end{aligned}$$

This leads to the interesting result that the operator, for which the wave function leads to an exact eigenvalue, exhibits no variance, as $\langle A^2\rangle = \langle A\rangle^2$. Thus, the uncertainty of measuring the expectation of this operator, with this wave function, is identically zero. Now suppose that we have two such wave functions for which the operator $A$ produces eigenvalues, and so

$$A\Psi_1(x,t) = a_1\Psi_1(x,t) \qquad A\Psi_2(x,t) = a_2\Psi_2(x,t). \qquad (5.29)$$

Let us post-multiply the first of these equations by the complex conjugate of $\Psi_2$, take the complex conjugate of the second and post-multiply by $\Psi_1$, then subtract the two and integrate the result. This leads to

$$\begin{aligned}(a_1 - a_2^*)&\int \Psi_2^*(x,t)\Psi_1(x,t)\,\mathrm{d}x \\ &= \int \Psi_2^*(x,t)A\Psi_1(x,t)\,\mathrm{d}x - \int A^*\Psi_2^*(x,t)\Psi_1(x,t)\,\mathrm{d}x \\ &= \int \Psi_2^*(x,t)A\Psi_1(x,t)\,\mathrm{d}x - \int \Psi_2^*(x,t)A\Psi_1(x,t)\,\mathrm{d}x = 0. \quad (5.30)\end{aligned}$$

We thus have two options for the integral on the left. The first option is that $\Psi_2 = \Psi_1$, which leads to $a_1 = a_2^* = a_1^*$. That is, if the two wave functions are the same, then the eigenvalue must be the same, and it must be real! The second option is that the eigenvalues actually are different, for which we must have

$$\int \Psi_2^*(x,t)\Psi_1(x,t)\,\mathrm{d}x = 0 \qquad (5.31)$$

which means that the wave functions are orthogonal to each other.

Thus, if an operator gives different eigenvalues when operating on different wave functions, these wave functions must be mutually orthogonal. It is this latter point that allows us to set up a 'coordinate' system with basis functions that are orthonormal (orthogonal and normalized). This is the purpose of the linear vector space approach. We can choose a set of orthogonalized (and normalized) wave functions $\psi_i(x)$ (we write in now only the spatial variation, as we will pursue the Heisenberg picture). These can correspond to the infinite set of wave functions found in the potential well problems of chapter 2, or the set of Hermite polynomials of chapter 4, as examples. Then, we expect that any operator $A$ will give a set of eigenvalues $a_i$ corresponding to each member of the set of wave functions. To illustrate this, let us expand the total wave function in the set of basis functions, as

$$\Psi(x) = \sum_i c_i \psi_i(x). \tag{5.32}$$

With this expansion, we can write the expectation value of the operator $A$ as

$$\begin{aligned}\langle A \rangle &= \int \sum_i c_i^* \psi_i^*(x) \sum_j A c_j \psi_j(x)\, \mathrm{d}x \\ &= \int \sum_i c_i^* \psi_i^*(x) \sum_j a_j c_j \psi_j(x)\, \mathrm{d}x \\ &= \sum_i c_i^* \sum_j a_j c_j \delta_{ji} = \sum_i a_i |c_i|^2. \end{aligned} \tag{5.33}$$

In going from the second line to the third, we have used the fact that the integral is zero (orthogonality) unless $i = j$. This provides the Kronecker delta function in the third line, which is then used in the summation over the index $j$. Thus, the coefficients provide the weight, or probability amplitude, for each of the individual eigenvalues of the basis set. The set of eigenvalues $a_i$ is termed the *spectrum* of the operator $A$. (No comment has been made about all of the $a_i$ being unique. This is not required, but can be achieved by techniques discussed below.) The connection with a normal spectrum is quite straightforward: in the normal case, the strength of a component at frequency $\omega$ is given by the corresponding Fourier coefficient. Here, the eigenvalue $a_i$ is the generalized Fourier coefficient giving the strength of the operator $A$ in the particular basis coordinate $\psi_i$. The coefficients $c_i$ give a normalization to the excitation of the particular mode $\psi_i$ by the total wave function $\Psi$.

If the operator is taken to be the unit scalar (non-operator) quantity, then each eigenvalue is also unity, and we are led to

$$\int \Psi^* \Psi \, \mathrm{d}x = \sum_i |c_i|^2 = 1. \tag{5.34}$$

If we know that the system is in one particular basis state, then we can say that $c_k = 1$, and all other $c_i = 0$ $(i \neq k)$. This is said to be a *pure state*. All other cases are taken to be the so-called mixed states.

The individual eigenvalues $a_i$ of the operator $A$ are taken to be the possible outcomes of measurements based upon the dynamic variable corresponding to this operator. The magnitude-squared terms $|c_i|^2$ are the individual probabilities corresponding to each member of the selected basis set of wave functions. In general, any function of the operator $A$ can be expanded in a Taylor series and evaluated term by term to lead to the form

$$\langle f(A) \rangle = f(a) = \sum_i f(a_i) |c_i|^2. \tag{5.35}$$

In many cases, the spectrum of the operator is a continuous function, such as the momentum in plane waves where the waves are unbounded. In general, the total wave function can be expanded in a sum over the discrete states plus an integral over the continuous states. The principles of orthonormality continue to be upheld. Thus, plane waves with different values of the momentum are orthogonal to each other and to any discrete (bounded) basis states.

We close this section by illustrating one final property of the basis states—the property of *closure*, which we used in the preceding chapters. The total wave function can easily be written as in (5.32) as an expansion in some arbitrary basis set of functions. These functions can be chosen for convenience, but should be a complete orthonormal set of such basis functions. Then, for a given wave function $\Psi(x)$, we can evaluate the coefficients as

$$c_i = \int \psi_i^*(x)\Psi(x)\,\mathrm{d}x. \tag{5.36}$$

If we re-insert this result in (5.32), we obtain

$$\begin{aligned}\Psi(x) &= \sum_i \psi_i(x) \int \psi_i^*(x')\Psi(x')\,\mathrm{d}x' \\ &= \int \sum_i \psi_i(x)\psi_i^*(x')\Psi(x')\,\mathrm{d}x'.\end{aligned} \tag{5.37}$$

If the right-hand side is to yield the left-hand side as it should, since we are making a completely circular argument, then we must have

$$\sum_i \psi_i(x)\psi_i^*(x') = \delta(x - x'). \tag{5.38}$$

This is the principle of closure: the summation over the complete set of the orthonormal basis states shown is zero unless the product of basis states is evaluated at the same position in space. This is another representation of the delta function itself in a discrete linear vector space.

### 5.2.3 On commutation relations

In the above, it was stated that operators that are simultaneously measurable must commute. We want now to show that this is indeed the case. Suppose that we have two operators, $A$ and $B$, for which the wave function $\Psi(x)$ is an eigenfunction; for example, both operators produce simple eigenvalues with this wave function:

$$A\Psi(x) = a\Psi(x) \qquad B\Psi(x) = b\Psi(x) \tag{5.39}$$

which implies that $\langle A\rangle = a$, $\langle B\rangle = b$. Then, we can form the product operations

$$\begin{aligned}BA\Psi(x) &= Ba\Psi(x) = aB\Psi(x) = ab\Psi(x) \\ AB\Psi(x) &= aB\Psi(x) = ab\Psi(x)\end{aligned} \tag{5.40}$$

which immediately shows that

$$[A, B] = AB - BA = 0. \tag{5.41}$$

In producing an eigenvalue according to (5.39), *both operators leave the wave function unchanged*. It is this property, of not changing the wave function, that is the important aspect of eigenvalues. Normally, an operator will take a wave function and change the functional form. This is not the case in the eigenvalue equation (5.39), which means that when the average is taken, the two wave functions maintain their normalization. The result (5.41) means that if we have two operators that both have this eigenvalue property of leaving the wave function unchanged, then the two operators by necessity must commute. This, in turn, implies that we can simultaneously determine their expectation values, and hence measure these average values. If the operators do not commute, they cannot be simultaneously measured. This is the essence of the uncertainty principle (Heisenberg 1927).

We now want to close this section by showing that when two operators do not commute, they satisfy the uncertainty relation determined by the value of the commutator relation itself. Consider two operators $A$ and $B$ that satisfy the commutator relation

$$AB - BA \equiv [A, B] = \mathrm{i}C. \tag{5.42}$$

The uncertainty in measuring the value of the operator $A$ is given by its variance which is found from (the variance is a $c$-number by construction)

$$\begin{aligned}(\Delta A)^2 &= \int \Psi^*(x)(A - \langle A\rangle)^2 \Psi(x)\,\mathrm{d}x \\ &= \int (A - \langle A\rangle)^* \Psi^*(x)(A - \langle A\rangle)\Psi(x)\,\mathrm{d}x \\ &= \int |(A - \langle A\rangle)\Psi(x)|^2\,\mathrm{d}x.\end{aligned} \tag{5.43}$$

Similarly, the uncertainty in measuring the expectation of $B$ is given by

$$(\Delta B)^2 = \int |(B - \langle B\rangle)\Psi(x)|^2\,\mathrm{d}x. \tag{5.44}$$

We can now invoke the Schwarz inequality

$$\left(\int |f|^2\,\mathrm{d}x\right)\left(\int |g|^2\,\mathrm{d}x\right) \geq \left|\int f^* g\,\mathrm{d}x\right|^2. \tag{5.45}$$

We now make the associations

$$f = (A - \langle A\rangle)\Psi(x) \qquad g = (B - \langle B\rangle)\Psi(x) \tag{5.46}$$

so the Schwarz inequality leads to

$$(\Delta A)^2(\Delta B)^2 \geq \left| \int \Psi^*(x)(A - \langle A\rangle)(B - \langle B\rangle)\Psi(x)\,\mathrm{d}x \right|^2. \tag{5.47}$$

Our task now is to put the operator product in this equation into a usable form. To do this, we will call this product $G$ and rewrite it as a symmetrized product plus a difference term in the following manner

$$\begin{aligned} G &= (A - \langle A\rangle)(B - \langle B\rangle) \\ &= \frac{(A - \langle A\rangle)(B - \langle B\rangle) + (B - \langle B\rangle)(A - \langle A\rangle)}{2} \\ &\quad + \frac{(A - \langle A\rangle)(B - \langle B\rangle) - (B - \langle B\rangle)(A - \langle A\rangle)}{2}. \end{aligned} \tag{5.48}$$

The (numerator of the) symmetrized product will yield the average $\langle AB + BA\rangle - 2\langle A\rangle\langle B\rangle$. Since both $A$ and $B$ are assumed to be Hermitian operators, which yield real numbers upon measurement, this product is self-adjoint and therefore gives an expectation value that is a real quantity (the order of the averages is unimportant in the overall average). We will call this symmetrized product $F$ for convenience, and $\langle F\rangle$ is real. The second term yields just $AB - BA$, which is $\mathrm{i}C$. Since $C$ is also Hermitian, and is usually a $c$-number, this latter term is completely imaginary. Thus, we can write (5.48) as

$$G = \tfrac{1}{2}(F + \mathrm{i}C) \tag{5.49}$$

and

$$(\Delta A)^2(\Delta B)^2 \geq \tfrac{1}{4}(\langle F\rangle^2 + \langle C\rangle^2) \geq \tfrac{1}{4}\langle C\rangle^2. \tag{5.50}$$

In the last form, we have dropped $\langle F\rangle$, since it does not invalidate the basic inequality. Using the last form, we arrive at the expression of the Heisenberg uncertainty principle for non-commuting operators:

$$(\Delta A)(\Delta B) \geq \tfrac{1}{2}|\langle C\rangle| \tag{5.51}$$

which we first used in (1.35). For position and momentum $C = \hbar$, and this leads to the normal form with which we familiar.

We note that in the previous paragraph the Heisenberg uncertainty relation holds only for non-commuting operators, such as position and momentum. What about the often cited uncertainty in energy and time? In the non-relativistic quantum mechanics with which we are dealing, time is not an operator, and therefore it commutes with the total-energy operator—the Hamiltonian itself. Thus, there is no Heisenberg uncertainty principle for energy and time! We will, however, see in a later chapter that an eigenstate that possesses a lifetime can give rise to an uncertainty relation that greatly resembles the Heisenberg uncertainty relation—but there is a basic difference. The Heisenberg uncertainty relation describes basic fundamental uncertainties, while this latter relationship is related to a measurement of the energy alone, and is not therefore a fundamental uncertainty relation (Landau and Lifshitz 1958).

## 5.3 Linear vector spaces

When we talk about linear vector spaces, we are combining the concepts of state variables, used in linear-systems theory, and vector calculus, used in electromagnetic fields. In vector calculus, we choose a set of basis vectors to define a coordinate system; these can be for example the unit vectors along the three Cartesian coordinate directions in a rectangular coordinate system. Then, any general vector can be written as the sum of components in each of these directions, as

$$\boldsymbol{B} = b_1 \boldsymbol{a}_x + b_2 \boldsymbol{a}_y + b_3 \boldsymbol{a}_z. \tag{5.52}$$

Here, the $\boldsymbol{a}_i$ are the unit vectors and the $b_i$ are the components of the vector $\boldsymbol{B}$ in each of the various directions defined by the unit vectors. The same ideas carry over to linear vector spaces of operators. We choose a basis set of functions that define the coordinates, which are usually infinite in number. Then, we can expand a general function in terms of the components as

$$\Psi(x) = b_1\psi_1(x) + b_2\psi_2(x) + b_3\psi_3(x) + \cdots = \sum_i b_i\psi_i(x). \tag{5.53}$$

Again, the $\psi_i(x)$ are the unit vectors of this space and the $b_i$ are the components of the 'vector' in each of the various directions defined by the unit vectors. The difference between (5.52) and (5.53) is that the Cartesian coordinate system has only three unit vectors (is only three dimensional) while the linear vector space generally has an infinite number of dimensions (von Neumann 1932).

In the rectangular coordinate system for vectors, the unit vectors satisfy the relationship $\boldsymbol{a}_i \cdot \boldsymbol{a}_j = \delta_{ij}$, that is they satisfy an orthonormality condition. Similarly, we require the basis vectors in the linear vector space to be orthonormal (normally, but there are cases where this is not required). This is expressed as

$$\int \psi_i^*(x)\psi_j(x)\,\mathrm{d}x = \delta_{ij}. \tag{5.54}$$

The basis set also satisfies closure as expressed in (5.38). The linear vector space has many standard properties associated with it being a *linear* vector space. Some of these are

$$\psi_i(x) + \psi_j(x) = \psi_j(x) + \psi_i(x) \tag{5.55$a$}$$

$$\psi_i(x) + (\psi_j(x) + \psi_k(x)) = (\psi_i(x) + \psi_j(x)) + \psi_k(x) \tag{5.55$b$}$$

$$\mu(\lambda\psi_i(x)) = (\mu\lambda)\psi_i(x) \qquad \mu, \lambda \text{ are } c\text{-numbers} \tag{5.55$c$}$$

$$\lambda(\psi_i(x) + \psi_j(x)) = \lambda\psi_i(x) + \lambda\psi_j(x). \tag{5.55$d$}$$

In addition, there exists a null element $\psi_0(x) = 0$ such that $\psi_i(x) + \psi_0(x) = \psi_i(x)$. Finally, the basis vectors are linearly independent of each other, so the statement

$$b_1\psi_1(x) + b_2\psi_2(x) + b_3\psi_3(x) + \cdots + b_k\psi_k(x) = 0 \tag{5.56}$$

implies explicitly that $b_k = 0$ for all values of $k$ and $x$. This can be seen by taking the inner product (5.54) of the expression with one of the basis vectors; at most, only one term remains after the integration which requires that particular coefficient to vanish.

This last expression raises a very important point. If we can write the wave function in terms of just one of the basis set, such as

$$\Psi(x) = b_i \psi_i(x) \tag{5.57}$$

then we say that the wave function is in a *pure* state. On the other hand, the general expression (5.56) is for a *mixed* state—the wave function has contributions from many of the basis states. More importantly, the wave function (5.56) is said to be an *entangled* state, in that it cannot be separated into any of the basis vectors.

If we have two wave functions, we can expand them in terms of the same linear vector space as

$$\begin{aligned} \Psi_a(x) &= \sum_i a_i \psi_i(x) \\ \Psi_b(x) &= \sum_i b_i \psi_i(x). \end{aligned} \tag{5.58}$$

Then, the inner product of the two wave functions is

$$\begin{aligned} (\Psi_a, \Psi_b) &= \int \mathrm{d}x \sum_i a_i^* \psi_i^*(x) \sum_j b_j \psi_j(x) \\ &= \sum_i a_i^* \sum_j b_j \delta_{ij} = \sum_i a_i^* b_i. \end{aligned} \tag{5.59}$$

Obviously, then,

$$(\Psi_a, \Psi_a) = \sum_i |a_i|^2 = 1. \tag{5.60}$$

In (5.37), we expanded the wave function in a linear vector space set of basis vectors. This can be re-expressed as

$$\begin{aligned} \Psi(x) &= \int \sum_i \psi_i(x) \psi_i^*(x') \Psi(x') \, \mathrm{d}x' \\ &= \sum_i \psi_i(x)(\psi_i, \Psi). \end{aligned} \tag{5.61}$$

The inner product in the last line is recognized as merely being the coefficient $b_i$. It is useful to write the individual terms in the summation as

$$P_i(\bullet) = \psi_i(x)(\psi_i, \bullet) \tag{5.62}$$

by which we define a projection operator $P_i$ onto the $i$th coordinate. The projection operator picks out that portion of the total wave function that has the variation of this coordinate, much as we project a vector onto a preferred axis in vector algebra. The projection operator has some interesting properties:

$$P_i P_i = P_i \qquad P_i P_j = 0 \qquad \sum_i P_i = 1. \tag{5.63}$$

The first property follows from the fact that the first projection operator picks out that part of the function that projects onto the desired coordinate, while the second one must yield unity as the entire result of the first projection is already onto the desired axis. The second property follows from the fact that the first projects onto a desired coordinate, so there is no part left to project onto any other coordinate. The first expression is a statement of normalization, the second expresses orthogonality, and the last expresses closure.

### 5.3.1 Some matrix properties

Once we have selected a basis set of functions, then it is no longer necessary to continue to restate these functions. Rather, we need only to keep track of the expansion coefficients for the wave function. This leads to matrix operations. Consider, for example, a wave function expanded in the series given in the first line of (5.58). If we operate on this wave function with the operator $A$, let us assume that it produces the new wave function given by the second line of (5.58), that is $A\Psi_a = \Psi_b$. Thus, we write

$$A\Psi(x) = \sum_i a_i A\psi_i(x) = \sum_i b_i \psi_i(x). \tag{5.64}$$

Let us pre-multiply the last two expressions by $\psi_j^*(x)$, and integrate this over all space. This gives the results

$$b_j = \sum_i a_i A_{ji} \qquad A_{ji} = \int \psi_j^*(x) A \psi_i(x)\,\mathrm{d}x. \tag{5.65}$$

The last expression defines the *matrix elements* of the operator $A$. This gives the matrix expression (we use the standard notation of only a single right bracket for a column matrix)

$$b] = [A]a]. \tag{5.66}$$

This leads to the expectation value of the operator $A$ as

$$\langle A \rangle = (\Psi, A\Psi) = a]^{\mathrm{T}}[A]a] \tag{5.67}$$

where the superscript T implies the transpose operation (which here produces the row matrix). Since the operator is Hermitian, the matrix representation of the operator is a Hermitian matrix, which implies that $A_{ij}^* = A_{ji}$.

Suppose that after carefully choosing a set of basis functions, we discover that another choice would have been preferred. Just as one can do coordinate transformations in vector algebra, one can do coordinate transformations in the linear vector spaces. Initially, the wave function is defined in terms of the basis set $\psi_{1i}(x)$ as

$$\Psi(x) = \sum_i a_i \psi_{1i}(x) \tag{5.68}$$

which we want to write in terms of a 'rotation' matrix with new coordinates defined by

$$\psi_{2i}(x) = \sum_j S_{ij} \psi_{1j}(x). \tag{5.69}$$

If the wave function is expanded in terms of the new coordinates as

$$\Psi(x) = \sum_k b_k \psi_{2k}(x) \tag{5.70}$$

the various coefficients are related by

$$a_i = \sum_k S_{ik} b_k \tag{5.71}$$

or

$$a] = [S]b]. \tag{5.72}$$

We now use our operator $A$ of (5.64) and operate on the various wave functions. Thus, we obtain

$$[A_a]a] = [S][A_b]b] \tag{5.73}$$

by which we mean that the operator acts upon the individual basis functions, and hence will move through the matrix sum. Since the basis functions are changed on the right-hand side, the result is different and leads to there being different matrix elements, which have been denoted by the subscripts corresponding to the old and new sets of basis functions. By using the inverse of (5.72), we can relate the two matrix representations (in the new and old basis sets) via

$$[A_a] = [S][A_b][S]^{-1} \tag{5.74}$$

where the last term is the inverse of the transformation matrix. Equation (5.74) defines a *similarity* transformation. The two basis sets are orthonormal, so

$$\begin{aligned}(\Psi, \Psi) &= \sum_k |b_k|^2 = \sum_i |a_i|^2 \\ &= \sum_{klij} S_{lk}^* b_k^* S_{ji} b_i (\psi_l, \psi_j) \\ &= \sum_{kij} S_{jk}^* b_k^* S_{ji} b_i. \end{aligned} \tag{5.75}$$

By comparing the first term on the right of the first line and the last line, it is obvious that the similarity transformation must produce

$$\sum_j S_{jk}^* S_{ji} = \delta_{ki}. \tag{5.76}$$

We can rewrite this expression as

$$([S]^+[S])_{ki} = \delta_{ki} \tag{5.77}$$

which means that the product of these two matrices must be a unit matrix (values of unity on the main diagonal and zeros off the diagonal), so

$$[S]^+ = [S]^{-1}. \tag{5.78}$$

The last expression is the relationship defining a *unitary matrix*. The similarity transformations from one coordinate (orthonormal basis) set to another must be unitary transformations, and thus are generalized rotations.

### 5.3.2 The eigenvalue problem

As the final part of this section, we want to address the transformations that apply to the eigenvalue problem, as it is often called. In previous chapters, and sections, it was assumed that one could find a basis set in which an operator returned the same basis function multiplied by a *c*-number constant, as in (5.26). In general, we would have to be quite prescient to choose the basis set properly for this to occur for all, or even a significant fraction, of the operators of interest. How then do we find the eigenvalues of operators, which become importantly related to the expectation values of the operators? We know from the arguments of section 5.2.2 that, if an operator produces two different values when operating upon a set of functions, these latter functions must be orthogonal to one another. This means, in general, that a selected basis set can be found that will lead to a diagonal matrix for any selected operator, but it is unlikely to be diagonal for other operators unless the latter commute with the selected operator. This is a very special situation, which is not found in general.

On the other hand, we have asserted above that the choice of the basis set is not particularly important to the physics, but is generally made based upon other considerations such as the ease of solving the boundary condition problem. Thus, in general, we should expect that some linear combination of the chosen basis states, much like a Fourier series representation, will describe the basis states for any operator. The connection between the selected basis states and those that yield the diagonal representation for an operator is given by the similarity transformations described in the previous subsection. In this section, we show that this is the case, and that one can in fact diagonalize the basis set, or linear vector space, through a coordinate rotation defined by the similarity transformations.

In order to describe any physical problem, the first step is to choose a basis set of wave functions $\{\psi_i(x)\}$ that describe the linear vector space. These functions are properly orthogonalized, so the inner product of any distinct pair is zero, but each is normalized. From this selection of a basis set, it is possible to determine the expectation value of a general operator in such a way that the operator lies in a space defined by these operators. It may be asserted that there exists a proper choice of basis functions, which describes a vector space rotated with respect to the initial choice, in which the operator is directed solely along one of the basis functions:

$$A\xi_j(x) = \lambda\xi_j(x). \tag{5.79}$$

The rotated basis function can be expressed in terms of the original set by the similarity transformation (5.69), or

$$\sum_i a_i A\psi_i(x) = \lambda \sum_i a_i\psi_i(x) \tag{5.80}$$

where we have used the coefficients $a_i$ in the actual expansion of the new basis functions in terms of the old. If we now multiply by any one of the original basis functions and integrate, it gives one row of the matrix equation

$$[A]a] = \lambda a]. \tag{5.81}$$

This latter is the *eigenvalue equation*. Since any choice of the $a_i$ produces one of the $\xi_j$, we do not expect the vector $a]$ to be a null vector. Thus, the only solution of (5.81) is the choice

$$\det|[A] - \lambda[1]| = 0. \tag{5.82}$$

In this last expression, [1] is the unit matrix which has the elements given by $\delta_{ij}$. Solving this algebraic expression gives the allowed values of the eigenvalues of the operator $A$. If the space has $n$ dimensions, there are then $n$ values of these eigenvalues. Once these values are found, $n - 1$ of these can be substituted, in turn, into (5.81) to find $n - 1$ equations for the various coefficients, which relate the new basis states to the old (the last solution arises from normalization). These equations define the similarity transformation, since it is assumed that each value of an eigenvalue determined from (5.82) corresponds to defining one basis state of the rotated coordinate system.

The above also is dependent upon the values obtained for the eigenvalues from (5.82) all being distinct. If two or more values of $\lambda$ are the same, then a problem arises. This case is called degeneracy, as two (or more) different basis functions in the rotated coordinate system yield the same value of the eigenvalue. The solution to this is to consider as a new sub-coordinate system only those degenerate functions for which a linear combination (a second rotation) can be used to lift the degeneracy of the functions. That is, we form a linear combination of the functions that lifts the degeneracy, proceeding precisely as above in a second iteration. We will see an example of this in the next chapter.

The presence of eigenvalues raises an important issue with regard to the uncertainty discussed in section 5.2.3. Suppose the wave function is such that it is an eigenvalue of the operator $A$, so that

$$A\Psi(x) = a\Psi(x). \tag{5.83}$$

Then, the uncertainty associated with this operator is *identically zero*. To see this, we note that

$$\begin{aligned} \langle A \rangle &= \int \Psi^+ A \Psi \, \mathrm{d}x = \int \Psi^+ a \Psi \, \mathrm{d}x = a \\ \langle A^2 \rangle &= \int \Psi^+ A^2 \Psi \, \mathrm{d}x = a \int \Psi^+ A \Psi \, \mathrm{d}x = a^2 \end{aligned} \tag{5.84}$$

so that

$$(\Delta A)^2 = \langle A^2 \rangle - \langle A \rangle^2 = 0. \tag{5.85}$$

*An operator which satisfies an eigenvalue equation has zero uncertainty*. This is especially important for the time-independent Schrödinger equation (2.23). This equation is an eigenvalue equation for the total energy. On the left-hand side is the Hamiltonian, or energy operator, while the energy value is on the right-hand side. Since this is an eigenvalue equation, there is no uncertainty in the energy that results from its solutions. Each energy value is well described and subject to no variation that can be connected to an uncertainty principle.

### 5.3.3 Dirac notation

A useful notation, which is very common, is called Dirac notation. Rather than write out all of the various symbols for the basis function, its subscript, and its variational functionality, we use only the descriptor by which the set is labelled; for example, we use the subscript (or expansion parameter) as the identifier. Thus, any single basis state from the set $\{\psi_i(x)\}$ is defined as $|i\rangle$. This formulation is called the *ket* vector. Its complex conjugate is called a *bra* vector $\langle i|$. Thus,

$$\psi_i(x) \to |i\rangle \qquad \psi_i^*(x) \to \langle i|. \tag{5.86}$$

The inner product is written as the product of a bra and a ket vector (a bracket!), as

$$(\psi_i, \psi_j) = \langle i|j\rangle = \delta_{ij}. \tag{5.87}$$

We note in the last expression that the double vertical bar is dropped in favour of a single one, and that any time a bra and a ket are combined as shown, there is an inferred integration over the functional variable. Thus, the matrix elements of the operator $A$ would be given by

$$(\langle A \rangle)_{ij} = \langle i|A|j\rangle. \tag{5.88}$$

The projection operator of (5.62) can be written in terms of the Dirac notation in a much simpler manner, as

$$P_i = |i\rangle\langle i| \tag{5.89}$$

and

$$\sum_i |i\rangle\langle i| = 1 \tag{5.90}$$

is said to be an expansion of unity in the linear vector space (or a resolution of the delta function). If we want to rotate the coordinate system to a new basis set described by the set $|i'\rangle$, we can use the projection operator quite effectively, as

$$\langle i'|i\rangle = \langle i'|\sum_{i''} |i''\rangle\langle i''|i\rangle = \sum_{i''}\langle i'|i''\rangle\langle i''|i\rangle. \tag{5.91}$$

Each bracket represents a similarity transformation matrix $[S]$. This can be seen quite easily by letting $i$ and $i'$ be the same set of basis states, and replacing $i$ in (5.91) by say $j'$, a member of the $i'$-set. The result is just

$$\sum_{i''}\langle i'|i''\rangle\langle i''|j'\rangle = \delta_{i'j'} \tag{5.92}$$

which is just (5.77) describing a basis property of the similarity transformations, in this case between the $i'$-set of basis functions and the $i''$-set of basis functions.

In the case of an infinite, continuous set of functions (the index is a continuous rather than a discrete one), the sums become integrals, and the Kronecker delta becomes a Dirac delta function:

$$\langle A|A'\rangle = \delta(A - A') \qquad \sum_i |i\rangle\langle i| \to \int |A\rangle\langle A|\,\mathrm{d}A. \tag{5.93}$$

In the case of the position representation used in the earlier chapters, $|x\rangle$ is a wave packet that is localized at the site $x$. Thus, we expect that

$$\langle x|x'\rangle = \delta(x - x'). \tag{5.94}$$

If we have a discrete function defined at a position, we can write this function as

$$|a\rangle = \int |x\rangle\,\mathrm{d}x\langle x|a\rangle \tag{5.95}$$

where we have introduced the projection operator for projection onto the coordinate (the resolution of the delta function) and we can recognize that

$$\psi_a(x) = \langle x|a\rangle. \tag{5.96}$$

If we operate with the position operator, it can be seen that

$$\langle x'|x|x\rangle = \hat{x}\langle x'|x\rangle = \hat{x}\delta(x - x'). \tag{5.97}$$

In general, any function of position can be expanded in a Taylor series, and (5.97) used to show that

$$\langle x'|f(x)|x\rangle = f(\hat{x})\langle x'|x\rangle = f(\hat{x})\delta(x - x'). \tag{5.98}$$

These general operations can be continued almost infinitely. It is important to remember that the Dirac notation is a shorthand, and should be used accordingly—that is, to simplify the equations, but not to obfuscate them.

## 5.4 Fundamental quantum postulates

One of the things that we would like to do is to show once again that the operator and matrix approach is fully compatible with the Schrödinger equation. To that end, we will consider here a number of operator functions and principles, all of which lead us to show that the Heisenberg picture is fully equivalent to the Schrödinger picture. At the end, we find a set of postulates that define the transition from classical dynamics to quantum dynamics.

### 5.4.1 Translation operators

Consider an operator that translates a wave packet in position by an infinitesimal amount $\xi$. In other words, we define a translation operator $D_\xi$ in a manner that produces the result

$$D_\xi|x\rangle = |x + \xi\rangle. \tag{5.99}$$

Now, this operator has the property of displacing the wave function in the positive direction, and this can be checked by noting that

$$\langle x'|D_\xi|x\rangle = \langle x'|x + \xi\rangle = \delta(x' - x - \xi). \tag{5.100}$$

In general $D_\xi D_\eta = D_{\xi+\eta} = D_\eta D_\xi$, so the displacement operator preserves orthonormality and completeness of the entire basis set of functions. Thus, $D_\xi$ is *unitary*. In general, then, we can write $D_\xi = \mathrm{e}^{\mathrm{i}A(\xi)}$, which ensures that the unitarity is preserved $(D^+D = 1)$. The properties of $D_\xi$ can now be used to determine just what value $A$ must have. We note that

$$(D_\xi)^2 = D_{2\xi} \tag{5.101}$$

and so

$$2A(\xi) = A(2\xi), \ldots, nA(\xi) = A(n\xi) \qquad \text{for any } n \tag{5.102}$$

and, by letting $\xi = 1$, we see that

$$A(n) = nA(1). \tag{5.103}$$

This, in turn, implies that $A(\xi) \propto \xi$, or $D_\xi = \mathrm{e}^{+\mathrm{i}k\xi}$. We have chosen the coefficient as the wave vector for the sake of units and taken the negative sign rather arbitrarily, but both can be checked by expanding in a Taylor series, as

$$\begin{aligned} D_\xi|x\rangle &= \mathrm{e}^{+\mathrm{i}k\xi}|x\rangle = (1 + \mathrm{i}k\xi + \tfrac{1}{2}(\mathrm{i}k\xi)^2 + \cdots)|x\rangle \\ &= (1 + \xi\boldsymbol{\nabla} + \tfrac{1}{2}|\xi\boldsymbol{\nabla}|^2 + \cdots)|x\rangle \\ &= (1 + (+\xi)\boldsymbol{\nabla} + \tfrac{1}{2}(+\xi)^2\nabla^2 + \cdots)|x\rangle \\ &= |x + \xi\rangle. \end{aligned} \tag{5.104}$$

We have used the exponential expansion in the first line, expressed $k$ in terms of the momentum operator in the second line, and recognized that the third line is the Taylor series for the fourth line. Thus, the choice of the sign satisfies the original definition (5.99). It may also be recognized that the infinitesimal translation operator is a momentum wave function corresponding to the small displacement.

We can expand the general idea that the translation operator is a momentum wave function by some simple arguments. First, we note that we can use the general orthonormality of the basis sets and (5.96) to allow us write that $\langle x|k'\rangle = g(k')\mathrm{e}^{\mathrm{i}k'x}$, where $g(k')$ is a scalar function. This leads to

$$\begin{aligned} \langle k|k'\rangle = \delta(k - k') &= \int \langle k|x\rangle\,\mathrm{d}x\langle x|k'\rangle \\ &= g^*(k)g(k')\int \mathrm{e}^{\mathrm{i}(k'-k)x}\,\mathrm{d}x \\ &= 2\pi g^*(k)g(k')\delta(k - k') \end{aligned} \tag{5.105}$$

from which we find that, within an arbitrary phase factor that will be ignored,

$$g(k) = \frac{1}{\sqrt{2\pi}} \qquad \langle x|k'\rangle = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{\mathrm{i}k'x}. \tag{5.106}$$

Thus, we can expand the momentum wave function as

$$|k'\rangle = \int |x\rangle\,\mathrm{d}x\langle x|k'\rangle = \frac{1}{\sqrt{2\pi}}\int \mathrm{e}^{\mathrm{i}k'x}|x\rangle\,\mathrm{d}x \tag{5.107}$$

which, if we make the identifications $\phi(-k) = |k'\rangle$ and $\psi(x) = |x\rangle$, is just the inverse of (1.26) (the first line of (1.37)). Hence, we are indeed back to momentum wave functions.

### 5.4.2 Discretization and superlattices

The displacement operators of the previous paragraphs point out an important point in regard to finite difference discretization that is used for numerical solutions of the Schrödinger equation. In (2.124), the finite difference version of the time-independent equation is just

$$-\frac{\hbar^2}{2ma^2}(\Psi_{i+1} + \Psi_{i-1} - 2\Psi_i) + V_i\Psi_i = E\Psi_i \tag{5.108}$$

or

$$-\frac{\hbar^2}{2ma^2}(\Psi(x+a) + \Psi(x-a) - 2\Psi(x)) + V(x)\Psi(x) = E\Psi(x). \tag{5.109}$$

Using the displacement operator $D_\mathrm{a} = \mathrm{e}^{\mathrm{i}ka}$ from (5.104), this may be written as

$$-\frac{\hbar^2}{2ma^2}(\mathrm{e}^{\mathrm{i}ka} + \mathrm{e}^{-\mathrm{i}ka} - 2)\Psi(x) + V(x)\Psi(x) = E\Psi(x) \tag{5.110}$$

or

$$E = V(x) + \frac{2\hbar^2}{ma^2}\sin^2(ka/2). \tag{5.111}$$

*This is not the proper energy relationship*. In the absence of the potential, the energy should just be the free-electron dispersion relation

$$E = \frac{\hbar^2 k^2}{2m}. \tag{5.112}$$

This can only be recovered from (5.111) if we insist that $ka/2 \ll 1$. That is, the discretization size $a$ must be much smaller than the inverse of the smallest momentum wave vector of interest

$$a \ll \frac{2}{k}. \tag{5.113}$$

Alternatively, we must require that the energy range of interest be sufficiently small that

$$E \ll \frac{2\hbar^2}{ma^2} \ll 4|S_{i,i\pm1}| \tag{5.114}$$

where the last form introduces the nearest-neighbour hopping energy from (2.126). If this is not strictly maintained, then the artificial band structure of (5.111) that is due to the discretization of the equation rather than the intrinsic physics will dominate the energy levels.

On the other hand, materials systems can be fabricated in which a superlattice is formed, such as with alternating layers of GaAs and AlAs. In this case, the parameter $a$ is the periodicity of this new superlattice, and the resulting band structure is quite real. The wide band gap material AlAs creates barriers to the conduction band electrons (and the valence band holes) in the GaAs, just as in the Kronig–Penney model of section 3.7. Translation from one GaAs layer to the next, at a distance of $a$, brings the wave function back to that of the GaAs electrons (or holes). Hence, in this case, the periodic structure produces a real mini-band system, in which the new allowed energy levels are described by (5.111). The band width of the new mini-band is just twice the hopping energy (2.126). It was in these structures that Esaki and Tsu sought to find Bloch oscillations, as discussed in section 3.7.1.

Imposing a magnetic field creates new effects in this mini-band system. Here, we will ignore the potential $V(x)$ by assuming that it has been taken care of in the superlattice structure that gives rise to the periodicity $a$. Then, we can rewrite (5.110) and (5.111) as

$$\frac{\hbar^2}{ma^2}[\cos(ka) - 1]\Psi(x) = E\Psi(x). \tag{5.115}$$

With the presence of the magnetic field, however, we must also treat the transverse direction, assuming the magnetic field is in the $z$-direction. Thus, we can extend (5.115) to the two-dimensional case as

$$\frac{\hbar^2}{ma^2}[\cos(k_x a) + \cos(k_y a)]\Psi(x, y) = E\Psi(x, y) \tag{5.116}$$

and the constant term has been absorbed into the energy as a shift of this energy by four times the hopping energy $S = \hbar^2/2ma^2$. Since the energy is relative to an arbitrary reference level, this shift is not important to the overall discussion. At this point, the magnetic field is introduced in the Landau gauge of section 4.7 as

$$\frac{\hbar^2}{ma^2}\left[\cos(k_x a) + \cos\left(k_y a - \frac{eBxa}{\hbar}\right)\right]\Psi(x, y) = E\Psi(x, y). \tag{5.117}$$

At this point, it is useful to return to the 'displaced' coordinates, and we may rewrite (5.117) as

$$\begin{aligned} E\Psi(x, y) = S[&\Psi(x + a, y) + \Psi(x - a, y) \\ &+ \mathrm{e}^{-\mathrm{i}eBxa/\hbar}\Psi(x, y + a) + \mathrm{e}^{\mathrm{i}eBxa/\hbar}\Psi(x, y - a)]. \end{aligned} \tag{5.118}$$

Here, the wave function is connected to its four nearest neighbours, exactly as in a finite difference discretization for the Schrödinger equation (but that is where we began). Indeed, this is the heart of the tight-binding method of band structure. To simplify this result, the fact that the wave functions are periodic in $x$ and $y$ with period $a$ suggests that we introduce the reduced coordinates

$$x = ra \qquad y = sa \qquad r, s = 0, \pm 1, \pm 2, \ldots. \tag{5.119}$$

In addition, the only remaining operator is in the $x$-direction, which is that of the initial superlattice, so this suggests that the $y$-variation be taken to be plane waves, and

$$\Psi(x, y) = \Psi(ra, sa) = \mathrm{e}^{\mathrm{i}\kappa s}\varphi(r) \qquad \kappa = k_y a. \tag{5.120}$$

With these substitutions, (5.118) becomes Harper's equation (Harper 1955)

$$\begin{aligned} E\varphi(r) &= S\left[\varphi(r+1)+\varphi(r-1)+\cos\left(\frac{eBa^2}{\hbar}r-\kappa\right)\varphi(r)\right] \\ &= S[\varphi(r+1)+\varphi(r-1)+\cos(2\pi\alpha r-\kappa)\varphi(r)]. \end{aligned} \tag{5.121}$$

Here,

$$\alpha = \frac{eBa^2}{h} = \frac{\Phi}{\Phi_0} \tag{5.122}$$

is the flux (in terms of the quantum unit of flux $h/e$) that is coupled through each 'unit cell' of the superlattice structure. Hofstadter (1976) studied the energy spectrum of (5.121) extensively, and found that the energy is periodic in $\Phi_0$ in magnetic field and in $2S$ in energy. As a general rule, the energy spectrum is *fractal*, having solutions only when the parameter $\alpha$ is a rational number (e.g., a ratio of two integers). Otherwise, the orbits do not close upon themselves. This rationality means that the magnetic length $l_\mathrm{B} = \sqrt{\hbar/eB}$ must be in a rational relationship with the periodicity $a$—that is, a fixed number of cycle lengths $l_\mathrm{B}$ must correspond to a fixed number of factors of $a$. The fact that the energy structure, and also the conductivity, is periodic in magnetic field was first noted by Azbel (1963), who seems to have not continued his investigations when he computed the needed magnetic field if $a$ corresponded to a regular inter-atomic spacing. However, with superlattices, it is now possible to study such effects (Ferry 1992).

### 5.4.3 Time as a translation operator

The results of the last subsection showed that we could describe the momentum wave function in a manner that produced a displacement, or translation, operator on the wave function in the position representation. Here, we want to examine the time evolution operator, an operator that differs somewhat from the Green's function kernel introduced in chapter 2. In particular, we want to examine a linear operator $T$ that gives the time evolution of the wave packet, or of any averages arrived at in the linear vector space. Then, we will turn to treating equations of motion that are obtained from the Heisenberg picture; for example, equations of motion obtained through forming commutators with the total Hamiltonian. These lead to an equivalence principle which will form a basis for quantum mechanics in the next section.

We recall that the Schrödinger equation is a linear first-order equation in its time evolution, so the specification of the initial state at $t = 0$ fully determines the time evolution. Thus, we can specify the state vectors as

$$\psi_a(x,t) = |a,t\rangle = T(t,t_0)|a,t_0\rangle \tag{5.123}$$

where $|a,t_0\rangle$ is the initial basis state and $T$ is an independent linear operator that produces the time evolution. Note that in section 2.8.2, the Green's function kernel needed to be integrated over all values of the initial position. Here, on the other hand, $T$ is an operator that produces the time evolution, and refers to a particular basis state $|a\rangle$ as one of the set of states forming the linear vector space. However, in the present context, the time variation remains with the basis states and not with the operators, although we will remove this shortly, returning to the operators as rotating functions.

Since the time evolution operator is a linear operator, it can be decomposed into products of similar operators in a manner consistent with the property of local time behaviour (Markovian behaviour, i.e., no memory). This may be stated as

$$|a, t\rangle = T(t, t_0)|a, t_0\rangle = T(t, t_1)T(t_1, t_0)|a, t_0\rangle \tag{5.124}$$

or

$$T(t, t_0) = T(t, t_1)T(t_1, t_0). \tag{5.125}$$

It is also obvious that $T(t, t) = 1$, so

$$T(t, t) = T(t, t_1)T(t_1, t) = 1 \tag{5.126}$$

and so

$$T(t, t_1) = [T(t_1, t)]^{-1}. \tag{5.127}$$

Now, let us define an operator $A$ such that

$$T(t + \varepsilon, t) = 1 - \frac{\mathrm{i}A\varepsilon}{\hbar} \tag{5.128}$$

where $\varepsilon$ is a very small parameter which we will ultimately let tend to zero. The left-hand side of (5.128) can also be written as

$$T(t + \varepsilon, t_0) = T(t + \varepsilon, t)T(t, t_0). \tag{5.129}$$

With these results, we can now write

$$\frac{\mathrm{d}T(t, t_0)}{\mathrm{d}t} = \lim_{\varepsilon \to 0} \frac{T(t + \varepsilon, t_0) - T(t, t_0)}{\varepsilon} = -\frac{\mathrm{i}A}{\hbar}T(t, t_0) \tag{5.130}$$

and

$$\mathrm{i}\hbar\frac{\mathrm{d}T(t, t_0)}{\mathrm{d}t} = AT(t, t_0). \tag{5.131}$$

Comparing this with the Schrödinger equation, it is immediately apparent that the operator $A$ is the Hamiltonian, the total-energy operator, and re-insertion of the basis function leads to the Schrödinger equation

$$\mathrm{i}\hbar\frac{\mathrm{d}|a\rangle}{\mathrm{d}t} = H|a, t\rangle. \tag{5.132}$$

Further, (5.131) is readily solved to give the time evolution operator as

$$T(t, t_0) = \exp\left[-\frac{\mathrm{i}H(t - t_0)}{\hbar}\right]. \tag{5.133}$$

Clearly, this also satisfies (5.127), and, in fact, $T$ is a unitary (and Hermitian) operator.

Now, let us see how this gives the equivalence principle. We recall that the time rate of change of the expectation value of an operator is given by

$$\begin{aligned} \mathrm{i}\hbar\frac{\mathrm{d}\langle A\rangle}{\mathrm{d}t} &= \mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\langle a,t|A|a,t\rangle \\ &= \left\langle [A,H] + \mathrm{i}\hbar\frac{\partial A}{\partial t}\right\rangle . \end{aligned} \tag{5.134}$$

In general, $A$ does not depend explicitly on time, so the last partial derivative in the bracket vanishes, and (5.18) is recovered. Thus, again, we are led to the conclusion that the time evolution of the average of an operator is determined by the manner in which that operator commutes with the Hamiltonian.

How can we use these operator relationships to establish the principles of quantum mechanics? The fact that we have established that a set of operators can lead to their respective average values, and that the time evolution is given by an operator $H$, which is equivalent to the total-energy operator in the Schrödinger equation, does not establish this as quantum mechanics yet. We must still invoke some quantum conditions that establish the system as a quantum mechanical one. In doing this, we invoke a correspondence principle: if a quantum system has a classical analogue, we must regain this classical analogue as $\hbar \to 0$. Further, we need a set of quantization rules. Certainly, we have these rules from earlier chapters—but suppose we did not. Can we proceed to show that they are consistent with this correspondence principle? The answer is, of course, yes. Consider the application of (5.134) to the position operator

$$\mathrm{i}\hbar\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t} = \langle [x,H]\rangle . \tag{5.135}$$

Classically, the same equivalent formula is (in one dimension)

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H}{\partial p}. \tag{5.136}$$

Therefore, the correspondence principle requires that we have

$$\lim_{\hbar\to 0}\frac{\langle xH - Hx\rangle}{\mathrm{i}\hbar} = \frac{\partial H}{\partial p}. \tag{5.137}$$

Similarly, we require that

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \lim_{\hbar\to 0}\frac{\langle pH - Hp\rangle}{\mathrm{i}\hbar} = -\frac{\partial H}{\partial x}. \tag{5.138}$$

We can achieve the quantization of the system by asserting two postulates, which are really all that are needed to introduce quantum mechanics. First, it is postulated that the Hamiltonian operator $H$ is a Hermitian operator that is identical *in form* to the classical one; for example, it represents simply a sum of the various energies, particularly (and most commonly) the kinetic and potential energies. The second postulate is that operators for conjugate variables, such as position and momentum, do not commute, but satisfy a commutator relationship given by

$$[x, p] = \mathrm{i}\hbar. \tag{5.139}$$

This commutator relationship is compatible with now using (5.138) to express the momentum operator, in position space, as

$$p = -\mathrm{i}\hbar\frac{\partial}{\partial x} \tag{5.140}$$

and using (5.137) to express the position operator, in momentum space, as

$$x = \mathrm{i}\hbar\frac{\partial}{\partial p}. \tag{5.141}$$

One problem that arises with the above is the ordering of products of non-commuting operators, such as terms like $xp$. Classically, the order of the terms is not important, but for non-commuting operators, the result of (5.134) depends critically upon whether such a term in the Hamiltonian is written as $xp$ or $px$. Often, symmetrized products such as $(xp + px)/2$ will be used to avoid this problem, but it is only a problem when such cross products arise.

To illustrate the results from these postulates, we consider the classical harmonic oscillator, in which the Hamiltonian is

$$H = \frac{p^2}{2m} + \frac{1}{2}m\omega^2 x^2. \tag{5.142}$$

By the first postulate, this carries over directly to the quantum mechanical case, but now the conjugate variables are operators. We can now use (5.134) to evaluate the time evolution of the expectation values for these operators. First, the position operator gives

$$\mathrm{i}\hbar\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t} = \langle [x, H]\rangle = \frac{1}{2}m\langle [x, p^2]\rangle = \frac{\mathrm{i}\hbar}{m}\langle p\rangle. \tag{5.143}$$

This result is just (2.96), but in a much more informative picture. Now, the time rate of change of the expectation value of the momentum is similarly given by

$$\mathrm{i}\,\hbar\frac{\mathrm{d}\langle p\rangle}{\mathrm{d}t} = \langle [p, H]\rangle = \frac{1}{2}m\omega^2\langle [p, x^2]\rangle = -\mathrm{i}\hbar m\omega^2\langle x\rangle \tag{5.144}$$

which again is just (2.99) for this particular potential energy. Thus, (5.134) for the momentum is equivalent to the Ehrenfest theorem as well. Here, however, the achievement of the relevant results is much more straightforward. We rely only upon the two postulates, which were fully used in the earlier chapters. It should also be noted that, in both of these cases, Planck's (reduced) constant drops out of the equations, so the classical result is directly obtained, except of course that we are dealing with operators.

The fact that the Hamiltonian is the major function in quantum mechanics raises an interesting issue that is important in terms of commutator relationships. We note that for non-commutating operators such as $x$ and $p$, their time rate of change is given by the expectation value of a derivative of the Hamiltonian with respect to their conjugate variable, according to (5.136) and (5.138). More importantly, this time rate of change is found from the commutation relationship with the Hamiltonian. What about an energy operator, such as the Hamiltonian? One often wants to express an energy–time uncertainty principle, although we pointed out in section 5.2.3 that this does not make any sense. The time-evolution operator is the Hamiltonian, but this certainly commutes with itself. In the reversible Schrödinger equation, energy is conserved as a constant of the motion. Therefore, there is no uncertainty in this variable. Time is a parameter by which the progress of the system is measured, and therefore is not a dynamical variable which can be used to infer an uncertainty relationship. As we pointed out in chapter 1, there is a *classical* problem in measuring the energy within a fixed amount of time, and this leads to an *indeterminism* from the Fourier transform relationship between time and frequency (energy). This classical problem is often confused as a quantum mechanical relationship, but this is erroneous and such usage should be avoided rigorously.

### 5.4.4 Canonical quantization

As a last consideration, we want to revisit the concepts that have been applied to a number of examples in previous chapters. In these examples, we arbitrarily (it seems) chose a set of conjugate operators, which were then used to introduce quantization through a commutation relationship. Is there any rationale in selecting these operators, and can we be sure that a particular selection is the correct one? In short, when a particular problem is approached, are there a set of coordinates, other than the normal position and momentum, that are useful in solving the problem? The answer is the same as is arrived at in classical mechanics: if there is a set of canonically conjugate generalized coordinates that simplify the system Hamiltonian, then it is fruitful to use these coordinates in the problem at hand. An example is to solve a problem with cylindrical symmetry in cylindrical coordinates, where the coordinates might for example be the azimuthal angle and the angular momentum. In the $LC$-circuit, the variables were charge and flux, and this pair forms another example.

We consider the position and momentum coordinates $x$ and $p$ to be a pair that satisfy the commutation relationship (5.139). However, let us assert that there is a conjugate pair of operators $q$ and $\pi$, which satisfy the transformations

$$q = x + \varepsilon \frac{\partial G}{\partial p} \qquad \pi = p - \varepsilon \frac{\partial G}{\partial x} \tag{5.145}$$

where $G$ is a new (classical) function which generates a transformation of the original Hamiltonian expressed in terms of $x$ and $p$. In general, we take the parameter $\varepsilon$ as a small parameter so that linear variations can be assumed. Then, the Hamiltonian can be expressed in terms of the new parameters through a Taylor series as

$$\begin{aligned} H'(q,\pi) &= H(x,p) + (q-x)\frac{\partial H}{\partial x} + (p-\pi)\frac{\partial H}{\partial p} + \cdots \\ &= H(x,p) + \varepsilon \frac{\partial G}{\partial p}\frac{\partial H}{\partial x} - \varepsilon \frac{\partial G}{\partial x}\frac{\partial H}{\partial p} + \cdots. \end{aligned} \tag{5.146}$$

The second and third terms in the last line of (5.146) form the classical Poisson bracket relationship, which goes over into the commutator relationship. In essence, the transformation between the two sides of (5.145) in quantum mechanics must be a unitary transformation, because it is primarily a coordinate transformation. The primary question is that of whether we can find a suitable function $G$ that makes the quantum mechanical problem simpler.

In carrying the problem over to quantum mechanics, we note that (5.145) is principally a classical statement. When the problem is made one of quantum mechanics, the partial derivative with respect to position is carried into the momentum operator, the partial derivative with respect to momentum is made the position operator, and the overall partial derivatives become commutators. Thus, the quantum mechanical version of (5.145) is written as

$$q = x + \frac{\varepsilon}{\mathrm{i}\hbar}[x, G] \qquad \pi = p + \frac{\varepsilon}{\mathrm{i}\hbar}[p, G] \tag{5.147}$$

and it is recognized now that $x$ and $p$ are operators as are $q$ and $\pi$. Now, since $\varepsilon$ is a small parameter, and we are linearizing the results, this last expression can be rewritten as

$$\begin{aligned} q &= \left(1 + \frac{\mathrm{i}\varepsilon}{\hbar}G\right) x \left(1 - \frac{\mathrm{i}\varepsilon}{\hbar}G\right) = U_\varepsilon \, x U_\varepsilon^+ \\ \pi &= \left(1 + \frac{\mathrm{i}\varepsilon}{\hbar}G\right) p \left(1 - \frac{\mathrm{i}\varepsilon}{\hbar}G\right) = U_\varepsilon p U_\varepsilon^+. \end{aligned} \tag{5.148}$$

Thus, each of the terms in large parentheses corresponds to an infinitesimal 'rotation' defined by the unitary transformation $U_\varepsilon$. These new conjugate coordinates still satisfy the commutation relationship, since

$$\begin{aligned} [q,\pi] &= U_\varepsilon x U_\varepsilon^+ U_\varepsilon p U_\varepsilon^+ - U_\varepsilon p U_\varepsilon^+ U_\varepsilon x U_\varepsilon^+ \\ &= U_\varepsilon x p U_\varepsilon^+ - U_\varepsilon p x U_\varepsilon^+ = U_\varepsilon [x,p] U_\varepsilon^+ = \mathrm{i}\hbar. \end{aligned} \tag{5.149}$$

While the transformation so far has been an infinitesimal transformation, it can be carried to higher orders. We note that if we apply the transformation repeatedly, we find that

$$\begin{aligned} U &= \lim_{N\to\infty}\left(1+\frac{\mathrm{i}\varepsilon}{\hbar}G\right)^N = \lim_{N\to\infty}\left(1+\frac{\mathrm{i}\lambda}{\hbar N}G\right)^N \\ &= \exp\left(\frac{\mathrm{i}\lambda}{\hbar}G\right). \end{aligned} \tag{5.150}$$

Between the first and second lines, we have made the substitution of $\varepsilon = \lambda/N$ as the small parameter, where $\lambda$ is not necessarily small. If the substitutions $G \to H$, $\lambda \to t$ are made, then the unitary transformation is just the time evolution operator. It is this relationship that led us to remark that the time evolution operator is a generalized rotation in the linear vector space, since the unitary transformations dealt with above are generalized rotations. Indeed, $q$ and $\pi$ could be the time-varying forms of $x$ and $p$, which implies that the time variation is a series of infinitesimal translations determined by unitary operators. In general, however, (5.150) is one class of important coordinate transformations. These transformations produce useful operators like the raising and lowering operators of the harmonic oscillator.

The above treatment also leads us to the realization that numerical simulation of quantum mechanics by infinitesimal steps (time transformations) is a quite viable procedure for finding solutions to complicated problems. This approach, which addresses the more general subject of computational quantum mechanics, allows us to attack much more difficult problems than can be solved with simple operators and coordinate systems.

We may summarize the above through the main principle in which we find a coordinate system in which there may be constants of the motion, such as energy and momentum. Then we choose 'natural' variables in this system that are conjugates of one another. From these, a coordinate transformation from more normal position and momentum coordinates leads to the equivalent quantization relationships. Solutions then follow much more easily, even if they must be obtained numerically. Our example of the previous chapter was the use of the raising (creation) and lowering (annihilation) operators. These provided a much easier method of solution, and yielded the solutions much more quickly than the troublesome determination of the Hermite polynomials.

## References

Azbel M Ya 1963 *J. Exp. Theor. Phys.* **44** 980
Ferry D K 1992 *Prog. Quantum Electron.* **16** 251
Harper P G 1955 *Proc. R. Soc.* A **68** 874
Heisenberg W 1925 *Z. Phys.* **33** 879
Heisenberg W 1927 *Z. Phys.* **43** 172

Hofstadter D R 1976 *Phys. Rev.* B **14** 2239

Kroemer H 1994 *Quantum Mechanics* (Englewood Cliffs, NJ: Prentice-Hall)

Landau L D and Lifshitz E M 1958 *Quantum Mechanics: Non-Relativistic Theory* (Reading, MA: Addison-Wesley) pp 150–3

Merzbacher E 1970 *Quantum Mechanics* (New York: Wiley)

Schiff L I 1955 *Quantum Mechanics* 2nd edn (New York: McGraw-Hill)

von Neumann J 1932 *Mathematische Grundlagen der Quantummechanik* (Berlin: Springer)

## Problems

1. A dynamic system has $H = p^2 + Ax + Bpx$. Calculate the equations of motion for the operators $x, p$; that is, compute the time derivatives of the expectation values for these operators.

2. A particular operator $A$ has been used to evaluate a dynamical property in a orthonormal set containing only three basis functions. When its expectation value is found, the resulting matrix for the operator is given by

$$[A] = \frac{1}{2}\begin{bmatrix} +\sqrt{2} & \sqrt{2} & 0 \\ -\sqrt{2} & \sqrt{2} & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Determine the eigenvalues of this operator.

3. Compare the different dynamics that results from Hamiltonians that are written as $2x^2p^2$, $x^2p^2 + p^2x^2$ and $2(xp)^2$. What can you conclude about the proper symmetrization of operators?

4. Two operators $A$, $B$ satisfy $[A, B] = C$. Compute the difference $A^3B^3 - (AB)^3$. Assume that $C$ is a $c$-number and not an operator.

5. A particular projection operator $P_j = |j\rangle\langle j|$ is defined on the basis set

$$\psi_k(x) = \sqrt{\frac{2}{a}}\sin\left(\frac{k\pi x}{a}\right)$$

for $0 \leq x \leq a$ ($k$, $j$ are integers). Expand the function $V(x)$ in the basis set $|j\rangle$ if $V(x) = 1$ for $0 \leq x \leq a$, and $V(x) = 0$ elsewhere; that is, determine the coefficients $c_j$ in the expansion

$$V(x) = \sum_j c_j|j\rangle.$$

6. Consider a computational scheme in which the real axis (the position coordinate) is discretized to points $x = na$, where $a$ is a ‘lattice’ constant. Instead of the continuous eigenvalue characteristics of (5.85), each lattice site must be characterized by a function that allows the orthogonality of (5.85) to be retained. Determine a function localized at each lattice site that satisfies normalization and is orthogonal with all of its neighbours. (The functions will yield an integrated product that is one of the generalized distribution forms of the delta function.)

7. Numerically solve equation (5.121) for the allowed energy values for a range of magnetic fields (reproduce the Hofstadter butterfly) over the range of $0 < B < \Phi_0/a^2, 0 < E < 2S$.

8. Consider a potential well which is described by the potential $V(x) = Fx$ for $0 < x < a$, and $V(x) \to \infty$ for $x < 0$ and $x > a$. Using the wave functions of the infinite potential well

$$\psi_n(x) = \sqrt{\frac{2}{a}}\sin\left(\frac{n\pi x}{a}\right)$$

compute the Hamiltonian matrix for the lowest five energy levels. Take $F = 0.02 \text{ eV nm}^{-1}$ and $a = 5 \text{ nm}$. Then, diagonalize the matix to determine the energy eigenvalues and their wave functions. (While a proper approach would use an infinite set, approximate results can be found from these lowest five wave functions.) Compare the results with those of problem 15 of chapter 2.

# Chapter 6

# Stationary perturbation theory

It is generally nice to be able to solve a problem exactly. However, this is often not possible, and some approximation scheme must be applied. One example of this is the triangular potential well of section 2.6 or the finite potential well of section 2.5. In the former example, complicated special functions were required to solve Schrödinger's equation. In the latter example, the solution could not be obtained in closed form, and graphical solutions were used to find the eigenvalues of the energy. These are examples of more common problems. In the previous chapter, however, it was pointed out that one could choose an arbitrary basis set of functions that formed a complete orthonormal set as a defining linear vector space. The only requirement on these functions (other than their properties as a set) was that they made the problem easy to solve. In the two previous examples mentioned, what choice would we have made for the set? One choice would have been to use the basis functions that arise from an infinitely deep potential well, but these are not really defined outside the range of the well itself. Nevertheless, it is usually found that it is difficult to find a proper basis set of functions for which the problem can be easily solved.

When the problem is not calculated easily with a direct method, one must then resort to approximate methods. One such method is *perturbation theory*. This approach is useful when the Hamiltonian can be split into two parts

$$H = H_0 + H_1 \tag{6.1}$$

in which the first part $H_0$ can be solved directly, and the second part $H_1$ is small. When this can be done, then the second part can be initially ignored, and the problem solved exactly in terms of a natural basis function set. After this is done, the second term in (6.1) is treated as a perturbation, and the eigenvalues and eigenfunctions are then developed in a perturbation series in terms of the natural basis function set.

In this chapter, we will develop the general methodology of the perturbation technique. We will then consider several examples of the technique. Finally, a totally different approach, the variational method, will be introduced. The latter is useful when the split of the Hamiltonian into two parts, as in (6.1), does not lead to a small perturbing term.

## 6.1 The perturbation series

As was discussed above, the basic approach revolves around being able to solve the quantum mechanical problem with only a portion of the total Hamiltonian, with the remainder being a small term. For this, we rewrite (6.1) as

$$H = H_0 + \lambda V \tag{6.2}$$

where $\lambda$ is a small parameter (if $V$ is small compared with $H_0$, we can eventually let $\lambda$ go to unity without introducing any inconsistency). The introduction of this parameter is solely to help us to develop the proper terms in the perturbation series. The term $V$ is the perturbing part of the Hamiltonian and contains all of the extra terms that have been removed from $H$ to produce $H_0$. Now, it is asserted that the solutions to $H_0$ are easily obtained in terms of a set of basis functions for which

$$H_0|k\rangle_0 = \mathcal{E}_k^{(0)}|k\rangle_0. \tag{6.3}$$

That is, we have a set of zero-order basis functions for which one can find the appropriate eigenvalues of $H_0$. If the deviation of the actual basis functions and eigenvalues from these zero-order ones is small, then we can write the basis functions as

$$|k\rangle = |k\rangle_0 + \lambda|k\rangle_1 + \lambda^2|k\rangle_2 + \cdots \tag{6.4}$$

and the eigenvalues as

$$\mathcal{E} = \mathcal{E}_k^{(0)} + \lambda\mathcal{E}_k^{(1)} + \lambda^2\mathcal{E}_k^{(2)} + \cdots. \tag{6.5}$$

To solve the overall problem, we insert the two series (6.4) and (6.5) into the Schrödinger equation

$$H|k\rangle = \mathcal{E}_k|k\rangle. \tag{6.6}$$

When this is done, all the terms with the same coefficient $\lambda^n$ are grouped together, and must vanish together. The lowest three equalities are

$$\lambda^0 : H_0|k\rangle_0 = \mathcal{E}_k^{(0)}|k\rangle_0 \tag{6.7}$$

$$\lambda^1 : H_0|k\rangle_1 + V|k\rangle_0 = \mathcal{E}_k^{(0)}|k\rangle_1 + \mathcal{E}_k^{(1)}|k\rangle_0 \tag{6.8}$$

$$\lambda^2 : H_0|k\rangle_2 + V|k\rangle_1 = \mathcal{E}_k^{(0)}|k\rangle_2 + \mathcal{E}_k^{(1)}|k\rangle_1 + \mathcal{E}_k^{(2)}|k\rangle_0. \tag{6.9}$$

The first equation (6.7) is obviously just the appropriate (6.3) for the unperturbed solutions. The second equation (6.8) gives the basis for *first-order* perturbation theory.

The initial corrections to the basis functions and to the eigenvalues are found by working with (6.8). This equation may be rearranged, to give a more direct approach to the solution, as

$$(H_0 - \mathcal{E}_k^{(0)})|k\rangle_1 = (\mathcal{E}_k^{(1)} - V)|k\rangle_0. \tag{6.10}$$

The basic assumption is that the perturbed basis functions are only small deviations from the unperturbed ones, so that it is fruitful to expand the deviations in terms of the unperturbed functions as

$$|k\rangle_1 = \sum_j a_j |j\rangle_0. \tag{6.11}$$

Introducing this expansion into (6.10) gives us

$$\sum_j a_j (H_0 - \mathcal{E}_k^{(0)})|j\rangle_0 = (\mathcal{E}_k^{(1)} - V)|k\rangle_0. \tag{6.12}$$

The left-hand side vanishes for the single term $j = k$, which will leave $a_k$ undetermined. The first-order deviation of the wave function $|k\rangle_1$ is made up of small admixtures of the other members of the unperturbed basis set. Yet, we want the new wave function to remain normalized. To achieve this, we will require that $a_k = 0$ by definition. Then, using (6.3), equation (6.12) can be rewritten as

$$\sum_{j\neq k} a_j (\mathcal{E}_j^{(0)} - \mathcal{E}_k^{(0)})|j\rangle_0 = (\mathcal{E}_k^{(1)} - V)|k\rangle_0. \tag{6.13}$$

In order to evaluate the change in the eigenvalue, let us multiply both sides of (6.13) by the adjoint function ${}_0\langle i|$, and perform the implied integral. The resulting Kronecker delta functions allow us to perform the summation, and rearrange (6.13) to give

$$\begin{aligned}\mathcal{E}_k^{(1)}\delta_{ik} &= V_{ik} + \sum_j a_j (\mathcal{E}_j^{(0)} - \mathcal{E}_k^{(0)})\delta_{ij} \\ &= V_{ik} + a_i(\mathcal{E}_i^{(0)} - \mathcal{E}_k^{(0)})\end{aligned} \tag{6.14}$$

where

$$V_{ik} = {}_0\langle i|V|k\rangle_0 \tag{6.15}$$

is the *matrix element* between states $i$ and $k$. This single equation (6.14) produces two sets of answers for us. If $i = k$, the last term on the right-hand side vanishes, and

$$\mathcal{E}_k^{(1)} = V_{kk} \tag{6.16}$$

is the shift of the energy level. Similarly, if $i \neq k$, the left-hand side vanishes, and the coefficient of the admixture of the different basis functions is given by

$$a_i = \frac{V_{ik}}{\mathcal{E}_k^{(0)} - \mathcal{E}_i^{(0)}}. \tag{6.17}$$

The choice $a_k = 0$ is now seen also to avoid an inconvenient zero in the denominator of this equation. Thus, to first order, the eigenvalues are ($\lambda = 1$)

$$\mathcal{E}_k = \mathcal{E}_k^{(0)} + V_{kk} \tag{6.18}$$

and the perturbed basis functions are

$$|k\rangle = |k\rangle_0 + \sum_{j\neq k} \frac{V_{jk}}{\mathcal{E}_k^{(0)} - \mathcal{E}_j^{(0)}} |j\rangle_0. \tag{6.19}$$

Let us now proceed to consider the second-order corrections to the eigenvalues and the basis functions. Equation (6.9) can be rearranged as

$$(H_0 - \mathcal{E}_k^{(0)})|k\rangle_2 = (V_{kk} - V)|k\rangle_1 + \mathcal{E}_k^{(2)}|k\rangle_0 \tag{6.20}$$

where (6.18) has been introduced for the first-order perturbation of the energy levels. Again, the second-order perturbation of the wave function is expanded in terms of the unperturbed basis set as

$$|k\rangle_2 = \sum_{s\neq k} b_s |s\rangle_0 \tag{6.21}$$

where again we omit the term $b_k\ (= 0)$. Now, (6.21) and the last term of (6.19) are inserted into (6.20) to give

$$\sum_{s\neq k} b_s (H_0 - \mathcal{E}_k^{(0)})|s\rangle_0 = \sum_{j\neq k} \frac{V_{jk}(V_{kk} - V)}{\mathcal{E}_k^{(0)} - \mathcal{E}_j^{(0)}} |j\rangle_0 + \mathcal{E}_k^{(2)}|k\rangle_0. \tag{6.22}$$

Again, in order to evaluate the change in the eigenvalue, let us multiply both sides of (6.22) by the adjoint function ${}_0\langle i|$, and perform the implied integral. The resulting Kronecker delta functions allow us to perform the summation, and rearrange (6.22) to give

$$b_i(\mathcal{E}_i^{(0)} - \mathcal{E}_k^{(0)}) = \sum_{j\neq k} \frac{V_{jk}(V_{kk}\delta_{ij} - V_{ij})}{\mathcal{E}_k^{(0)} - \mathcal{E}_j^{(0)}} + \mathcal{E}_k^{(2)}\delta_{ik}. \tag{6.23}$$

There are several possibilities in this equation, but again it will determine all that we need to know. First, consider the case for which $i = k$. For this case, the left-hand side vanishes (particularly as we have directly omitted such a term from the left-hand summation prior to the last step), but the term in $V_{kk}$ in the numerator of the summation does not contribute ($j \neq k$), and the second-order change in the energy is just

$$\mathcal{E}_k^{(2)} = \sum_{j\neq k} \frac{V_{kj}V_{jk}}{\mathcal{E}_k^{(0)} - \mathcal{E}_j^{(0)}}. \tag{6.24}$$

When $i \neq k$, the last term on the right-hand side of (6.23) does not contribute, and it is straightforward now to determine the expansion coefficient for the wave functions. Some care must be taken with the diagonal terms (they are of the same order of magnitude and may cancel, especially when the first-order energy perturbation does not depend upon the index). This gives the final result as

$$b_i = \sum_{j \neq k} \frac{V_{ij} V_{jk}}{[E_k^{(0)} - E_j^{(0)}][E_k^{(0)} - E_i^{(0)}]} - \frac{V_{jk} V_{kk}}{[E_k^{(0)} - E_j^{(0)}]^2} \delta_{ij} \qquad i \neq k. \tag{6.25}$$

Thus, to second order in the perturbation, the energies are ($\lambda = 1$)

$$\mathcal{E}_k = \mathcal{E}_k^{(0)} + V_{kk} + \sum_{j \neq k} \frac{V_{kj} V_{jk}}{\mathcal{E}_k^{(0)} - \mathcal{E}_j^{(0)}} \tag{6.26}$$

and the basis functions are

$$\begin{aligned} |k\rangle = |k\rangle_0 &+ \sum_{j \neq k} \frac{V_{jk}}{\mathcal{E}_k^{(0)} - \mathcal{E}_j^{(0)}} |j\rangle_0 \\ &+ \sum_{\substack{j \neq k \\ i \neq k}} \frac{(V_{ij} - V_{kk}\,\delta_{ij}) V_{jk}}{(\mathcal{E}_k^{(0)} - \mathcal{E}_j^{(0)})(\mathcal{E}_k^{(0)} - \mathcal{E}_i^{(0)})} |i\rangle_0 . \end{aligned} \tag{6.27}$$

This procedure can obviously be extended to many higher orders of perturbation. In general, each higher order adds another summation over an intermediate state. For example, in (6.27), the first-order correction just couples two states, with a summation over all other states that couple to the $k$th one. At second order, however, this coupling generally goes through an intermediate state, with an additional ratio of a matrix element to an energy difference (in the denominator). Thus, at third order, it would be expected that an additional summation would appear as the connection moves through two intermediate states. In the world of perturbation theory, it is of course quite important to be sure that the correction terms are indeed quite small. If they are not, it is possible that the series does not converge, which means the entire approach violates the assumptions and is invalid. It is usual to check the results at an order above the desired one to make sure that the solution is stable. Otherwise other approaches must be pursued.

In the above discussion, it has been assumed that the eigenvalues $\mathcal{E}_k^{(0)}$ are all discrete, or at least all different if the spectrum is continuous. This may not always be the case and there may be several members of the set of wave functions that have the same energy. Then, the approaches described above must be modified to 'lift' this degeneracy prior to proceeding with the perturbation series in order to avoid unwanted zeros in the denominators of the summations. Let us consider the case where there are several different wave functions that all have the same energy value $\mathcal{E}_k^{(0)}$. The correct wave functions in the zero-order approximation are then some linear combinations of these wave functions

$$c_k |k\rangle_0 + c_{k'} |k'\rangle_0 + c_{k''} |k''\rangle_0 + \cdots . \tag{6.28}$$

For the constants $c_k$, it suffices to take their zero-order values as the solutions of the process. We now will write out a set of equations with $i = k, k', k''$, etc and substitute in them, as the first approximation, $\mathcal{E}_k = \mathcal{E}_k^{(0)} + \mathcal{E}_k^{(1)}$ obtained from the $H|k\rangle_0$ term in (6.10), which then gives

$$\mathcal{E}_k^{(1)} c_k = \sum_{k'} V_{kk'} c_{k'}. \tag{6.29}$$

Here, $k$ and $k'$ take on all the possible values that span the set of degenerate wave functions. This leads to a system of homogeneous equations

$$\sum_{k'} (V_{kk'} - \mathcal{E}_k^{(1)} \delta_{kk'}) c_{k'} = 0. \tag{6.30}$$

This system of homogeneous equations has solutions for the various values of the coefficients $c_k$ if the determinant vanishes:

$$|V_{kk'} - \mathcal{E}_k^{(1)} \delta_{kk'}| = 0. \tag{6.31}$$

If there are $n$ degenerate eigenvalues, then the equation (6.31) is of order $n$. The $n$ roots of the equation then give the values of $\mathcal{E}_k^{(1)}$. The equation (6.31) is called the *secular equation*. We note that the sum of the various new eigenvalue corrections is given by the sum of the diagonal matrix elements, a general result in degenerate perturbation theory. This process can be continued to second order if the degeneracy is not fully lifted by the first-order treatment.

## 6.2 Some examples of perturbation theory

In this section, we now want to consider a series of examples that will serve to illustrate the stationary perturbation technique. We will first consider a trapezoidal potential well, one with a linear potential term. Then we turn to a shifted harmonic oscillator, again obtained as a result of adding a linear potential to the harmonic oscillator. Next, we consider two coupled quantum wells, a system that we first treated in section 2.7. Finally, we treat the *scattering* of a plane wave by a Coulomb potential, a problem important to impurity scattering in semiconductors.

### 6.2.1 The Stark effect in a potential well

The system that we consider is that shown in figure 6.1, where we have an infinitely deep potential well (one with infinitely high barriers, corresponding to section 2.4), but with a linear potential existing within the well. Here, we take the unperturbed Hamiltonian as the quantity

$$H_0 = \frac{p^2}{2m} + V_0 \tag{6.32}$$

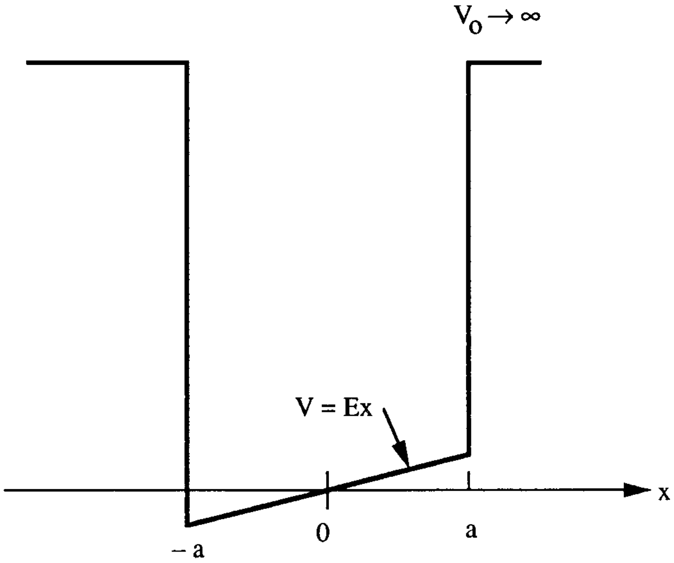


**Figure 6.1.** The trapezoidal well, in which a linear potential appears at the bottom.

where

$$V_0 = 0 \quad \text{for } |x| \leq a \quad \text{and} \\ \rightarrow \infty \quad \text{elsewhere.} \tag{6.33}$$

For this value of the unperturbed Hamiltonian, the eigenvalues and eigenfunctions are found from the solutions of section 2.4 to be

$$\psi_n(x) = \begin{cases} \dfrac{1}{\sqrt{a}} \sin\left[\dfrac{n\pi}{2a}(x+a)\right] & |x| \leq a \\ 0 & \text{elsewhere} \end{cases} \tag{6.34$a$}$$

$$\mathcal{E}_n = \frac{n^2\pi^2\hbar^2}{8ma^2}. \tag{6.34$b$}$$

The perturbing Hamiltonian is then the potential

$$V = eEx \tag{6.35}$$

and the entire problem is reduced to determining the matrix elements of this potential between the various basis states. There are two sets to be determined, the diagonal ones and the off-diagonal ones. First, the diagonal matrix elements are simply

$$V_{nn} = \langle n|V|n\rangle = \langle n|eEx|n\rangle$$

$$
\begin{aligned}
&= \frac{eE}{a}\int_{-a}^{a} x\sin^2\left[\frac{n\pi}{2a}(x+a)\right]\,\mathrm{d}x \\
&= \frac{eE}{2a}\int_{-a}^{a} x\left\{1-\cos\left[\frac{n\pi}{a}(x+a)\right]\right\}\,\mathrm{d}x = 0.
\end{aligned}
\tag{6.36}
$$

This result is a particular result of choosing the potential to be symmetrical about the centre of the well, so that the average under the perturbing potential is zero. Thus, there is no first-order shift in the energy levels themselves. In other words, the particular choice of the potential results in the centre of the well not moving, which means that, on average, the energy levels are not shifted to first order. The off-diagonal terms, however, are non-zero, and we find that

$$
\begin{aligned}
V_{ij} &= \langle i|V|j\rangle = \langle i|eEx|j\rangle \\
&= \frac{eE}{a}\int_{-a}^{a} x\sin\left[\frac{i\pi}{2a}(x+a)\right]\sin\left[\frac{j\pi}{2a}(x+a)\right]\,\mathrm{d}x \\
&= \frac{eE}{2a}\int_{-a}^{a} x\left\{\cos\left[\frac{(i-j)\pi}{2a}(x+a)\right]-\cos\left[\frac{(i+j)\pi}{2a}(x+a)\right]\right\}\,\mathrm{d}x \\
&= \begin{cases} -\dfrac{16eEa}{\pi^2}\dfrac{ij}{(i^2-j^2)^2} & (i\pm j)\ \text{odd} \\ 0 & (i\pm j)\ \text{even.} \end{cases}
\end{aligned}
\tag{6.37}
$$

The last result tells us that the odd potential must mix only those states that have different parity. Thus, the odd potential produces matrix elements only between an even-parity state and an odd-parity state. (The plus/minus sign just reminds us that the if the difference between two integers is odd, then the sum of these same two integers is also odd, and thus the two terms that result from the two integrals in (6.37) have been combined in the final result.)

Normally, the addition of an electric field produces a shift of the energy levels, an effect called the Stark effect. Here, because the potential has been taken to be an anti-symmetric one, the linear shift in the energy levels vanishes, and there is no first-order Stark effect. However, this is merely because of the choice of the reference level for the potential. If we had taken the zero of energy to lie at the point $x = -a$, then all the energy levels would have been pushed upward by the value of the potential at the centre of the well. We can see this by taking the potential as shown in figure 6.2. Now, the basis functions are given by

$$
\psi_n(x) = \frac{1}{\sqrt{a}}\sin\left[\frac{n\pi x}{2a}\right]
\tag{6.38}
$$

although the energy levels do not change (we are only moving the reference in the axis, and the well does not change when the perturbation is not present). However, the symmetry of the perturbing potential has been changed. Now, the diagonal elements become

$$
V_{nn} = \langle n|V|n\rangle = \langle n|eEx|n\rangle
$$

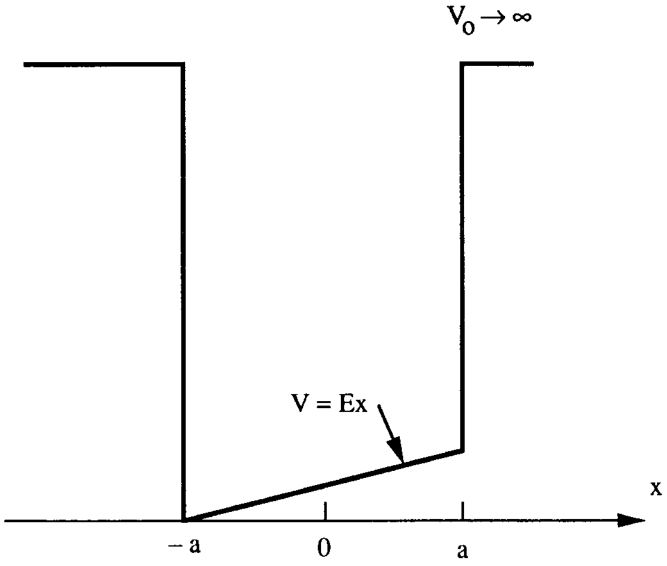


**Figure 6.2.** A trapezoidal well that is not anti-symmetrical.

$$
\begin{aligned}
&= \frac{eE}{a}\int_{-a}^{a} x\sin^2\left[\frac{n\pi x}{2a}\right]\,\mathrm{d}x \\
&= \frac{eE}{2a}\int_{-a}^{a} x\left\{1-\cos\left[\frac{n\pi x}{a}\right]\right\}\,\mathrm{d}x = eEa. \qquad (6.39)
\end{aligned}
$$

Now, the energy levels are all shifted to first order by the same amount $eEa$, which is the potential at the centre of the well, as suggested above. This is the linear Stark shift of the energy levels. However, the off-diagonal elements do not change from (6.37), which is a reflection of the fact that both energy levels shift, but their difference does not (the linear Stark shift is the same for all energy levels). It may be noted that the denominator of (6.37), other than for associated constants, is the difference (squared) of two energy levels.

Thus, whether or not we see a uniform shift of the energy levels in a perturbation approach often depends upon the reference level chosen for the overall potential. However, this does not appear in the off-diagonal matrix elements. This latter is significant, as it is only the off-diagonal matrix elements that appear in the summations for the various orders of perturbation theory (the term in (6.27) involving the difference between two diagonal matrix elements is identically zero due to the fact that all of these diagonal elements are equal).

### 6.2.2 The shifted harmonic oscillator

As a second example of the application of perturbation theory, we want to consider a harmonic oscillator that is subjected to a linear potential that can arise from an applied electric field. Then, the Hamiltonian can be written as

$$\begin{aligned} H &= \frac{p^2}{2m} + \frac{1}{2}m\omega^2 x^2 + eEx \\ &= \hbar\omega(a^+ a + \tfrac{1}{2}) + eEx. \end{aligned} \tag{6.40}$$

In the last line, we have used the operator results of chapter 4. The application of the linear potential still produces a harmonic oscillator, but one that is shifted in both position and energy. This can be seen, as the first line of (6.40) can be rewritten as

$$H = \frac{p^2}{2m} + \frac{1}{2}m\omega^2(x+x_0)^2 - \frac{e^2E^2}{2m\omega^2} \tag{6.41}$$

where

$$x_0 = \frac{eE}{m\omega^2}. \tag{6.42}$$

The last term in (6.41) is the downward shift of the quadratic potential, while (6.42) is the shift of the centre of the potential well. This is shown in figure 6.3. There are now two ways to approach the solution of the new Hamiltonian in (6.40). In the first approach, the linear potential is taken to be the perturbing potential, and the wave functions of the harmonic oscillator centred at $x = 0$ are used. In the second approach, the perturbing potential is the uniform shift of the harmonic oscillator downward by the constant energy term, which is the last term on the right-hand side of (6.41). The wave functions are those corresponding to the harmonic oscillator centred at $x = -x_0$. We take the last approach first.

If we treat the harmonic oscillator directly as a shifted harmonic oscillator with the minimum centred at $x = -x_0$, then the wave functions are merely shifted from those in (4.75) according to

$$\Psi_n = \frac{(a^+)^n}{\sqrt{n!}}\Psi_0 = \frac{1}{\sqrt{n!}}\left(\frac{m\omega}{2\hbar}\right)^{n/2}\left(x + x_0 - \mathrm{i}\frac{p}{m\omega}\right)^n \Psi_0(x+x_0) \tag{6.43}$$

where the creation operator is now

$$a^+ = \left(\frac{m\omega}{2\hbar}\right)^{1/2}\left(x + x_0 - \mathrm{i}\frac{p}{m\omega}\right). \tag{6.44}$$

The matrix elements are now

$$V_{ij} = -\langle i|\frac{e^2E^2}{2m\omega^2}|j\rangle = -\frac{e^2E^2}{2m\omega^2}\langle i|j\rangle = -\frac{e^2E^2}{2m\omega^2}\,\delta_{ij}. \tag{6.45}$$

Thus, only the diagonal elements are non-zero. *A constant shift of the Hamiltonian provides only a uniform and constant shift of all energy levels. It*

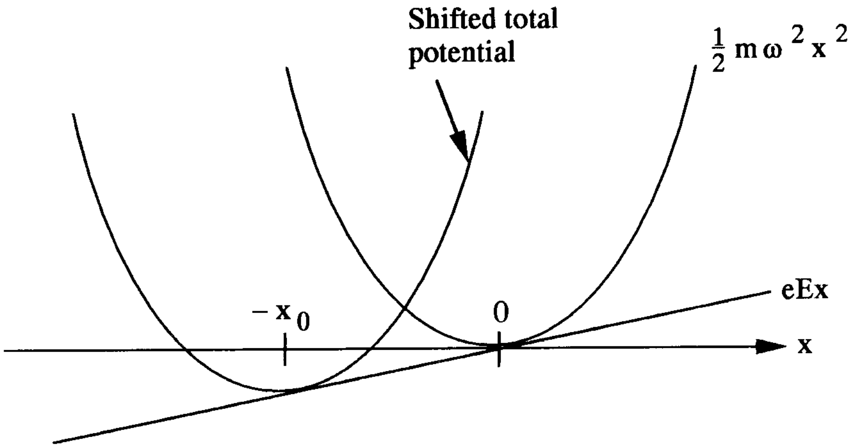


**Figure 6.3.** The combination of a quadratic potential and a linear potential produces a shifted quadratic potential.

*provides no mixing of the states*. Thus, when the linear potential is applied to the harmonic oscillator, the only result is a shift downward of all energy levels by the shift in the minimum, and a shift of the centroid of the wave functions to the new centre of the harmonic oscillator.

The first approach to the solution should yield a similar answer. For this, we can now compute the matrix elements using the linear potential as the perturbation. Then, the matrix elements can be computed using (4.80$a$) and (4.79) as

$$\begin{aligned} V_{ij} &= \langle i|eEx|j\rangle = eEx_{ij} = eE\sqrt{\frac{\hbar}{2m\omega}}(a+a^{+})_{ij} \\ &= eE\sqrt{\frac{\hbar}{2m\omega}}(\sqrt{i}\delta_{j+1,i} + \sqrt{i+1}\delta_{j-1,i}). \end{aligned} \tag{6.46}$$

In this approach, there is no first-order shift of the energy, but the second-order shift involves only two terms in the summation. In each term, the product of matrix elements connects a given level only with its neighbours just above and just below, so only one of the two delta functions in (6.46) appears in each matrix element. We compute the second-order shift to be (the signs arise from the energy denominators)

$$\mathcal{E}_i^{(2)} = (eE)^2\frac{\hbar}{2m\omega}\left[\frac{-(i+1)+i}{\hbar\omega}\right] = -\frac{e^2E^2}{2m\omega^2}. \tag{6.47}$$

This produces the full shift of the energy levels, and it also is independent of the value of the energy level under consideration. Thus, higher-order corrections to the energy levels should vanish, and this can be checked by examining the fourth-order term, which is the next one expected. However, it cannot always be ensured that the higher-order terms vanish, but rational thinking for this problem suggests that they should sum to zero, when all orders are included. The corrections to the wave functions are just those that try to move the centroid to the new centre of the harmonic potential. This is a perfect example of a case where low-order perturbation theory does not produce a result that is close to the exact answer (at least for the wave functions), which we know from being able to solve the problem exactly. The first-order shift in the wave functions can also be obtained by using the matrix elements above. For the lowest energy level,

$$|0\rangle^{(1)} = -\frac{eE}{\hbar\omega}\sqrt{\frac{\hbar}{2m\omega}}|1\rangle^{(0)} \tag{6.48}$$

and for the higher levels, we obtain

$$|i\rangle^{(1)} = -\frac{eE}{\hbar\omega}\sqrt{\frac{\hbar}{2m\omega}}\left[\sqrt{i}|i-1\rangle^{(0)} - \sqrt{i+1}|i+1\rangle^{(0)}\right] \qquad i > 0. \tag{6.49}$$

To obtain the exact wave functions, we will need to sum to higher orders.

### 6.2.3 Multiple quantum wells

We now want to turn to a consideration of coupled quantum wells, which we will take each to be quadratic in nature. This treatment provides us with a discussion of *degenerate perturbation theory*, a process in which we decide how two levels that otherwise would have the same energy interact to cause a small splitting. This was discussed at the end of section 6.1. Consider, for example, two harmonic wells that are displaced from one another by a distance $d$, as shown in figure 6.4. Here, the unperturbed wave functions in each well are at the same energy, so the degeneracy is between wave functions centred in each of the two wells. The interaction between the two causes only a small barrier between the two wells, so the wave function of an electron localized in one of the wells actually extends somewhat into the other well (see, e.g., section 2.7). Thus, the two wave functions overlap with each other and there is an interaction between them. It is this interaction that causes a modification of the two wave functions, lifts the degeneracy and leads to two new levels, which are then separated slightly in energy. One of the levels has a wave function composed of the constructive interference of the two individual wave functions, while the second (the higher energy state) has a wave function that is anti-symmetric in the combined wells and thus arises from the destructive interference of the two individual wave functions.

In each of the two wells, the unperturbed problem is just that of a harmonic oscillator, centred at the particular well, so the Hamiltonian and lowest wave function of the well at $x = 0$ are given by

$$H_{0,1} = \frac{p^2}{2m} + \frac{1}{2}m\omega^2 x^2$$

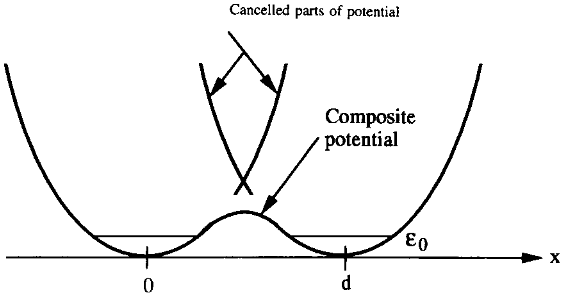


**Figure 6.4.** The composite potential that arises from two overlapping harmonic potentials that are displaced slightly.

$$H_{0,1}|0(0)\rangle^{(0)} = \mathcal{E}_0|0(0)\rangle^{(0)}. \tag{6.50}$$

Similarly, those at the second well are given by

$$\begin{aligned} H_{0,2} &= \frac{p^2}{2m} + \frac{1}{2}m\omega^2(x-d)^2 \\ H_{0,2}|0(d)\rangle^{(0)} &= \mathcal{E}_0|0(d)\rangle^{(0)}. \end{aligned} \tag{6.51}$$

We note that both wave functions, one centred in each well, have the same eigenvalues for the unperturbed energy. Thus, these two eigenvalues are said to be *degenerate*. The task here is to use perturbation theory to lift this degeneracy. We are not so much interested in how the interaction with higher levels changes the energy levels as we are with how these two wave functions interact with each other to change the energy levels. While we work with the energy level that we have indicated is the lowest isolated well level, this is not a limitation, and we could work with any of the levels of the isolated well that lie below the central peak between the two wells (for those levels above this peak, perturbation theory is not appropriate). Our interest then is in the matrix element

$$V_{00} = {}^{(0)}\langle 0(0)|V|0(d)\rangle^{(0)} = \int_{-\infty}^{\infty} \Psi_0^*(x)V(x)\Psi_0(x-d)\,\mathrm{d}x. \tag{6.52}$$

Here, the potential is the actual total potential in the problem, but the main part is that part of the potential that lies between the two wells. Equation (6.52) is the *overlap* integral defining the *mixing* of the two wave functions.

The composite wave functions should be able to be written as a sum of the two individual wave functions, as

$$|0\rangle = A|0(0)\rangle^{(0)} + B|0(d)\rangle^{(0)}. \tag{6.53}$$

When we operate first with the conjugate of the wave function of the left-hand well, and then with the conjugate of the wave function of the right-hand well, we produce two equations. This is done by neglecting the variation of the potential difference from the single well for the diagonal terms (wave functions on the same sites). The process produces the resulting matrix equation

$$\begin{bmatrix} E_0 & V_{00} \\ V_{00}^* & E_0 \end{bmatrix} \left. \begin{matrix} A \\ B \end{matrix} \right] = E \left. \begin{matrix} A \\ B \end{matrix} \right] . \tag{6.54}$$

If this is to have a solution, then we must arrange for the determinant

$$\begin{vmatrix} (\mathcal{E}_0 - \mathcal{E}) & V_{00} \\ V_{00}^* & (\mathcal{E}_0 - \mathcal{E}) \end{vmatrix} = 0. \tag{6.55}$$

This leads to the new energy levels

$$\mathcal{E} = \mathcal{E}_0 \pm |V_{00}|. \tag{6.56}$$

Thus, the two individual energy levels $\mathcal{E}_0$, which were degenerate, are now split by the amount $2|V_{00}|$. It is easy then to determine that the additive overlap of the two individual wave functions gives the lower energy level, while the subtractive overlap of the two individual wave functions gives the upper energy level. The lower is said to be a bonding wave function, while the upper is an anti-symmetric anti-bonding wave function.

The above treatment can be extended to three potential wells, under the assumption that only the two nearest neighbours interact. This then leads to a 3 × 3 matrix equation, for which the solution is given by the determinant equation

$$\begin{vmatrix} |(\mathcal{E}_0 - \mathcal{E}) & V_{00} & 0 \\ V_{00}^* & (\mathcal{E}_0 - \mathcal{E}) & V_{00} \\ 0 & V_{00}^* & (\mathcal{E}_0 - \mathcal{E}) \end{vmatrix} = 0. \tag{6.57}$$

This leads to the equation

$$(\mathcal{E}_0 - \mathcal{E})^3 - 2V_{00}V_{00}^*(\mathcal{E}_0 - \mathcal{E}) = 0 \tag{6.58}$$

which gives the eigenvalues

$$\mathcal{E} = \mathcal{E}_0 \qquad \mathcal{E}_0 \pm \sqrt{2}|V_{00}|. \tag{6.59}$$

The lowest energy level has the additive sum of the three individual wave functions, while the other two levels are various other combinations. It is important to note that the resulting wave functions all have contributions in each of the three potential wells and are no longer localized in a single well.

The result (6.59) illustrates another important point of the degenerate perturbation approach. The band widths in (6.59) and in (6.56) differ. The difference between the highest energy level and the next lowest is smaller, however. This is true in nearly all such near-neighbour systems. As the number of wells builds up, the *density* of the energy levels increases, until it is almost a continuum—an energy band as it were. This is how the energy bands in solids are formed. The width of the so-called energy band does not increase without limit, but approaches a limiting value as the number of wells increases without limit. In terms of the matrix element $V_{00}$, the 'band width' for two, three, and four potential wells is 2, 2.828, 3.236. In fact, the limiting value is $4V_{00}$, which is a property of the matrix rather than the physics.

### 6.2.4 Coulomb scattering

Here, we will discuss the scattering of an incoming plane wave, which varies as $\mathrm{e}^{\mathrm{i}kz}$, by a charged atom through the Coulomb potential. This atom can be considered to be an impurity atom in the host semiconductor. In any treatment of electron scattering from a Coulomb potential, it is necessary to consider the long-range nature of the potential. If the interaction is summed over all space, the integral diverges and a cutoff mechanism must be invoked to limit the integral. One approach is just to cut off the integration at the mean impurity spacing, the so-called Conwell–Weisskopf (1950) approach. A second approach is to invoke screening of the Coulomb potential by the free carriers. In this case, the potential is induced to fall off much more rapidly than a bare Coulomb interaction, due to the Coulomb forces from surrounding carriers (such as the one creating the incident plane wave). The screening is effective over a distance on the order of the Debye screening length in non-degenerate materials. This screening of the repulsive Coulomb potential results in an integral for the scattering cross section which converges without further approximations (Brooks 1955).

For spherical symmetry about the scattering centre, or ion location, the potential is screened in a manner that gives rise to the potential

$$\Phi(\boldsymbol{r}) = \frac{e^2}{4\pi\varepsilon_\infty r}\mathrm{e}^{-q_{\mathrm{D}}r} \tag{6.60}$$

where the Debye wave vector $q_{\mathrm{d}}$ is the inverse of the screening length, and is given by

$$q_{\mathrm{D}}^2 = \frac{ne^2}{\varepsilon_\infty k_{\mathrm{B}}T}. \tag{6.61}$$

Here $\varepsilon_\infty$ is the high-frequency permittivity and $n$ is the density of charge carriers in the semiconductor of interest.

In treating the scattering from this screened Coulomb potential, we use a wave scattering approach and compute the *scattering cross section* $\sigma(\theta)$, which gives the angular dependence of the scattering. This is the perturbation theory for a continuum of energy (or momentum) states as represented by the continuous variable $k$. It is assumed the incident wave is a plane-wave, and the scattered wave is also a plane-wave. The total wave function is written as

$$\Psi(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}kz} + v(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}}. \tag{6.62}$$

Here, $\boldsymbol{k} = k\boldsymbol{a}_z$ orients the incident wave along the polar axis in a spherical coordinate system, and the second term represents the scattered wave. That is, the exponential represents the final state of the perturbation, and the coefficient $v(r)$ represents the matrix element coupling this state to the initial state. Equation (6.62) is inserted into the Schrödinger equation, neglecting terms of second or higher order in the scattered wave, in keeping with first-order perturbation theory, and

$$\nabla^2 v(\boldsymbol{r}) + k'^2 v(\boldsymbol{r}) = \frac{2m^*}{\hbar^2}\frac{e^2}{4\pi\varepsilon_\infty r}\mathrm{e}^{-q_\mathrm{D} r}\mathrm{e}^{\mathrm{i}kz}. \tag{6.63}$$

If the terms on the right-hand side are treated as a charge distribution, the normal results from electromagnetic field theory can be used to write the solution as

$$v(\boldsymbol{r}) = -\frac{m^* e^2}{8\pi^2\varepsilon_\infty}\int \frac{\mathrm{d}^3\boldsymbol{r}'}{r'|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{e}^{\mathrm{i}kz' - q_\mathrm{D} r'}\,\mathrm{e}^{\mathrm{i}k'|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{6.64}$$

To proceed, it is assumed that $r \gg r'$, and the polar axis in real space is taken to be aligned with $r$. Further, the scattering wave vector is taken to be $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'$, so that

$$\int_0^\pi \sin\theta\,\mathrm{d}\theta\,\mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}'} = \frac{\sin(qr')}{qr'} \tag{6.65}$$

the $\phi$ integration is simple and can be done quickly, while the remaining integration becomes

$$v(\boldsymbol{r}) \cong -\frac{m^* e^2}{2\pi\varepsilon_\infty\hbar^2 qr}\int_0^\infty \sin(qr')\mathrm{e}^{-q_\mathrm{D} r'}\,\mathrm{d}r' = \frac{m^* e^2}{2\pi\varepsilon_\infty\hbar^2 qr(q^2+q_\mathrm{D}^2)}. \tag{6.66}$$

Now $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'$, but here we assume that $k = k'$, and $q = 2k\sin(\theta/2)$, where $\theta$ is the angle between $\boldsymbol{k}$ and $\boldsymbol{k}'$. The assumption is that the scattering does not change the energy of the carrier, but only changes its direction. If we write the scattered wave function $v(\boldsymbol{r})$ as $f(\theta)/r$, then we recognize that the factor $f(\theta)$ is the matrix element, and the cross section is defined as

$$\sigma(\theta) = |f(\theta)|^2 = \left(\frac{m^* e^2}{8\pi\hbar^2 k^2\varepsilon_\infty}\right)^2 \frac{1}{[\sin^2(\theta/2) + q_\mathrm{D}^2/4k^2]^2}. \tag{6.67}$$

The total scattering cross section (*for the relaxation time*) is found by integrating over $\theta$. Thus

$$\begin{aligned}\sigma_\mathrm{c} &= 2\pi\int_0^\pi \sigma(\theta)\cos\theta\,\mathrm{d}\theta = 16\pi\int_0^{\pi/2}\sigma\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right)\,\mathrm{d}\left(\sin\frac{\theta}{2}\right)\\ &= \frac{\pi}{2}\left(\frac{m^* e^2}{2\pi\hbar^2 k^2\varepsilon_\infty}\right)^2\left[\frac{16k^4}{q_\mathrm{D}^2(4k^2+q_\mathrm{D}^2)}\right].\end{aligned} \tag{6.68}$$

The cross section describes the fraction of each incoming carrier that is lost to other wave states, and has the units of area (such as $\mathrm{cm}^2$). It can be connected to a scattering *rate*, which describes the number of scattering events per second, by multiplying by the number of scattering centres per unit volume and by the velocity of the incoming carrier (which converts a loss per centimetre into a loss per second). First, the loss per unit length of the incoming carrier to scattering events is just

$$\alpha(k) = N\sigma_{\mathrm{c}} = \frac{Ne^4 m^{*2}}{2\pi\varepsilon_\infty^2 \hbar^4 k^4}\left[\frac{4k^4}{(4k^2+q_{\mathrm{D}}^2)q_{\mathrm{D}}^2}\right]. \tag{6.69}$$

The scattering rate is now the product of the attenuation per unit length and the velocity of the carrier, or

$$\Gamma(k) = N\sigma_{\mathrm{c}}\boldsymbol{v} = \frac{Ne^4 m^{*}}{2\pi\varepsilon_\infty^2 \hbar^3 k^3}\left[\frac{4k^4}{(4k^2+q_{\mathrm{D}}^2)q_{\mathrm{D}}^2}\right]. \tag{6.70}$$

## 6.3 An alternative technique—the variational method

There are often times when the perturbation approach just does not work, because the basis set cannot be determined consistently with (6.3). An example is the triangular potential well of section 2.6. The problem with this particular barrier is that the solution of the Schrödinger equation is complicated and the wave functions are special functions that, in general, are evaluated numerically. Often, however, the exact eigenvalues and eigenfunctions are not required; rather, only good approximations to them are required. To accomplish this, another approximation technique is generally used—the Rayleigh–Ritz method, or variational method. Here, it may be assumed that we cannot begin with (6.3) or any equivalent form. Nevertheless, we may take an approximate wave function as an expansion in a convenient basis set:

$$\psi(x) = \sum_i a_i \psi_i(x). \tag{6.71}$$

It may be assumed then that, if this is an orthonormal basis set, the expectation value of the Hamiltonian is given by

$$\langle H\rangle = \sum_i \mathcal{E}_i |a_i|^2 \tag{6.72}$$

where we have constructed the expectation value by pre-multiplying by the conjugate of (6.71), post-multiplied by (6.71), integrated and used the orthogonality properties of the basis set. An inequality can be created by assuming that only the lowest energy level is used in the expansion (using the lowest energy level for each of the eigenstates), and

$$\langle H\rangle \geq \mathcal{E}_0 \sum_i |a_i|^2 = \mathcal{E}_0 \tag{6.73}$$

since the sum over all the basis states is unity to ensure the normalization of the wave function in (6.71).

The principle of the Rayleigh–Ritz method is that we can adopt a trial wave function, which may have some parameters in it, and then minimize the energy calculated with this wave function. This energy will still be above the actual lowest energy level $\mathcal{E}_0$. This gives a good approximation to the actual energy level. The next higher state can be developed by using a second parametrized basis function, which is made orthogonal to the first. The energy level of this wave function is then found by the same adjustment technique. Let us illustrate this with the actual triangular-potential-well problem of section 2.6.

To begin, we note that the wave function vanishes at the origin, as the potential is typically defined by

$$V(x) = eEx \quad \text{for } x > 0 \quad \text{and} \quad V(x) \to \infty \quad \text{for } x \leq 0. \tag{6.74}$$

Thus, we seek a trial function that vanishes at $x = 0$, and also vanishes for large $x$, where the energy lies at a lower level than the potential. For this, we take

$$|0\rangle = Ax\mathrm{e}^{-bx} \tag{6.75}$$

which satisfies both limiting requirements. The pre-factor coefficient is determined by normalization to be

$$\langle 0|0\rangle = 1 = \int_0^\infty A^2 x^2 \mathrm{e}^{-2bx}\,\mathrm{d}x = \frac{A^2}{4b^3} \tag{6.76}$$

or

$$A = 2b^{3/2}. \tag{6.77}$$

Then, the expectation value of the Hamiltonian is

$$\begin{aligned}\langle H\rangle &= 4b^3 \int_0^\infty x\mathrm{e}^{-bx}\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + eEx\right] x\mathrm{e}^{-bx}\,\mathrm{d}x \\ &= \frac{\hbar^2 b^2}{2m} + \frac{3eE}{2b}.\end{aligned} \tag{6.78}$$

This value is now above the actual lowest energy level, and we can minimize this value by setting the derivative with respect to $b$ to zero. This then leads to (we take the first level as the lowest level in keeping with the general results of section 2.6)

$$b = \left(\frac{3eEm}{2\hbar^2}\right)^{1/3} \tag{6.79$a$}$$

$$\mathcal{E}_0 \leq \frac{3\hbar^2}{2m}\left(\frac{3eEm}{2\hbar^2}\right)^{2/3}. \tag{6.79$b$}$$

This result should be compared with (2.83). The present result differs from (2.83) only by a factor of $3(2/3\pi)^2/3 \simeq 1.07$, or about a 7% error, which is quite small considering the approximations. As expected, the right-hand side of (6.79$b$) lies above the actual value, given by (2.83).

To compute the next level, we must assume a wave function that is orthogonal to (6.75). The wave function of (6.75) has a single maximum, decaying away to zero at the origin and for large values of $x$. The next level should have a wave function with two maxima, but also decaying away for large values of $x$ and at the origin. For this purpose, we write the first-excited-state wave function as

$$|1\rangle = a(x - cx^2)\mathrm{e}^{-fx}. \tag{6.80}$$

To begin, we make this wave function orthogonal to $|0\rangle$, through

$$\begin{aligned}\langle 0|1\rangle &= 2ab^{3/2}\int_0^\infty x(x - cx^2)\mathrm{e}^{-(b+f)x}\,\mathrm{d}x \\ &= 2ab^{3/2}\left[\frac{2}{(b+f)^3} - \frac{6c}{(b+f)^4}\right] = 0\end{aligned} \tag{6.81}$$

or

$$c = \frac{b+f}{3}. \tag{6.82}$$

With this new wave function, we can now compute the normalization:

$$\langle 1|1\rangle = 1 = \frac{a^2}{4f^5}(f^2 - 3cf + 3c^2) \tag{6.83}$$

or

$$a^2 = \frac{12f^5}{f^2 - bf + b^2}. \tag{6.84}$$

so, once we have evaluated $f$ by minimizing the energy level, we can compute the normalization as well. Actually, we will use the result in our evaluation of the energy, which is

$$\begin{aligned}\langle H\rangle = \frac{\hbar^2 f^2}{6m}&\left(1 + \frac{15f^2}{f^2 - bf + b^2}\right) \\ &+ \frac{eE}{2f}\left(1 + \frac{3b^2}{2(f^2 - bf + b^2)}\right).\end{aligned} \tag{6.85}$$

This can now be minimized, to yield the value of $f$ as above. While the tendency might be to ignore the second terms in each of the large parentheses, this would be a mistake, as it would lead to an energy level below the lowest level. These terms are actually the major terms. Unfortunately, this leads to a complicated, high-order algebraic equation for $f$ in terms of $b$. This can be solved most easily by numerical calculations, which lead to the result $f \simeq 0.73b$, which in turn leads to $\mathcal{E}_1 \simeq 2.4\mathcal{E}_0$.

A similar procedure can now be used to generate a variational wave function for the next higher eigenstate. A wave function, probably with a cubic term in the pre-factor, is assumed and the coefficients are adjusted first to make it orthogonal to each of the two lower eigenfunctions. Then, the energy is minimized to provide an estimate of the energy level. Of course, each successive higher level becomes more complicated to determine, and most efforts using this method are addressed only to the lowest energy level.

## References

Ando T, Fowler A and Stern F 1982 *Rev. Mod. Phys.* **54** 437
Brooks H 1955 *Adv. Electr. Electron Phys.* **8** 85
Conwell E M and Weisskopf V 1950 *Phys. Rev.* **77** 368
Schiff L I 1955 *Quantum Mechanics* 2nd edn (New York: McGraw-Hill)
Stern F and Howard W E 1967 *Phys. Rev.* **163** 816

## Problems

1. In an infinite square well, in which $V(x) \to \infty$ for $x < 0$ and $x > a$, $V(x) = 0$ for $0 \leq x \leq a$, the basis states are given by $|n\rangle = \sqrt{2/a}\sin(n\pi x/a)$, and $\mathcal{E}_n = \hbar^2\pi^2 n^2/(2ma^2)$. For the introduction of a perturbing potential of $V_1 = (x-a)^2/2$, calculate the changes in the wave functions up to first order and in the energy levels up to second order.

2. Determine the third-order correction to the energy in perturbation theory.

3. Determine the corrections to second order necessary to lift the degeneracy of a set of states in an arbitrary perturbing potential.

4. Except for certain accidents, degeneracy usually does not arise in a one-dimensional problem. Degeneracy is usually found when the dimension is raised. Consider the two-dimensional infinite square well

$$V(x,y) = \begin{cases} 0 & \text{for } |x| \leq a/2 \text{ and } |y| \leq a/2 \\ \to \infty & \text{for } |x| \geq a/2 \text{ or } |y| \geq a/2. \end{cases}$$

Using this potential, separate the Schrödinger equation into two one-dimensional equations and solve for the allowed energy values using the fact that the potential can be written as the sum of two one-dimensional potentials. The solutions can then be written down using the results of chapter 2. Consider the application of the simple perturbation $V_1(x,y) = V_0$ for $|x| \leq a/4$ and $|y| \leq a/4$, and is zero elsewhere (the perturbing potential exists only in the centre of the well). Compute the splitting of the lowest two unperturbed energy levels.

5. Using the harmonic oscillator raising and lowering operators, compute directly the matrix element for $\langle n|x^4|n\rangle$. Then, introducing a complete set of states (as a resolution of the delta function) break this into two products $x^2x^2$, and compute the matrix element from

$$\langle n|x^4|n\rangle = \sum_j \langle n|x^2|j\rangle\langle j|x^2|n\rangle.$$

6. A one-dimensional harmonic oscillator is perturbed by a term $\alpha x^3$. Calculate the change in each energy level to second order.

7. Calculate the lowest energy level of a harmonic oscillator using the Gaussian wave function as a variational wave function. Let the width of the Gaussian be the parameter that is varied.

8. To compute a momentum relaxation time, a factor of $(1 - \cos\theta)$ is included in the integral (6.68). Plot the results with this inclusion versus the results that result from (6.68) directly—plot the scattering rate versus the energy.

# Chapter 7

# Time-dependent perturbation theory

In many physical systems, we are interested in small time-dependent changes in the state. The approach used in the last chapter cannot deal with this situation, and we must develop perturbation theory further. This will be done in the present chapter. In fact, we will develop the approach twice, first for the Schrödinger representation and then for the Heisenberg representation in a linear vector space. This double approach is followed for two reasons. First, it is important to grasp the concept of the time-dependent perturbation approach in a direct method, without the encumbrance of the mathematics that accompanies the latter approach—the interaction representation. Second, it is important to learn the mathematics of the interaction representation, but in doing so it is quite helpful to understand just where we are heading. The double approach achieves both objectives.

In time-dependent perturbation theory, we continue to assume that the Hamiltonian can be split into two parts

$$H = H_0 + H_1 \tag{7.1}$$

in which the first part $H_0$ can be solved directly and the second part $H_1$ is small. When this can be done, then the second part can be initially ignored, and the problem solved directly to give the wave function, or its expansion in a natural basis function set. After this is done, the second term in (7.1) is treated as a perturbation, and the eigenvalues and eigenfunctions are then developed in a time-dependent perturbation approach in terms of the natural basis function set.

In this chapter, we will first develop the approach in the Schrödinger representation and consider an example due to a harmonic potential. Then the interaction representation will be developed and the perturbation theory redeveloped in the Heisenberg representation. Our attention will then be turned to an extended approach in which the initial state of the system has an exponential decay. It is this latter approach that allows us to return once again to the idea of an energy–time relationship as a result of the lifetime of the initial state. Finally, the scattering matrix approach (the $T$-matrix) is developed. One note of importance is that it will be assumed that the total Hamiltonian is a *time-independent* quantity, while the perturbation term $H_1$ will be allowed to have an oscillatory nature.

## 7.1 The perturbation series

If the Schrödinger equation is used with a time-independent and unperturbed Hamiltonian ($H_1 = 0$), for which it may readily be solved, the solution may be written from (2.93) as (the position coordinate is suppressed)

$$\Psi^{(0)}(t) = \sum_n c_n \mathrm{e}^{-\mathrm{i}\omega_n t}\psi_n \tag{7.2}$$

where

$$H_0\psi_n = \mathcal{E}_n\psi_n \qquad \omega_n = \mathcal{E}_n/\hbar \tag{7.3}$$

and

$$c_n = (\psi_n, \Psi^{(0)}(0)). \tag{7.4}$$

The determination of the coefficients has been done at $t = 0$ for ease of notation.

In the presence of the perturbation $H_1$, we seek a solution that has the same general form as (7.2)–(7.4), and this can be achieved by making the assumption that the perturbation $H_1 = V$ is small and produces a slow variation in the coefficients $c_n$. Then, we may write the new solution as

$$\Psi(t) = \sum_n c_n(t)\mathrm{e}^{-\mathrm{i}\omega_n t}\psi_n. \tag{7.5}$$

The problem is now to determine the time variation of the coefficients. To proceed, the assumed solution (7.5) is introduced into the Schrödinger equation

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = (H_0 + V)\Psi. \tag{7.6}$$

This leads to

$$\begin{aligned}\sum_n c_n(t)\mathcal{E}_n\mathrm{e}^{-\mathrm{i}\omega_n t}\psi_n + \mathrm{i}\hbar\sum_n \frac{\mathrm{d}c_n}{\mathrm{d}t}\mathrm{e}^{-\mathrm{i}\omega_n t}\psi_n \\ = \sum_n H_0 c_n(t)\mathrm{e}^{-\mathrm{i}\omega_n t}\psi_n + \sum_n V c_n(t)\mathrm{e}^{-\mathrm{i}\omega_n t}\psi_n.\end{aligned} \tag{7.7}$$

Here, we shall assume that the potential $V$ is time varying as

$$V = V(t) = V_0\mathrm{e}^{\mathrm{i}\omega_0 t}. \tag{7.8}$$

The first term on the left and the first term on the right drop out by virtue of (7.3). We now multiply through by an arbitrary basis function $\psi_k^*$, and integrate over all space, using the orthonormality of these functions. This leads to

$$\mathrm{i}\hbar\frac{\mathrm{d}c_k}{\mathrm{d}t}\mathrm{e}^{-\mathrm{i}\omega_k t} = \sum_n V_{kn}c_n(t)\mathrm{e}^{-\mathrm{i}\omega_n t+\mathrm{i}\omega_0 t} \tag{7.9}$$

where

$$V_{kn} = \langle k|V_0|n\rangle \tag{7.10}$$

is the matrix element. On introducing the difference frequency

$$\omega_{kn} = \frac{\mathcal{E}_k - \mathcal{E}_n}{\hbar} \tag{7.11}$$

equation (7.8) can be rewritten as

$$\mathrm{i}\hbar\frac{\mathrm{d}c_k}{\mathrm{d}t} = \sum_n V_{kn}c_n(t)\mathrm{e}^{\mathrm{i}\omega_{kn}t+\mathrm{i}\omega_0 t}. \tag{7.12}$$

What we now have is a complicated set of equations for the coefficients and their time dependence. In general, this can be solved by matrix techniques when the matrix elements are known.

The approach that we want to follow, however, assumes that the perturbation is very small. Moreover, we will also assume that the system is in one single eigenstate at the initial time, which will be taken as $t = 0$:

$$c_s(0) = 1 \qquad c_k(0) = 0 \quad \text{for all } k \neq s. \tag{7.13}$$

To be consistent with this assumption, we also will assume that the perturbation $V$ is zero for $t < 0$. In (7.10), it has been assumed that the perturbing potential is constant, but it would be easy to change this and to include a perturbing potential varying with a particular frequency. This frequency factor would then appear with the matrix element. We will continue with the time-independent approach, but the reader should be aware that the inclusion of the time variation can easily be incorporated, as has been done. With the above approximations, (7.12) now becomes a simple equation

$$\mathrm{i}\hbar\frac{\mathrm{d}c_k}{\mathrm{d}t} = V_{ks}c_s(0)\mathrm{e}^{\mathrm{i}\omega_{ks}t+\mathrm{i}\omega_0 t} = V_{ks}\mathrm{e}^{\mathrm{i}\omega_{ks}t+\mathrm{i}\omega_0 t}. \tag{7.14}$$

This can now be easily solved to give

$$\begin{aligned} c_k &= -\frac{\mathrm{i}}{\hbar}\int_0^t V_{ks}\mathrm{e}^{\mathrm{i}\omega_{ks}t'+\mathrm{i}\omega_0 t'}\,\mathrm{d}t' \\ &= -\frac{1}{\hbar\omega_{ks}+\hbar\omega_0}V_{ks}(\mathrm{e}^{\mathrm{i}\omega_{ks}t+\mathrm{i}\omega_0 t}-1). \end{aligned} \tag{7.15}$$

The change in the occupancy of the state is mainly determined by the change in the magnitude squared of the wave function. Thus, the quantity of interest is the magnitude squared of (7.15), which becomes

$$|c_k|^2 = \frac{4|V_{ks}|^2}{\hbar^2(\omega_{ks}+\omega_0)^2}\sin^2[\tfrac{1}{2}(\omega_{ks}+\omega_0)t]. \tag{7.16}$$

This is now the probability that density moves from the initial state $s$ to the state $k$. Our interest is in the *rate* at which this probability transfers, or the *transition rate*

$$\Gamma_{ks} = \frac{|c_k|^2}{t}. \tag{7.17}$$

For large values of time, we note that (Landau and Lifshitz 1958)

$$\lim_{t\to\infty} \frac{\sin^2(at)}{a^2 t} = 2\pi\delta(a). \tag{7.18}$$

When $a \neq 0$, the limit is zero and the transition rate vanishes (we associate $a$ with $(\omega_{ks} + \omega_0)/2$). Thus, we note that the delta function serves to ensure that the transition rate is non-vanishing only for $a = 0$. Thus, the transition rate can be written as

$$\Gamma_{ks} = \frac{2\pi}{\hbar^2}|V_{ks}|^2\delta(\omega_{ks} + \omega_0) = \frac{2\pi}{\hbar}|V_{ks}|^2\delta(\hbar\omega_{ks} + \omega_0). \tag{7.19}$$

The last form in (7.19) is often referred to as the *Fermi golden rule*. The transition rate is given by a numerical factor times the squared magnitude of the matrix element and a delta function that conserves the energy in the transition. The latter is important, as it ensures that time-dependent variations arise only from those states in which the transition can conserve energy.

## 7.2 Electron–phonon scattering

Scattering of the electrons, or the holes, from one state to another by the lattice vibrations is one of the most important processes in the transport of the carriers through a semiconductor. In one sense, it is the scattering that limits the velocity of the charge carriers in the applied fields. Transport is seen as a balance between accelerative forces and dissipative forces (the scattering). In general, the electronic motion is separated from the lattice motion. It is the adiabatic principle, where the atomic motion is supposed to be slow relative to the electronic motion, which allows separation of the electronic motion from the lattice motion. There remains one term in the total system Hamiltonian that couples the electronic motion to the lattice motion. This term gives rise to the electron–phonon interaction. However, this is not a single interaction term. Rather, the electron–phonon interaction can be expanded in a power series in the scattered wave vector $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'$, and this process gives rise to a number of terms, which correspond to the number of phonon branches and the various types of interaction term. There can be acoustic phonon interactions with the electrons, and the optical interactions can be either through the polar interaction (in compound semiconductors) or through the non-polar interaction. These are just the terms up to the harmonic expansion of the lattice; higher-order terms give rise to higher-order interactions.

In this section, the basic *elastic* electron–phonon (which may also be hole–phonon) interaction is treated in a general sense and we develop the acoustic phonon scattering rate as an example of time-dependent perturbation theory.

The treatment followed here is based on the simple assumption that vibrations of the lattice cause small shifts in the energy bands. These vibrations were discussed in section 4.6, and the Fourier components are simple harmonic oscillators. Thus these oscillators induce periodic deviations in the bands from their equilibrium positions. Deviations of the bands due to these small shifts from the frozen lattice positions lead to an additional potential that causes the scattering process. The scattering potential is then used in time-dependent, first-order perturbation theory to find a rate at which electrons are scattered out of one state $\boldsymbol{k}$ and into another state $\boldsymbol{k}'$, while either absorbing or emitting a phonon of wave vector $\boldsymbol{q}$. Each of the different processes, or interactions, leads to a different 'matrix element' in terms of its dependence on these three wave vectors and their corresponding energy. These are discussed in the following sections, but here the treatment will retain just the existence of the scattering potential $\delta E$ which leads to a matrix element

$$V_{kk'} \to M(\boldsymbol{k}, \boldsymbol{k}') = \langle \Psi_{\boldsymbol{k}',\boldsymbol{q}} | \delta E | \Psi_{\boldsymbol{k},\boldsymbol{q}} \rangle \tag{7.20}$$

and the subscripts indicate that the wave function involves both the electronic and the lattice coordinates. Normally, the electronic wave functions are taken to be Bloch functions that exhibit the periodicity of the lattice. In addition, the matrix element usually contains the momentum conservation condition. Here this conservation condition leads to

$$\boldsymbol{k} - \boldsymbol{k}' \pm \boldsymbol{q} = \boldsymbol{G} \tag{7.21}$$

where $\boldsymbol{G}$ is a vector of the reciprocal lattice. In essence, the presence of $\boldsymbol{G}$ is a result of the Fourier transform from the real space lattice to the momentum space lattice, and the fact that we can only define the crystal momentum $\boldsymbol{k}$ within a single Brillouin zone. For the upper sign, the final state lies at a higher momentum than the initial state, and therefore also at a higher energy. This upper sign must correspond to the absorption of a phonon by the electron. The lower sign leads to the final state being at a lower energy and momentum, hence corresponds to the emission of a phonon by the electrons.

Straightforward time-dependent, first-order perturbation theory then leads to the equation for the scattering rate, in terms of the *Fermi golden rule* (7.19):

$$P(\boldsymbol{k}, \boldsymbol{k}') = \frac{2\pi}{\hbar} |M(\boldsymbol{k}, \boldsymbol{k}'')|^2 \delta(E_{\boldsymbol{k}} - E_{\boldsymbol{k}'} \pm \hbar\omega_{\boldsymbol{q}}) \tag{7.22}$$

and the signs have the same meaning as in the preceding paragraph: for example, the upper sign corresponds to the absorption of a phonon and the lower sign corresponds to the emission of a phonon.

The scattering rate out of the state defined by the wave vector $\boldsymbol{k}$ and the energy $E_{\boldsymbol{k}}$ is obtained by integrating (7.22) over all final states. Because of the momentum conservation condition (7.21), the integration can be carried out over either $\boldsymbol{k}'$ or $\boldsymbol{q}$ with the same result (omitting the processes for which the reciprocal lattice vector $\boldsymbol{G} \neq \boldsymbol{0}$). For the moment, the integration will be carried out over the final state wave vector $\boldsymbol{k}'$, and ($\Gamma = 1/\tau$ is the scattering rate, whose inverse is the scattering time $\tau$ used in previous paragraphs)

$$\Gamma(\boldsymbol{k}) = \frac{2\pi}{\hbar} \sum_{\boldsymbol{k}'} |M(\boldsymbol{k}, \boldsymbol{k}')|^2 \delta(E_{\boldsymbol{k}} - E_{\boldsymbol{k}'} \pm \hbar\omega_{\boldsymbol{q}}). \tag{7.23}$$

In those cases in which the matrix element $M$ is independent of the phonon wave vector, the matrix element can be removed from the summation, which leads to just the density of final states

$$\Gamma(\boldsymbol{k}) = \frac{2\pi}{\hbar} |M(\boldsymbol{k})|^2 \rho(E_{\boldsymbol{k}} \pm \hbar\omega_{\boldsymbol{q}}). \tag{7.24}$$

This has a very satisfying interpretation: the total scattering rate is just the product of the square of the matrix element and the total number of final states. For these cases the scattering angle is a random variable that is uniformly distributed across the energy surface of the final state. Thus any state lying on the final energy surface is equally likely, and the scattering is said to be isotropic.

One of the most common phonon-scattering processes is the interaction of the electrons (or holes) with the acoustic modes of the lattice through a deformation potential. Here, a long-wavelength acoustic wave moving through the lattice can cause a local strain in the crystal that perturbs the energy bands due to the lattice distortion. This change in the bands produces a weak scattering potential, which leads to a perturbing energy (Shockley and Bardeen 1950)

$$\delta E = \Xi_1 \Delta = \Xi_1 \nabla \cdot \boldsymbol{u}_{\boldsymbol{q}}. \tag{7.25}$$

Here, $\Xi_1$ is the *deformation potential* for a particular band and $\Delta$ is the *dilation* of the lattice produced by a wave, whose Fourier coefficient is $\boldsymbol{u}_{\boldsymbol{q}}$. We note here that any static displacement of the lattice is a displacement of the crystal as a whole and does not contribute, so that it is the wavelike variation of the amplitude within the crystal that produces the local strain in the bands. This variation is represented by the dilation, which is just the desired divergence of the wave. The amplitude $\boldsymbol{u}_{\boldsymbol{q}}$ is a relatively uniform Fourier coefficient for the overall lattice wave, and may be expressed as (4.89) with $\boldsymbol{u}_{\boldsymbol{q}}$ taken from (4.80) as

$$\boldsymbol{u}_{\boldsymbol{q}} = \left(\frac{\hbar}{2\rho_m V \omega_{\boldsymbol{q}}}\right)^{1/2} [a_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} + a_{\boldsymbol{q}}^{+} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}] \boldsymbol{e}_{\boldsymbol{q}} \mathrm{e}^{-\mathrm{i}\omega_{\boldsymbol{q}} t} \tag{7.26}$$

where $\rho_m$ is the mass density, $V$ is the volume, $a_{\boldsymbol{q}}$ and $a_{\boldsymbol{q}}^{+}$ are the annihilation and creation operators of section 4.4 for phonons, $\boldsymbol{e}_{\boldsymbol{q}}$ is the polarization vector, and the plane-wave factors have been incorporated along with the normalization factor for completeness. Because the divergence produces a factor proportional to the component of $\boldsymbol{q}$ in the polarization direction (along the direction of propagation), only the longitudinal acoustic modes couple to the carriers in a spherically symmetric band. The fact that the resulting interaction potential is now proportional to $q$ (i.e., to first order in the phonon wave vector) leads to this term being called a *first-order* interaction.

The matrix element may now be calculated by considering the proper sum over both the lattice and the electronic wave functions. The second term in (7.26), the term for the emission of a phonon by the carrier, leads to the matrix element squared, as

$$|M(\boldsymbol{k},\boldsymbol{q})|^2 = \frac{\hbar\Xi_1^2 q^2}{2\rho_m V\omega_q}(N_q+1)I_{\boldsymbol{k},\boldsymbol{q}}^2 \tag{7.27}$$

where $N_{\boldsymbol{q}}$ is the Bose–Einstein distribution function for the phonons, and

$$I_{\boldsymbol{k},\boldsymbol{q}} = \int_{\Omega} u_{\boldsymbol{k}-\boldsymbol{q}}^{+} u_{\boldsymbol{k}}\,\mathrm{d}^3\boldsymbol{r} \tag{7.28}$$

is the *overlap* integral between the cell portions of the Bloch waves (unfortunately, similar symbols are used, but the $u_{\boldsymbol{k}}$ in this equation is the cell periodic part of the Bloch wave and not the phonon amplitude given earlier) for the initial and final states, and the integral is carried out over the cell volume $\Omega$. For elastic processes, and for both states lying within the same 'valley' of the band, this integral is unity. Essentially, exactly the same result (7.27) is obtained for the case of the absorption of phonons by the electrons, with the single exception that $(N_{\boldsymbol{q}}+1)$ is replaced by $N_{\boldsymbol{q}}$.

One thing that should be recognized is that the acoustic modes have very low energy. If the velocity of sound is $5 \times 10^5$ cm s$^{-1}$, a wave vector corresponding to 25% of the zone edge yields an energy only of the order of 10 meV. This is a very large wave vector, so for most practical cases the acoustic mode energy will be less than a millivolt. Scattering processes in which the phonon energy may be ignored are termed *elastic* scattering events. Of more interest here is the fact that these energies are much lower than the thermal energy except at the lowest temperatures, and the Bose–Einstein distribution can be expanded under the equipartition approximation as

$$N_q = \frac{1}{\exp\left(\frac{\hbar\omega_q}{k_\mathrm{B}T}\right)-1} \sim \frac{k_\mathrm{B}T}{\hbar\omega_q} \gg 1. \tag{7.29}$$

Since this distribution is so large and the energy exchange so small, it is quite easy to add the two terms for emission and absorption together, and use the fact that $\omega_q = q\boldsymbol{v}_\mathrm{s}$, where $\boldsymbol{v}_\mathrm{s}$ is the velocity of sound, to achieve

$$|M(\boldsymbol{k})|^2 \approx \frac{\Xi_1^2 k_\mathrm{B}T}{\rho_m V\boldsymbol{v}_\mathrm{s}^2}. \tag{7.30}$$

For electrons in a simple, spherical energy surface and parabolic bands, this leads to

$$\begin{aligned}\Gamma(k) &= \frac{2\pi}{\hbar}\frac{\Xi_1^2 k_{\mathrm{B}}T}{\rho_m V \boldsymbol{v}_{\mathrm{s}}^2}\frac{v}{4\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} E^{1/2} \\ &= \frac{\Xi_1^2 k_{\mathrm{B}}T(2m^*)^{3/2}}{2\pi\hbar^4\rho_m\boldsymbol{v}_{\mathrm{s}}^2}E^{1/2}. \qquad (7.31)\end{aligned}$$

It has been assumed that the interaction does not mix spin states, and this factor is accounted for in the density of states. Although most of the parameters may easily be obtained for a particular semiconductor, it is found that the deformation potential itself is almost universally of the order of 7 to 10 eV for nearly all semiconductors. Nevertheless, this result for the scattering rate of electrons by acoustic phonons is a straightforward example of the application of the Fermi golden rule.

## 7.3 The interaction representation

We now want to revisit the idea of time-dependent perturbation theory, but in an approach that is based upon the Heisenberg representation introduced in chapter 5. To review, for this approach, the wave function is expanded in a linear vector space of basis functions, just as in the approach leading to (7.3), but now it is assumed that the basis set is not time varying, at least in the unperturbed state. Rather, the time variation is attached to the operators themselves, which rotate in the fixed linear vector space (which forms our coordinate system) as a function of time according to

$$A(t) = \mathrm{e}^{\mathrm{i}H_0 t/\hbar} A \mathrm{e}^{-\mathrm{i}H_0 t/\hbar} \qquad (7.32)$$

with

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar}[H_0, A]. \qquad (7.33)$$

This is the basis of the Heisenberg representation.

In the presence of a time-varying perturbation, however, it will now be assumed that the perturbation introduces a slow variation in the basis set itself. Just as we earlier assumed that the coefficients of the expansion were now time varying, we will assume that this slow variation is entirely due to the perturbing potential $V(t)$ defined according to (7.32) as

$$V(t) = \mathrm{e}^{\mathrm{i}H_0 t/\hbar} V \mathrm{e}^{-\mathrm{i}H_0 t/\hbar} \qquad (7.34)$$

except that we will also assume below that there is an explicit time variation to this perturbation (such as might arise if the perturbation were due to an interaction with a harmonic oscillator). Hence, there is a slow variation of the wave function in the linear vector space that is introduced by $V$, while all other operators still respond to the unperturbed Hamiltonian according to (7.33). This mixed representation is termed the *interaction representation*.

In keeping with the development leading to (5.124), it may be assumed that there is a unitary operator $U(t, t_0)$ that describes the slow time evolution of the wave function according to

$$\Psi(t) = U(t, t_0)\Psi(t_0) \tag{7.35}$$

from which we may write the equivalent Schrödinger equation as

$$\mathrm{i}\hbar\frac{\mathrm{d}U}{\mathrm{d}t} = V(t)U(t, t_0). \tag{7.36}$$

The approach now is to define the properties of the unitary operator $U(t, t_0)$. First, we note that for $t = t_0$,

$$U(t_0, t_0) = 1 \tag{7.37}$$

and, in general,

$$U(t_1, t_0) = U(t_1, t_2)U(t_2, t_0). \tag{7.38}$$

Using (7.37), we can directly integrate (7.36) to give

$$U(t, t_0) = 1 - \frac{\mathrm{i}}{\hbar}\int_{t_0}^{t} V(t')U(t', t_0)\,\mathrm{d}t'. \tag{7.39}$$

Obviously, this integral equation is not easily solved (because of the two-time behaviour, it is not a simple convolution integral). Thus, the normal method is to develop a perturbation series of various orders of iteration. For example, in the first iteration, it is assumed that only the first term of (7.39) is used inside the integral; then this result is used within the integral to produce the next iterate, and so on. This procedure gives the sequence of approximations to $U(t, t_0)$ as

$$U_0(t, t_0) = 1 \tag{7.40$a$}$$

$$U_1(t, t_0) = 1 - \frac{\mathrm{i}}{\hbar}\int_{t_0}^{t} V(t')U_0(t', t_0)\,\mathrm{d}t' = 1 - \frac{\mathrm{i}}{\hbar}\int_{t_0}^{t} V(t')\,\mathrm{d}t' \tag{7.40$b$}$$

$$U_2(t, t_0) = 1 - \frac{\mathrm{i}}{\hbar}\int_{t_0}^{t} V(t')\,\mathrm{d}t' + \left(\frac{\mathrm{i}}{\hbar}\right)^2\int_{t_0}^{t} V(t')\int_{t_0}^{t'} V(t'')\,\mathrm{d}t''\,\mathrm{d}t' \tag{7.40$c$}$$

and so on. We leave it as a problem at the end of the chapter to show that the infinite iterate is the exponential operator

$$U_\infty(t, t_0) = \exp\left[-\frac{\mathrm{i}}{\hbar}\int_{t_0}^{t} V(t')\,\mathrm{d}t'\right]. \tag{7.40$d$}$$

In keeping with the approach begun in section 7.1, we will assume that the linear vector space is rotated so that the wave function initially is completely in state

$s$. Then, we can take the inner product with state $k$ to obtain the equivalent coefficient:

$$\begin{aligned}(\psi_k, \Psi(t)) &= (\psi_k, U(t, t_0)\psi_s)\\ &= \langle k|s\rangle - \frac{\mathrm{i}}{\hbar}\int_{t_0}^{t} V_{ks}(t')\,\mathrm{d}t'\\ &\quad + \left(\frac{\mathrm{i}}{\hbar}\right)^2 \int_{t_0}^{t}\int_{t_0}^{t'} \langle k|V(t')V(t'')|s\rangle\,\mathrm{d}t''\,\mathrm{d}t' + \cdots. \end{aligned}\tag{7.41}$$

Obviously, the first term on the right-hand side vanishes by the principle of orthogonality, while the second term is essentially the first line of (7.15) if we recognize that the exponential is buried in the time dependence of the perturbing potential. We note that the left-hand side is the coefficient $c_k$ defined in section 7.1. The result here tells us that the result obtained earlier is just the first term in a perturbation series expansion, which may be terminated only so long as the terms decrease sufficiently rapidly in order.

Here, we would like to take into account the specific time variation of the perturbing potential according to the simple rule

$$V = V_0 \mathrm{e}^{\pm \mathrm{i}\omega t}. \tag{7.42}$$

The matrix element is then

$$\begin{aligned}\langle k|V(t)|s\rangle &= \langle k|\mathrm{e}^{\mathrm{i}H_0 t/\hbar} V_0 \mathrm{e}^{\pm\mathrm{i}\omega t}\mathrm{e}^{-\mathrm{i}H_0 t/\hbar}|s\rangle\\ &= V_{0,ks}\mathrm{e}^{\mathrm{i}(\omega_{ks}\pm\omega)t}.\end{aligned}\tag{7.43}$$

This leads to the coefficient ($t_0 = 0$)

$$\begin{aligned}c_k(t) = (\psi_k, \Psi(t)) &\simeq -\frac{\mathrm{i}}{\hbar}\int_0^t V_{0ks}\mathrm{e}^{\mathrm{i}(\omega_{ks}\pm\omega)t'}\,\mathrm{d}t'\\ &= -\frac{V_{0,\,ks}}{\hbar}\frac{\mathrm{e}^{\mathrm{i}(\omega_{ks}\pm\omega)t}-1}{\omega_{ks}\pm\omega}\end{aligned}\tag{7.44}$$

and

$$|c_k(t)|^2 = \frac{4}{\hbar^2}|V_{0,ks}|^2\frac{\sin^2[(\omega_{ks}\pm\omega)t]}{(\omega_{ks}\pm\omega)^2}. \tag{7.45}$$

Using the result (7.18), we obtain the transition rate as

$$\Gamma_{ks} = \frac{2\pi}{\hbar}|V_{0,ks}|^2\delta[\hbar(\omega_{ks}\pm\omega)] \tag{7.46}$$

which is just (7.19). The delta function conserves energy in the transition rate, so we have to interpret (7.46) as declaring that the states $k$ and $s$, which are coupled by the perturbation, must have energies that differ from each other by the energy $\hbar\omega$ of the photon associated with the oscillation in the potential. Thus, a single photon of energy $\hbar\omega$ is absorbed (or emitted) in moving the state from coordinate $s$ (and energy $\hbar\omega_s$) to the state with coordinate $k$ (and energy $\hbar\omega_k$). The initial and final states are no longer degenerate, but differ in energy by the energy of the perturbing potential. The second-order term in (7.41) is a two-photon transition, and each higher-order term includes another photon in the process. (We use the term photon here, but the type of particle depends upon the perturbation and could be a photon, a phonon, or whatever.)

## 7.4 Exponential decay and uncertainty

In the preceding sections, it has been assumed that the initial state remains at unit amplitude. In fact, this is really unlikely, and it is now desirable to see how the decay of the initial state affects the transition probability. Each approach yields the same form for the transition rate, so we will use the direct form of (7.14)–(7.15); we rewrite this as

$$c_k = -\frac{\mathrm{i}}{\hbar}\int_0^t V_{ks} c_s \mathrm{e}^{\mathrm{i}\omega_{ks}t'}\,\mathrm{d}t' \tag{7.47}$$

where it is assumed that the perturbing potential is time invariant. The inclusion of any time variation of this potential is easily re-inserted using the results of (7.43). In a similar manner, (7.12) must also be written for the initial state as

$$\mathrm{i}\hbar\frac{\mathrm{d}c_s}{\mathrm{d}t} = \sum_k V_{sk} c_k(t) \mathrm{e}^{\mathrm{i}\omega_{sk}t}. \tag{7.48}$$

We need first to solve for the time variation of the initial state $c_s$. Combining the last two equations, we find that

$$\frac{\mathrm{d}c_s}{\mathrm{d}t} = -\frac{1}{\hbar^2}\sum_k |V_{sk}|^2 \int_0^t c_k(t')\mathrm{e}^{\mathrm{i}\omega_{ks}(t'-t)}\,\mathrm{d}t'. \tag{7.49}$$

The last term is a convolution integral of the coefficient $c_s$ and the exponential time variation. This suggests that a transform approach is the best way to solve the equation for the time variation of this coefficient. For this purpose, the Laplace transform will be used, but the transform variable will be taken to be $z$, as the traditional $s$ has been used for the state notation. Thus, using (7.13), the transform of (7.49) leads to the equation

$$C_s(z)\left[z + \frac{\mathrm{i}}{\hbar}V_{ss} + \frac{1}{\hbar^2}\sum_{k\neq s}|V_{sk}|^2\frac{1}{z - \mathrm{i}\omega_{ks}}\right] = 1 \tag{7.50}$$

where the term for $k = s$ has been separated out from the rest of the sum. To clarify this result, the last term in the square brackets will be rationalized, and

$$C_s(z)\left[z + \frac{1}{\hbar^2}\sum_{k\neq s}|V_{sk}|^2\frac{z}{z^2+\omega_{ks}^2} + \frac{\mathrm{i}}{\hbar}\left\{V_{ss} + \frac{1}{\hbar}\sum_{k\neq s}\frac{|V_{sk}|^2\omega_{ks}}{z^2+\omega_{ks}^2}\right\}\right] = 1. \tag{7.51}$$

In the end, we will be taking the long-time limit, which corresponds to $z \to 0$ (the multiplication by $z$ is in the finished radical which will not affect the approximations we are about to make). In this limit, the second term in the large bracket becomes just (7.19) summed over the states $k$, as the fraction becomes the delta function; that is, the second term becomes just half of

$$\Gamma_s = \frac{2\pi}{\hbar} \sum_{k \neq s} |V_{sk}|^2 \delta(\hbar\omega_{ks}) \tag{7.52}$$

which is the total *out-scattering* rate from state $s$. The term in the curly brackets is just the energy shift in (6.25) found to second order in time-independent perturbation theory. Thus, we may re-transform (7.51) as

$$c_s(t) = \exp\left[-\left(\frac{\Gamma_s}{2} + \frac{\mathrm{i}}{\hbar}\Delta\mathcal{E}_s\right) t\right]. \tag{7.53}$$

The out-scattering process causes a decay of the state amplitude, as expected, but there is also a shift in the frequency of the state, which is often referred to as the *self-energy correction*. We note that the probability amplitude of the initial state decays as

$$P_s(t) = |c_s(t)|^2 = \mathrm{e}^{-\Gamma_s t}. \tag{7.54}$$

The result (7.53) can now be used in (7.47) to compute the rate of scattering into the state $k$. This gives

$$c_k = -\frac{\mathrm{i}}{\hbar} \int_0^t V_{ks} \mathrm{e}^{-\Gamma_s t'/2 + \mathrm{i}\Omega_{ks} t'} \, \mathrm{d}t' \tag{7.55}$$

where

$$\Omega_{ks} = \omega_{ks} - \frac{\mathrm{i}}{\hbar}\Delta\mathcal{E}_s = \frac{\mathcal{E}_k - \mathcal{E}_s - \Delta\mathcal{E}_s}{\hbar} \tag{7.56}$$

is the shift in frequency due to the self-energy correction of the initial state. Equation (7.55) is readily integrated as

$$c_k = -\frac{V_{ks}}{\hbar} \frac{\mathrm{e}^{-\Gamma_s t/2 + \mathrm{i}\Omega_{ks} t} - 1}{\Omega_{ks} + \mathrm{i}\Gamma_s/2} \tag{7.57}$$

and

$$|c_k|^2 = \frac{|V_{ks}|^2}{\hbar^2} \frac{1 - 2\mathrm{e}^{-\Gamma_s t/2}\cos(\Omega_{ks} t) + \mathrm{e}^{-\Gamma_s t}}{\Omega_{ks}^2 + (\Gamma_s/2)^2}. \tag{7.58}$$

After a very long time compared with the *lifetime* $\hbar/\Gamma_s$ (we have previously taken the long-time limit), we obtain the occupancy of the $k$th state as

$$|c_k|^2 = \frac{|V_{ks}|^2}{\hbar^2[\Omega_{ks}^2 + (\Gamma_s/2)^2]} \tag{7.59}$$

which exhibits the bell-shaped resonance behaviour with a peak height of $4/\Gamma_s^2$ at the new resonance position $\Omega_{ks} = 0$ (which is shifted from the normal position of the delta function by the self-energy of the state $s$).

One problem with the form of (7.59) is that the peaked shape of the function, which is a Lorentzian line shape, does not integrate out to be equivalent to (7.19). In the case of (7.19), the quantity $|c_k|^2$ was divided by $t$ in order to determine the transition probability because it increased linearly with time. If (7.59) is 'integrated' over all energies $\hbar\Omega_{ks}$, it is found that the result is

$$\frac{2\pi}{\hbar\Gamma_s}|V_{ks}|^2 \tag{7.60}$$

which differs from the result obtained from (7.19) by the quantity $\Gamma_s$. It is clear from this that instead of dividing by the time, in this decaying case the effective time is the reciprocal of the lifetime of the state $s$, which is $\Gamma_s$. Thus, the probability of transition into state $k$ is given by

$$\Gamma_{ks} = \Gamma_s|c_k|^2 = \frac{|V_{ks}|^2}{\hbar^2}\frac{\Gamma_s}{\Omega_{ks}^2 + (\Gamma_s/2)^2}. \tag{7.61}$$

As the lifetime reduces toward zero, the Lorentzian lineshape becomes an approximation to a delta function, which reproduces (7.19).

We now see that the decay of the initial state $s$ gives rise to an effective 'uncertainty' relationship between this lifetime and the energy. The half-amplitude points of the Lorentzian curve are the values of frequency such that $\Omega_{ks} = \Gamma_s/2$. Then it is possible to define an energy width $\Delta\mathcal{E} = \hbar\Gamma_s$ (corresponding to twice the frequency width defined in the previous sentence) or, for $\tau = 1/\Gamma_s$,

$$\tau\Delta\mathcal{E} \geq \hbar \tag{7.62}$$

which is a factor of two larger than a true uncertainty principle would produce. One should not make the mistake of calling this an *uncertainty principle*, however (Landau and Lifshitz 1958). In arriving at (7.62), the limit $t \to \infty$ has been invoked. There is no uncertainty, or standard deviation, of the time parameter; it has been sent off to infinitely large values. Instead, the lifetime of the initial state $s$ has induced a measurement problem in the determination of the energy corresponding to this state (due to the self-energy shift of the initial state). Our measurement time is limited by this lifetime; thus the accuracy with which the energy can be measured is limited. However, this is not a fundamental uncertainty principle, since these are not non-commuting operators. Rather, this result is more a property of Fourier transform pairs, or to be more strictly accurate, the problem of taking the Fourier transform in a finite time window. Here, the time window is set by the lifetime. Using a window function in the Fourier transform causes the actual transform to be convolved by the transform of the window function, which limits the resolution in the spectral domain. Here, the lifetime limits the resolution of the energy scale. Again, this is not a fundamental uncertainty principle, but a limitation placed on the experiment by the finite measurement time, as defined by the lifetime.

## 7.5 A scattering-state basis—the $T$-matrix

In treating the Fermi golden rule transition rate of the last few sections, it was assumed that the perturbation was turned on at $t = 0$. This was done simply to avoid facing the evaluation of the limit of the integrals at $t \to -\infty$. The latter limit is problematic when the excitation is taken to be a simple sinusoid or a constant function. This has also allowed us to use the Laplace transforms where necessary. A somewhat different approach is often used, in which Fourier transforms are utilized as necessary. In order to avoid the problem of the limit at large negative time, a different strategy is adopted. Here, let us assume that the perturbation is slowly turned on over a very long time period, but in such a manner that the matrix element can be written as

$$V_{ks} = T_{ks}\mathrm{e}^{\alpha t} \tag{7.63}$$

with $\alpha > 0$ for $t < 0$ and $\alpha = 0$ for $t > 0$. In general, the factor $\alpha$ can be set to zero after the lower limit of the integrals is evaluated. With this approach, (7.15) becomes (we include a delta function for the case where $k = s$)

$$c_k = -\frac{\mathrm{i}T_{ks}}{\hbar}\int_{-\infty}^{t} \mathrm{e}^{\mathrm{i}\omega_{ks}t'+\alpha t'}\,\mathrm{d}t' = \frac{T_{ks}}{\hbar}\frac{\mathrm{e}^{\mathrm{i}\omega_{ks}t'+\alpha t'}}{\mathrm{i}\alpha - \omega_{ks}} + \delta_{ks}. \tag{7.64}$$

The last form is only valid for $t \ll 1/\alpha$, so the growing exponential is not to be considered as a major factor. From this, we find that

$$|c_k|^2 = \frac{|T_{ks}|^2}{\hbar^2}\frac{\mathrm{e}^{2\alpha t}}{\alpha^2 + \omega_{ks}^2} \tag{7.65}$$

and

$$\begin{aligned}\Gamma_{ks} &= \lim_{\alpha\to 0}\frac{\mathrm{d}|c_k|^2}{\mathrm{d}t} = \lim_{\alpha\to 0}\frac{|T_{ks}|^2}{\hbar^2}\frac{2\alpha\mathrm{e}^{2\alpha t}}{\alpha^2+\omega_{ks}^2}\\ &\to \frac{2\pi|T_{ks}|^2}{\hbar^2}\delta(\omega_{ks})\end{aligned} \tag{7.66}$$

where we have assumed that $\alpha$ is small and taken only the derivative of the exponential term. This result is just (7.19), the Fermi golden rule.

### 7.5.1 The Lippmann–Schwinger equation

The matrix of quantities $T_{ks}$ is usually called the $T$-matrix, and it was introduced in the above discussion in a rather *ad hoc* manner. Here, we want to pursue it a little more deeply and find out how it is related to the matrix elements $V_{ks}$. To proceed, let us insert the basic equation (7.64) into the earlier (7.12) for the coefficients of the expansion series. This leads to

$$T_{ks}\mathrm{e}^{\mathrm{i}\omega_{ks}t} = \frac{1}{\hbar}\sum_n \frac{V_{kn}T_{ns}\mathrm{e}^{\mathrm{i}(\omega_{kn}+\omega_{ns})t+\alpha t}}{\mathrm{i}\alpha - \omega_{ns}} + V_{ks}\mathrm{e}^{\mathrm{i}\omega_{ks}t}. \tag{7.67}$$

We note that the frequency term in the exponent reduces to $\omega_{ks}$, which relates the initial and final states described by the $T$-matrix term on the left-hand side of the equation. For all practical purposes, we can now set $\alpha = 0$ in the exponent, and the exponential factors drop out of the equation. The summation includes terms for which $n = s$, which lead to singularities, that carry significant weight if the energy spectrum is a set of discrete energy levels. On the other hand, if the energy spectrum is continuous, any single point in the spectrum does not carry much weight, and the singularity is not serious. For this reason, we will assume that the following is for a system in which the energy spectrum is continuous, such as a free-electron gas.

Let us now define a new total wave function $\psi_n^{(+)}$ which describes an outgoing, causal wave that arises from the scattering process. We will define this new outgoing wave in terms of the $T$-matrix through the matrix elements

$$T_{ks} = (\psi_k, V\psi_s^{(+)}) = \sum_n (\psi_k, V\psi_n)(\psi_n, \psi_s^{(+)}) \tag{7.68}$$

where an expansion of the delta function has been inserted in the last line. Using this definition, (7.67) can be rewritten as

$$(\psi_k, V\psi_s^{(+)}) = \sum_n \frac{(\psi_k, V\psi_n)(\psi_n, V\psi_s^{(+)})}{\hbar(\mathrm{i}\alpha - \omega_{ns})} + (\psi_k, V\psi_s). \tag{7.69}$$

If we now let $k = s$, this becomes

$$(\psi_s, V\psi_s^{(+)}) = \sum_n \frac{(\psi_s, V\psi_n)(\psi_n, V\psi_s^{(+)})}{\mathrm{i}\hbar\alpha - \mathcal{E}_n + \mathcal{E}_s} + (\psi_s, V\psi_s). \tag{7.70}$$

This can be recognized as just taking the product of the wave function with the perturbing potential and the unperturbed wave function and integrating, for example,

$$\begin{aligned}\psi_s^{(+)} &= \sum_n \psi_n \frac{(\psi_n, V\psi_s^{(+)})}{\mathrm{i}\hbar\alpha - \mathcal{E}_n + \mathcal{E}_s} + \psi_s \\ &= \psi_s + \sum_n \frac{1}{\mathrm{i}\hbar\alpha - H_0 + \mathcal{E}_s}\psi_n(\psi_n, V\psi_s^{(+)}).\end{aligned} \tag{7.71}$$

In the last line, we have recognized that (7.3) can be used to replace the energy with the Hamiltonian operator (the new ordering is very important). The summation over $\psi_n$ is just the resolution of the delta function (we recognize the presence of the projection operator), and since the energy denominator no longer has a dependence on the summation index, we can write (7.71) as

$$\psi_s^{(+)} = \psi_s + \frac{1}{\mathrm{i}\hbar\alpha - H_0 + \mathcal{E}_s} V \psi_s^{(+)}. \tag{7.72}$$

This is the Lippmann–Schwinger equation, and relates the outgoing scattered wave to the incoming wave at the same enegy level. As in section 7.3, the inverse energy operator can be expanded into a power series that will yield the perturbation series. However, it is possible to use (7.72) to find the exact solution for the scattered wave function in many cases. It is easy to show that the result is the exact solution. Let us operate on (7.72) with the operator $H_0 - \mathcal{E}_s$ in the limit in which $\alpha = 0$. This leads to

$$(H_0 - \mathcal{E}_s)\psi_s^{(+)} = (H_0 - \mathcal{E}_s)\psi_s - V\psi_s^{(+)} \tag{7.73}$$

or, since the first term on the right-hand side is zero by (7.3),

$$(H_0 + V)\psi_s^{(+)} = \mathcal{E}_s \psi_s^{(+)}. \tag{7.74}$$

Thus, the solutions for the outgoing waves are exact solutions to the total Hamiltonian.

### 7.5.2 Coulomb scattering again

In section 7.2, the scattering of an incoming plane wave by a Coulomb potential was first discussed. The configuration of this scattering problem is pictured in figure 7.1. There, we also introduced the cross section $\sigma$ of the scattering potential in terms of the impact parameter $\rho$. In this section, we would like to calculate this scattering cross section from the $T$-matrix approach. To begin, we note that the Fermi golden rule can be used to describe the total scattering, where we assume that the initial state $s$ is the incident plane wave state. Then, using (7.66), we find that the total transition probability, and hence the total scattering-out rate from state $s$, is given by

$$\Gamma = \sum_k \frac{\mathrm{d}|c_k|^2}{\mathrm{d}t} \to \frac{2\pi}{\hbar} \int |T_{ks}|^2 \delta(\mathcal{E}_k - \mathcal{E}_s)\,\mathrm{d}\mathcal{E}_k. \tag{7.75}$$

We have assumed that there is a near continuum of energy levels, so it is likely that there is significant degeneracy at any given energy level, and the integral will count the number of final states for which $\mathcal{E}_k = \mathcal{E}_s$, with each weighted by the factor $|T_{ks}|^2$. Our first task is then to evaluate the $T$-matrix elements and ascertain just how they weight these states over which the sum is performed.

The individual matrix elements of the $T$-matrix were defined for us in (7.68), which we can rewrite in terms of an incident plane wave, as

$$T_{ks} = \left(\psi_k, V\psi_s^{(+)}\right) = \frac{1}{\sqrt{L^3}} \int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} V(\boldsymbol{r})\psi_s^{(+)}\,\mathrm{d}^3\boldsymbol{r} \tag{7.76}$$

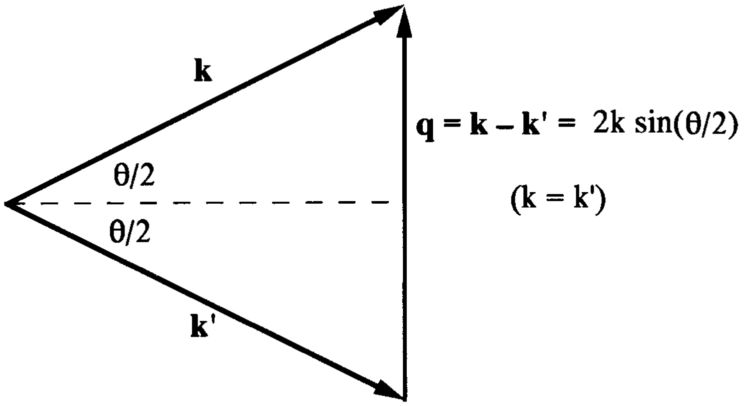


**Figure 7.1.** The scattering geometry for the Coulomb scattering of a plane wave.

where $L^3$ is the volume normalization for the plane wave that is incident. In addition, it has been assumed that the incident plane wave is described by the wave vector $\boldsymbol{k}$ of energy $\mathcal{E}_k$. How are we to describe the scattered wave function $\psi_s^{(+)}$? It may be assumed that the scattering is small, so we can approximate this wave function as a plane wave as well, but will take the scattering into a new wave vector $\boldsymbol{k}'$, describing the energy $\mathcal{E}_s$. The delta function in (7.75) ensures that these two energies are equal, which means that the two wave functions must have equal amplitudes (but not necessarily the same directions). Thus, we assume that

$$\psi_s^{(+)} \simeq \frac{1}{\sqrt{L^3}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}} \tag{7.77}$$

and so

$$T_{ks} = \frac{1}{L^3} \int \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{r}} V(\boldsymbol{r})\, \mathrm{d}^3\boldsymbol{r} \tag{7.78}$$

which is just the Fourier transform of the Coulomb potential. This can be evaluated by standard techniques (assuming a convergence factor to eliminate the problem of $r \to -\infty$, where the scattering is assumed to vanish), which leads to

$$T_{ks} = \frac{e^2}{\varepsilon L^3 |\boldsymbol{k} - \boldsymbol{k}'|^2} \tag{7.79}$$

where $\varepsilon$ is the dielectric permittivity of the medium. Hence, there is a strong dependence upon the magnitude of the scattered wave vector in the matrix element and we have to take care in evaluating the integral in (7.75).

While we have indicated an integration over the energy coordinate, the result for the matrix element of the $T$-matrix indicates that the integration should be more carefully carried out in momentum space. For this purpose, the delta function and differential can be transformed to

$$\delta(\mathcal{E}_k - \mathcal{E}_s) = \frac{m}{\hbar^2 k}\delta(\boldsymbol{k} - \boldsymbol{k}') \tag{7.80}$$

and

$$\mathrm{d}\mathcal{E}_k = \left(\frac{L}{2\pi}\right)^3 k^2 \,\mathrm{d}k \sin\theta \,\mathrm{d}\theta \,\mathrm{d}\phi. \tag{7.81}$$

The direction of the scattering vector $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'$ is found subject to the delta function which gives the scattering geometry shown in figure 7.1. This leads to

$$\begin{aligned}\Gamma &= \frac{2\pi}{\hbar}\left(\frac{L}{2\pi}\right)^3 \int \left(\frac{e^2}{\varepsilon L^3 |\boldsymbol{k} - \boldsymbol{k}'|^2}\right)^2 \frac{m}{\hbar^2 k}\delta(k - k')\, k^2 \,\mathrm{d}k \sin\theta \,\mathrm{d}\theta \,\mathrm{d}\phi \\ &= \frac{2\pi m}{\hbar^3 k}\left(\frac{L}{2\pi}\right)^3 \left(\frac{e^2}{2\varepsilon L^3}\right)^2 2p \int \frac{\sin\theta \,\mathrm{d}\theta}{\sin^2(\theta/2)}.\end{aligned} \tag{7.82}$$

The last integral has a problem with the lower limit of the $\theta$-integration, where it diverges. This divergence is usually removed via the following argument: the transition rate of interest is really that at which the momentum of the incoming wave is reduced, which leads to the introduction of an additional factor of $(1 - \cos\theta)$ into the argument of the integral. This factor arises from projecting $\boldsymbol{k}'$ onto $\boldsymbol{k}$, so the change in the initial momentum is proportional to $|\delta k| = k - k'\cos\theta = k(1 - \cos\theta)$, which gives the factor of $(1 - \cos\theta)$ to be included within the integral. Thus, each scattering process reduces the momentum by only a small amount, and it is the integrated sum of many scattering processes that relaxes the momentum. It is really this momentum decay rate that is of interest. The result would then properly be called a momentum transition rate, or momentum scattering rate. We will pursue this approach to complete the illustration of the scattering cross section.

Using the convergence factor to describe the momentum scattering rate, the transition rate becomes

$$\begin{aligned}\Gamma &= \frac{4\pi^2 m}{\hbar^3 k}\left(\frac{L}{2\pi}\right)^3 \left(\frac{e^2}{2\varepsilon L^3}\right)^2 \int \frac{(1 - \cos\theta)\sin\theta \,\mathrm{d}\theta}{\sin^2(\theta/2)} \\ &= \frac{32\pi^2 m}{3\hbar^3 k}\left(\frac{L}{2\pi}\right)^3 \left(\frac{e^2}{2\varepsilon L^3}\right)^2 = \frac{e^4 m}{3\varepsilon^2 \hbar^3 k L^3}.\end{aligned} \tag{7.83}$$

This result is the scattering rate for a single Coulomb centre, and needs of course to be modified if there are multiple scattering centres. Normally, one uses the density of these centres, which eliminates the volume term remaining in (7.83). We can now retreat from this to obtain the cross section through the definition

$$\sigma = \frac{\Gamma}{v/L^3} = \frac{e^4 m^2}{3\varepsilon^2 \hbar^4 k^2}. \tag{7.84}$$

Generally, one goes in the opposite direction, and computes the scattering rate from the cross section through $\Gamma = Nv\sigma$, where $N$ is the density of scattering centres.

### 7.5.3 Orthogonality of the scattering states

The Lippmann–Schwinger equation can now be used to show that the scattering states remain normalized and orthogonal. Before doing this, though, we diverge to illustrate another representation of the equation. If the resolvent operator, or Green's function (in this case a retarded Green's function due to the sign on the imaginary part), is defined by

$$G(\mathcal{E}_s) = \frac{1}{\mathcal{E}_s - H_0 + \mathrm{i}\hbar\alpha} \tag{7.85}$$

then (7.72) can be rewritten as

$$\psi_s^{(+)} = \frac{1}{1 - G(\mathcal{E}_s)V\psi_s}. \tag{7.86}$$

This latter form is termed Dyson's equation, and clearly illustrates how the perturbation series is obtained by expanding the denominator in a power series. In fact, the form (7.86) is often termed the re-summed perturbation series.

Let us now rewrite (7.72) to obtain another useful result. We pre-multiply all the terms by the denominator of the last term to give

$$(\mathcal{E}_s - H_0 + \mathrm{i}\hbar\alpha)\psi_s^{(+)} = (\mathcal{E}_s - H_0 + \mathrm{i}\hbar\alpha)\psi_s = V\psi_s^{(+)} \tag{7.87}$$

which can be rearranged to give

$$(\mathcal{E}_s - H_0 - V + \mathrm{i}\hbar\alpha)\psi_s^{(+)} = (\mathcal{E}_s - H_0 + \mathrm{i}\hbar\alpha)\psi_s \tag{7.88}$$

or

$$\begin{aligned}(\mathcal{E}_s - H + \mathrm{i}\hbar\alpha)\psi_s^{(+)} &= (\mathcal{E}_s - H_0 + \mathrm{i}\hbar\alpha)\psi_s \\ &= (\mathcal{E}_s - H + \mathrm{i}\hbar\alpha)\psi_s + V\psi_s.\end{aligned} \tag{7.89}$$

This can now be rearranged to give

$$\psi_s^{(+)} = \psi_s + \frac{1}{\mathcal{E}_s - H + \mathrm{i}\hbar\alpha}V\psi_s. \tag{7.90}$$

With a little algebra, it is clear that this form is merely the Dyson equation, but rearranged slightly.

Finally, we want to show that the scattering states retain their orthogonality and orthonormality. To begin, we will use (7.90) to describe just one of the states according to

$$\begin{aligned}(\psi_r^{(+)}, \psi_s^{(+)}) &= \left(\psi_r + \frac{1}{\mathcal{E}_r - H + \mathrm{i}\hbar\alpha}V\psi_r, \psi_s^{(+)}\right) \\ &= \left(\psi_r, \psi_s^{(+)} + V\frac{1}{\mathcal{E}_r - H - \mathrm{i}\hbar\alpha}\psi_s^{(+)}\right) \\ &= \left(\psi_r, \psi_s^{(+)} + V\frac{1}{\mathcal{E}_r - \mathcal{E}_s - \mathrm{i}\hbar\alpha}\psi_s^{(+)}\right)\end{aligned} \tag{7.91}$$

where we have used (7.74) to replace the total Hamiltonian with the energy of the scattered state. We can now factor out the minus sign of the denominator, and since the fraction is now a $c$-number, we can reverse the order of the terms multiplying the last scattering state. We then take the total operator *back* to the adjoint term, where the $\mathcal{E}_r$-term becomes the operator $H_0$. Then, returning this operator to the position shown in (7.91), this latter equation can be written as

$$(\psi_r^{(+)}, \psi_s^{(+)}) = \left(\psi_r, \psi_s^{(+)} - \frac{1}{\mathcal{E}_s - H_0 + \mathrm{i}\hbar\alpha} V \psi_s^{(+)}\right). \tag{7.92}$$

Now, using (7.72), this finally becomes

$$(\psi_r^{(+)}, \psi_s^{(+)}) = (\psi_r, \psi_s) = \delta_{rs} \tag{7.93}$$

which establishes the orthonormality of the scattering states.

## References

Ferry D K 2000 *Semiconductor Transport* (London: Taylor and Francis)

Landau L D and Lifshitz E M 1958 *Quantum Mechanics* (London: Pergamon) pp 146, 150–3

Merzbacher E 1970 *Quantum Mechanics* (New York: Wiley)

Shockley W and Bardeen J 1950 *Phys. Rev.* **77** 407; **80** 72

## Problems

1. For the unperturbed wave functions and energies of section 6.2.1, estimate the decay rate

$$\Gamma_s = \frac{\mathrm{d}}{\mathrm{d}t}\sum_k |c_k|^2 .$$

2. The propagator $U(t, t_0)$ satisfies the integral equation (7.36). If $V(t) = V_0\delta(t - t_1)$, show that

$$U(t, t_0) = \frac{1}{1 + (\mathrm{i}V_0/\hbar)\Theta(t - t_1)}$$

where $\Theta(x)$ is the Heaviside step function; $\Theta(x) = 1$ for $x > 0$, and $\Theta(x) = 0$ for $x < 0$. (Hint: break up the integral into segments $0 < t < t_1 - \delta$, $t_1 - \delta < t < t_1 + \delta$, $t_1 + \delta < t < \infty$, where $\delta$ is a very small parameter that is assumed to vanish, and use the property (7.35).)

3. Verify (7.37$d$). (Hint: the multiple time integrals in each term must be transformed all to have the same limits of integration. This introduces a factorial in the pre-factor and also changes the multiple integrals from a nested set to a product of individual integrals. The series is then summed.)

4. At $t < 0$, an electron is assumed to be in the $n = 3$ eigenstate of an infinite square potential well, which extends from $-a/2 < x < a/2$. At $t = 0$, an electric field is applied, with the potential $V = Ex$. The electric field is then removed at time $\tau$. Determine the probability that the electron is in any other state at $t > \tau$. Do not make any assumptions about the relative size of $\omega_{k3}\tau$. What are the differences for the cases in which this latter quantity is small or is large?

5. Assume that an incident electron in a solid, in a state characterized by the wave vector $\boldsymbol{k}$, is scattered by the motion of the atoms of the solid. This scattering from the acoustic waves in the solid is characterized by a perturbing potential that is independent of the scattering wave vector $\boldsymbol{q}$. Convert the sum over the final states into a sum over the final-state energies, and show that the scattering rate (the total transition rate $\Gamma_s$) is directly proportional to the magnitude of the incident wave vector.

6. Using only ionized impurity scattering (chapter 6) and acoustic deformation potential scattering, so that the average relaxation time can be easily computed, analyse the data of Tyler W W and Woodbury H H 1956 *Phys. Rev.* **102** 647, for n-type germanium. Treat the impurity concentration and the deformation potential as adjustable parameters for each sample and tabulate the results (you should have a single value for each of these two numbers for each sample and the deformation potential should not vary from sample to sample).

7. A particular semiconductor has a zero-field mobility composed of ionized impurity scattering of $3500\ \mathrm{cm}^2\ \mathrm{V}^{-1}\ \mathrm{s}^{-1}$ and of acoustic scattering of

$4500\ \mathrm{cm}^2\ \mathrm{V}^{-1}\ \mathrm{s}^{-1}$. Assume that the average energy of the carriers is given approximately by

$$\tfrac{3}{2}k_{\mathrm{B}}(T_e - T) = e\mu F^2\tau_E$$

with $\tau_E = 10^{-12}$ s ($\tau_E$ is an energy relaxation time). Plot $\mu$ as a function of the electric field at 22 K.

# Chapter 8

# Motion in centrally symmetric potentials

The problem of interacting particles can usually be reduced in quantum mechanics to that of one particle, as can be done in classical mechanics. Normally, this problem is one of an electron orbiting around, or being affected by, a positive atomic core or a scattering centre with the interaction governed by the Coulomb potential. In general, this is a multi-dimensional problem, and not one of the simpler one-dimensional problems with which we have been concerned in the previous chapters. Once we begin to treat multiple dimensions, then state degeneracy begins to arise more frequently, and the most common problem treated is that of the hydrogen atom. These extra dimensions provide more degrees of freedom and more complexity. In this chapter, we want to discuss the motion of a charged particle in a centrally symmetric potential, and so will discuss the hydrogen atom. First, however, we want to begin to understand how the degeneracies arise and just what they mean. To facilitate this, we will treat first the harmonic oscillator for a two-dimensional motion and potential. We will then consider the manner in which the degeneracies are split by a magnetic field. Following this, we will be ready to discuss the hydrogen atom with its three-dimensional potential. Finally, we will briefly discuss the energy levels that arise in atoms more complex than the hydrogen atom with real, but non-coulombic potentials.

## 8.1 The two-dimensional harmonic oscillator

The harmonic oscillator in one dimension was discussed in chapter 4. This simple problem of a particle in a quadratic potential energy is one of the typical problems of quantization. In two (or more) dimensions, the problem is more interesting, but not particularly more complicated. We want to treat only two dimensions here in order to understand better just how the degeneracies arise. However, as we will see, the two-dimensional problem is of particular interest for electrons in semiconductor interfaces. Here, this typical two-dimensional problem, similar to the one that we will consider, can arise when electrons are confined in an inversion layer at the interface between, for example, silicon and silicon dioxide or at the interface of a heterostructure or in a potential well in the direction normal to that considered here (see section 2.6 for the first example). In any case, it is assumed that there is no $z$-motion, due to confinement of the carriers in this dimension. Further, it will be assumed for simplicity that centrally symmetric properties of the potential require that the 'spring' constants of the harmonic potential are the same in the two free coordinates.

### 8.1.1 Rectangular coordinates

We begin by treating the two dimensions as existing along two perpendicular axes in a normal rectangular coordinate system. Thus, the Hamiltonian may be obtained by expanding (4.9) to two dimensions, and becomes

$$H = -\frac{\hbar^2}{2m}\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{\mathrm{d}^2}{\mathrm{d}y^2}\right] + \frac{1}{2}m\omega^2(x^2 + y^2). \tag{8.1}$$

In order to simplify the results, we will introduce the creation and annihilation operators from (4.60), but with one set for the $x$-coordinates (denoted with a subscript $x$) and one set for the $y$-coordinates (with a subscript $y$). With the introduction of these operators, equation (8.1) may simply be expressed as

$$H = \hbar\omega(a_x^+ a_x + \tfrac{1}{2}) + \hbar\omega(a_y^+ a_y + \tfrac{1}{2}) = \hbar\omega(n_x + n_y + 1) \tag{8.2}$$

where we have introduced the number operator $n = a^+ a$, introduced in section 4.4.

We note from (8.2) that the lowest energy level $\mathcal{E}_0$ is just $\hbar\omega$, which arises for $n_x = n_y = 0$. This level arises from just the zero-point motion of the harmonic oscillators, and there is just one possible state that contributes to this level (the state where both number operators are identically zero). Thus, there is no degeneracy in this lowest energy level. On the other hand, the next highest energy level

$$\mathcal{E}_1 = 2\hbar\omega \tag{8.3}$$

has a double degeneracy since it can arise from two combinations of number operators, $n_x = 1$ and $n_y = 0$, as well as $n_x = 0$ and $n_y = 1$. The third energy level has the value

$$\mathcal{E}_2 = 3\hbar\omega \tag{8.4}$$

and is triply degenerate, since it arises from the three combinations of values that yield $n_x + n_y = 2$. This discussion can be continued to higher energy levels, from which it is found that the energy value of the $n$th level is

$$\mathcal{E}_n = (n + 1)\hbar\omega \tag{8.5}$$

and has an $(n + 1)$-fold degeneracy from the multiple ways in which the two number operators can be combined to yield $n_x + n_y = n$.

How do we determine the $x$- and $y$-axes? The problem has no central property that allows us to determine uniquely the orientation of either of these axes. Thus, the selection of some arbitrary direction for the $x$-axis is one of convenience, but not one of basic physics. This means that the basic properties of the harmonic oscillator are not those associated with these axes. Rather, we will find that the various degeneracies arise from the angular motion of the particle, while the energy level is basically set by the radial motion (in cylindrical coordinates). To illustrate this better, we will change variables and work the problem in cylindrical coordinates, and this is done in the next section.

### 8.1.2 Polar coordinates

If we specify the oscillator merely by its energy Hamiltonian (8.5), we cannot specify a complete set of commuting observables; that is, there is some degeneracy in the specification. This is because there are a number of combinations of $x$- and $y$-oscillator states that can combine into any single value of $n$. If we expand the total Hamiltonian into polar coordinates (cylindrical coordinates with no $z$-variation), and separate the resulting Schrödinger equation into the radial and the angular parts, we will find that the radial part generally determines the energy level, and the degeneracy goes into the various angular variations that result. For example, if we consider the $n = 2$ level of (8.5), one of the solutions arises for $n_x = n_y = 1$. Thus, each of the two one-dimensional harmonic oscillators is in the first excited state. How do we adjust the phase difference between the two oscillations? This phase difference can make the total oscillation actually rotate in the $(x, y)$ plane as a phasor. The rotation of the oscillation amplitude corresponds to angular momentum, and it is this momentum operator that arises from the angular parts of the Hamiltonian in polar coordinates. By treating this angular momentum, we can understand the overall motion of the two-dimensional harmonic oscillator in polar coordinates.

Angular momentum in classical mechanics arises from the motion of a body around some centre of rotation. This is defined by the vector

$$\boldsymbol{L} = \boldsymbol{r} \times \boldsymbol{p}. \tag{8.6}$$

In the two-dimensional harmonic oscillator in quantum mechanics, these variables become operators, but the only component (there is no $z$-variation in either position or momentum in our current approach) that arises is $L_z$. This is given by

$$L_z = xp_y - yp_x. \tag{8.7}$$

We want to put this into the operator notation used in the past section, and so we introduce the operators defined in (4.60) for each coordinate as

$$\begin{aligned} L_z &= -\mathrm{i}\frac{\hbar}{2}(a_x + a_x^+)(a_y - a_y^+) + \mathrm{i}\frac{\hbar}{2}(a_y + a_y^+)(a_x - a_x^+) \\ &= -\mathrm{i}\hbar(a_x^+ a_y - a_y^+ a_x) \end{aligned} \tag{8.8}$$

where we have used the various commutation relations for these operators

$$[a_x^+, a_y^+] = [a_x^+, a_y] = [a_y^+, a_x] = [a_y, a_x] = 0 \tag{8.9}$$

and the commutators between operators with the same subscript satisfy (4.60). We note that these operators also satisfy

$$[a_x^+ a_y, a_x^+ a_x + a_y^+ a_y] = 0 \tag{8.10$a$}$$

$$[a_y^+ a_x, a_x^+ a_x + a_y^+ a_y] = 0. \tag{8.10$b$}$$

Using the total Hamiltonian (8.2) and (8.8), this last result shows that the $z$-component of the angular momentum commutes with the Hamiltonian. Since the energy (8.5) has a set of degeneracies, these must correspond to different values of the angular momentum. In the lowest level ($n = 0$), there is only one state which must correspond to a single angular momentum value. Because there is automatically a degeneracy between positive and negative angles of rotation (e.g., for every state with positive angular momentum, there must be a state with the opposite negative value of angular momentum, as there is nothing in the problem to give a preferred rotation direction), this single level must have $L_z = 0$. In the next energy level ($n = 1$), there are two degenerate states, one for each of the $x$- and $y$-axes. These must also correspond, in polar coordinates, to a non-zero value of angular momentum, with one state for each direction of rotation. This can be continued to the higher energy levels. We shall therefore seek to find the basis of the eigenvalues for the total Hamiltonian $H$ and the angular momentum $L_z$. We will basically follow the treatment of Cohen-Tannoudji *et al* (1977).

In electromagnetic fields, it is useful to decompose linear polarized waves into left- and right-circularly polarized waves. These circularly polarized waves have angular momentum just as we are discussing here. This suggests that we introduce rotating creation and annihilation operators

$$a = \frac{1}{\sqrt{2}}(a_x - \mathrm{i}a_y) \tag{8.11$a$}$$

$$b = \frac{1}{\sqrt{2}}(a_x + \mathrm{i}a_y) \tag{8.11$b$}$$

and comparably for their adjoints

$$a^+ = \frac{1}{\sqrt{2}}(a_x^+ + \mathrm{i}a_y^+) \tag{8.12$a$}$$

$$b^+ = \frac{1}{\sqrt{2}}(a_x^+ - \mathrm{i}a_y^+). \tag{8.12$b$}$$

We note from these definitions that both $a$ and $b$ will produce the result that

$$a|n_x, n_y\rangle = c_1|n_x - 1, n_y\rangle + c_2|n_x, n_y - 1\rangle \tag{8.13}$$

where the $c_i$ are constants. This means that the operation of $a$ or $b$ produces a linear combination of states that are at the energy level one quantum ($\hbar\omega$) down, and so removes one quantum of energy from the system just as the isolated component of an annihilation operator in a single dimension does from that coordinate's harmonic oscillator. Similarly, the adjoint operators (creation operators) increase the energy of the state by one unit of energy $\hbar\omega$. We will see below that this energy reduction is accompanied by a corresponding change in the angular momentum of the state.

The rotational operators satisfy a commutation relation that can be obtained from those of the independent ones for each axis, and

$$[a, a^+] = [b, b^+] = 1 \tag{8.14}$$

$$[a, b] = [a, b^+] = [a^+, b] = [a^+, b^+] = 0. \tag{8.15}$$

We also note that (8.11) and (8.12) can be used to give

$$a^+ a = \tfrac{1}{2}(a_x^+ a_x + a_y^+ a_y - \mathrm{i}a_y^+ a_x + \mathrm{i}a_x^+ a_y) \tag{8.16$a$}$$

$$b^+ b = \tfrac{1}{2}(a_x^+ a_x + a_y^+ a_y + \mathrm{i}a_y^+ a_x - \mathrm{i}a_x^+ a_y). \tag{8.16$b$}$$

Thus, we can write the Hamiltonian as

$$H = (a^+ a + b^+ b + 1)\hbar\omega = (n_a + n_b + 1)\hbar\omega. \tag{8.17}$$

Here, we have introduced the equivalent number operators for each of the rotational operator pairs. In addition, we can rewrite the angular momentum as

$$L_z = (a^+ a - b^+ b)\hbar = (n_a - n_b)\hbar. \tag{8.18}$$

Here, the Hamiltonian remains in a form as simple as that for rectangular coordinates, and the angular momentum has been considerably simplified.

Using the operators $a$ and $b$, we can go through the complete arguments of section 4.4 to determine the wave functions. This will lead to the definition of the orthonormal states

$$|n_a, n_b\rangle = \frac{1}{\sqrt{(n_a)!(n_b)!}}(a^+)^{n_a}(b^+)^{n_b}|0, 0\rangle \tag{8.19}$$

and these states are eigenstates of *both* the Hamiltonian and the angular momentum. The normal parameters, such as the energy index in (8.6), are given by the integers

$$n = n_a + n_b \qquad m = n_a - n_b. \tag{8.20}$$

We note that if we act with the operator $a^+$ on (8.19), we not only raise the energy by one unit (we increase $n_a$ by one unit), we also raise the value of $m$ by one unit, hence increasing the angular momentum by one unit of Planck's (reduced) constant. Note that operating with $b^+$ raises the energy but reduces the angular momentum by decreasing $m$ through increasing $n_b$. This tells us that this operator corresponds to the positive (anti-clockwise) direction of rotation, while the other operator set refers to the negative (clockwise) direction of rotation. The eigenvalues of the angular momentum are then $m\hbar$. The eigenvalues with a given value of $n$ are $(n+1)$-fold degenerate, since we can have

$$\begin{aligned} &n_a = n && n_b = 0 \\ &n_a = n-1 && n_b = 1 \\ &n_a = n-2 && n_b = 2 \\ &\ldots && \\ &n_a = 1 && n_b = n-1 \\ &n_a = 0 && n_b = n. \end{aligned} \tag{8.21}$$

Now, $m$ can be a positive or negative integer, since it can range from $n_{a,\max}$ to $-n_{b,\max}$, or from $n$ to $-n$. For $n = 0$, there is only a single level that requires $m = 0$. For the next energy level, $n = 1$ and there are two levels. Thus, $m = \pm 1$. For the third level, $n = 2$ and there are three levels. Here, $m$ takes on the values 0 and $\pm 2$. Hence, for a given energy level, the allowed values of $m$ are

$$m = n, n-2, n-4, \ldots, -n+2, -n. \tag{8.22}$$

Hence, any particular state is specified by the integers $n$ and $m$, and the wave function is defined through

$$\left| n_a = \frac{n+m}{2}, n_b = \frac{n-m}{2} \right\rangle . \tag{8.23}$$

We now need to turn our attention to finding the lowest wave function, as the higher wave functions can be obtained from this lowest level with (8.19).

In seeking the wave function for the ground state, we will now introduce the polar coordinates, through the quantities

$$\begin{aligned} x &= r\cos\phi && r \geq 0 \\ y &= r\sin\phi && 0 \leq \phi \leq 2\pi. \end{aligned} \tag{8.24}$$

Using (4.60) and (8.11), we find that

$$a = \frac{\mathrm{e}^{-\mathrm{i}\phi}}{2}\left[Br + \frac{1}{B}\frac{\partial}{\partial r} - \frac{\mathrm{i}}{Br}\frac{\partial}{\partial \phi}\right] \tag{8.25}$$

and

$$b = \frac{\mathrm{e}^{\mathrm{i}\phi}}{2}\left[Br + \frac{1}{B}\frac{\partial}{\partial r} + \frac{\mathrm{i}}{Br}\frac{\partial}{\partial \phi}\right]. \tag{8.26}$$

Similarly,

$$a^{+} = \frac{\mathrm{e}^{\mathrm{i}\phi}}{2}\left[Br - \frac{1}{B}\frac{\partial}{\partial r} - \frac{\mathrm{i}}{Br}\frac{\partial}{\partial \phi}\right] \tag{8.27}$$

and

$$b^{+} = \frac{\mathrm{e}^{-\mathrm{i}\phi}}{2}\left[Br - \frac{1}{B}\frac{\partial}{\partial r} + \frac{\mathrm{i}}{Br}\frac{\partial}{\partial \phi}\right] \tag{8.28}$$

where

$$B = \sqrt{m\omega/\hbar}. \tag{8.29}$$

Note, that because of the rotating coordinates, *the adjoint operators are not simply complex conjugates of the annihilation operators*, but must be carefully calculated from the definitions themselves. Now, either $a$ or $b$ will attempt to lower the energy and change the angular momentum. The lowest eigenstate should satisfy both of these operators, as

$$a|0,0\rangle = b|0,0\rangle = 0. \tag{8.30}$$

The two operators $a$ and $b$ differ only by the sign of the imaginary part. Since (8.30) must satisfy both the real and imaginary parts simultaneously, it leads to

$$\frac{1}{Br}\frac{\partial}{\partial\phi}|0,0\rangle = 0 \tag{8.31$a$}$$

$$\frac{1}{B}\frac{\partial}{\partial r}|0,0\rangle = -Br|0,0\rangle. \tag{8.31$b$}$$

Equation $(8.31a)$ just leads to a function that is independent of the angle $\phi$. The second of these equations leads to $\exp(-B^2r^2/2)$ behaviour, just as for the linear harmonic oscillator. After normalization, the ground-state wave function is just

$$|0,0\rangle = \frac{B}{\sqrt{\pi}}\mathrm{e}^{-B^2r^2/2}. \tag{8.32}$$

The higher-lying levels are now found by using (8.19). We note that angular variation is introduced into the wave function by the operators $a^{+}$ and $b^{+}$ themselves through the exponential pre-factors. For example, the wave function for the $n = m = 1$ state is given by

$$|1,0\rangle = a^{+}|0,0\rangle = \frac{B}{\sqrt{\pi}}Br\mathrm{e}^{-B^2r^2/2}\mathrm{e}^{\mathrm{i}\phi} \tag{8.33}$$

while the state $|0,1\rangle$ ($n = 1$, $m = -1$) is given by the complex conjugate of this expression. Thus, the two angular momentum states have counter-rotating properties, but both are made up of linear combinations of the linear harmonic oscillators along the two axes. In the next section, an additional potential due to a magnetic field will be used to raise the degeneracy of the two angular momentum states.

### 8.1.3 Splitting the angular momentum states with a magnetic field

The motion of an electron in a magnetic field was discussed earlier in section 4.7, where it was shown that the magnetic motion also introduces a harmonic oscillator potential. Here, we want to examine the coupling of the magnetic harmonic oscillator potential with the two-dimensional harmonic oscillator potential of (8.1). The first people to study the two-dimensional harmonic oscillator in a magnetic field were apparently Fock (1928) and Darwin (1931), but the operator approach that we want to use follows the treatment of Rössler (1991). The total Hamiltonian can be found by coupling (8.1) with (4.107), and

$$H = \frac{1}{2m}(\boldsymbol{p} + e\boldsymbol{A})^2 + \frac{1}{2}m\omega^2(x^2 + y^2). \tag{8.34}$$

Because of the two-dimensional nature of the electrostatic harmonic oscillator, it is more convenient to take the vector potential in the symmetric gauge, rather than the Landau gauge used in section 4.7, with $\boldsymbol{A} = (-By, Bx, 0)/2$, so that the magnetic field is oriented in the $z$-direction, normal to the two-dimensional plane of the motion ($\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A} = B\boldsymbol{a}_z$). The momentum term can now be expanded as

$$(\boldsymbol{p} + e\boldsymbol{A})^2 = -\hbar^2 \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) - \mathrm{i}\hbar eB \left( x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x} \right) + \frac{1}{4}e^2B^2(x^2 + y^2). \tag{8.35}$$

The last term can be combined with the electrostatic harmonic oscillator if we define the new oscillator 'spring' frequency

$$\Omega = \sqrt{\omega^2 + \left( \frac{\omega_\mathrm{c}}{2} \right)^2} \tag{8.36}$$

where $\omega_\mathrm{c} = eB/m$ is the cyclotron frequency introduced near (4.112). The second term in (8.35), which is linear in the magnetic field, can be rewritten using (8.7) as

$$eBL_z. \tag{8.37}$$

With these definitions, the Hamiltonian, with the change in frequency in the definitions of the operators as $\omega \to \Omega$, can then be rewritten as

$$H = (n_a + n_b + 1)\hbar\Omega + \tfrac{1}{2}\omega_\mathrm{c}L_z = (n + 1)\hbar\Omega + \tfrac{1}{2}\hbar\omega_\mathrm{c}m \tag{8.38}$$

where $m = n_a - n_b$ is the angular momentum quantum number, and not the mass that appeared in the some of the early equations in this section. The reader should check that this new Hamiltonian still commutes with the angular momentum, and hence that both are simultaneously measurable. However, the energy now has a contribution from the angular momentum directly. This term raises the degeneracy of the previous eigenvalues (energies).

For small values of the magnetic field, $\omega_\mathrm{c} \ll \Omega$, each energy level is split into $n + 1$ levels as the degeneracy is raised by the magnetic field. Each pair of these levels is separated by the amount $\hbar\omega_{\mathrm{c}}$, since $\delta m$ was determined in the previous section to be 2. As the magnetic field value is increased, this spread in energies changes significantly.

For slightly larger values of the magnetic field, but still with $\omega_{\mathrm{c}} \ll \Omega$, the uppermost energy level is described by the frequency

$$\omega_{+} = (n+1)\left(\omega + \frac{\omega_{\mathrm{c}}^2}{4\omega}\right) + \frac{m\omega_{\mathrm{c}}}{2} \tag{8.39}$$

which is only slightly above the degenerate energy level in the absence of the magnetic field. The lowest energy level is now described by the frequency

$$\omega_{-} = (n+1)\left(\omega + \frac{\omega_{\mathrm{c}}^2}{4\omega}\right) - \frac{m\omega_{\mathrm{c}}}{2} \tag{8.40}$$

so all the levels begin to show a weak quadratic upward motion away from the linear spreading of the levels for very small magnetic fields.

In the case of very large magnetic fields, $\omega_{\mathrm{c}} \gg \Omega$, the energy levels are quite different. The energy levels are now given by the frequencies

$$\omega_{n,m} = (n+m+1)\frac{1}{2}\omega_{\mathrm{c}} + (n+1)\frac{\omega^2}{\omega_{\mathrm{c}}}. \tag{8.41}$$

If we neglect the last term, we recover the Landau level energies, since $n + m = 2n_a$ is an even integer. Moreover, we note that for example for the lowest Landau level, where $n + m = 0$, one level from each of the electrostatic harmonic oscillator levels (when $B = 0$) merges into the Landau level (see figure 8.1). Similarly, for the next higher Landau level, one level from each electrostatic harmonic oscillator level for which $n \geq 1$ converges into the Landau level. This continues upward throughout the spectrum, with the $i$th Landau level being formed from states arising from the levels for which $n \geq i$. Thus, as the magnetic field is increased in size, the energy levels move smoothly from those values associated with the electrostatic harmonic oscillator to those values associated with the Landau levels.

While a given Landau level has contributions from all equal-index and higher-index harmonic oscillator levels, a given harmonic oscillator level contributes to only a fixed number of Landau levels. The lowest harmonic oscillator level, for example, contributes only to the lowest Landau level, since it has only one non-degenerate state. This is the case for which $n + m = 0$. Similarly, each harmonic oscillator level contributes its lowest level to the lowest Landau level. However, its highest level ($m = n$) goes into the $n$th Landau level. So the first harmonic oscillator level contributes to the lowest two Landau levels, the second to the lowest three Landau levels, and so on. The spacing of the levels that merge into a particular Landau level is given by the frequency

$$\omega_{-} = \frac{\omega^2}{\omega_{\mathrm{c}}} \tag{8.42}$$

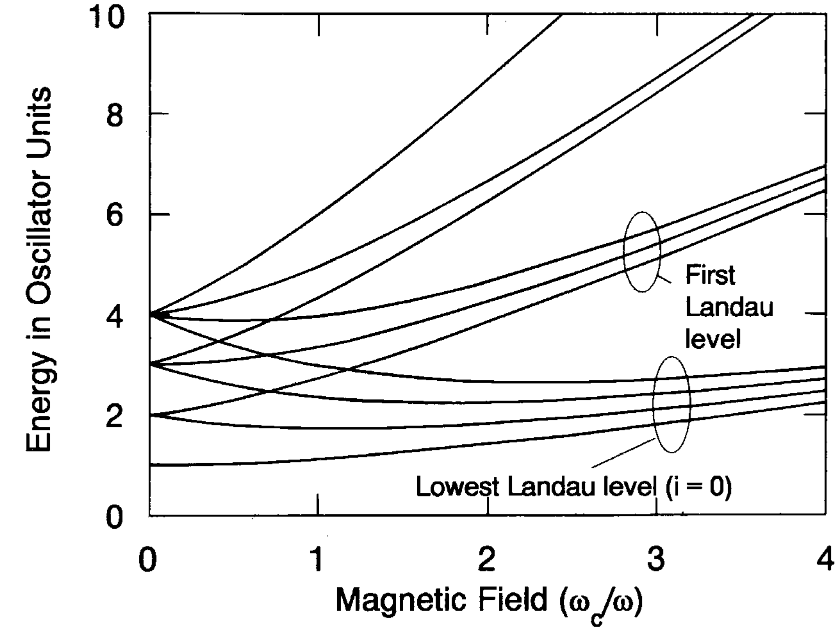


**Figure 8.1.** The energy levels of the two-dimensional harmonic oscillator in a magnetic field. Only the lowest four electrostatic levels are shown, for simplicity, and these are plotted in units of $\hbar\omega$.

so the electrostatic harmonic oscillator potential splits the degeneracy of the Landau levels. The spectral positions of these 'quantum box' levels have been measured by far-infrared absorption measurements for InSb by Sikorski and Merkt (1989) and via the conductance arising from single-electron charging in GaAs by McEuen *et al* (1991); we discuss this in the next section.

### 8.1.4 Spectroscopy of a harmonic oscillator

There has been considerable interest in small quantum-structured devices in semiconductors for some years, particularly where the size scale is much smaller than the inelastic mean free path. The latter quantity is the characteristic length over which an electron will lose phase coherency due to scattering from impurities or interaction with other electrons. Most of these studies are done at low temperature, so scattering by the lattice is greatly reduced, and these mean free paths may be a micrometre or many tens of micrometres in a quantizing magnetic field such as that being discussed above. One area of study has been the transport of electrons through nanometre-scale electron confinement structures that have been lithographically patterned, referred to as quantum dots. The structures are typically heterostructures between GaAs and AlGaAs, in which electrons move from donors in the AlGaAs to the GaAs and form an inversion layer at the interface. This inversion layer has confinement quantization in the direction normal to the heterostructure interface, just as discussed for Si–SiO$_2$ in section 2.6. The transport is then allowed in the two dimensions in the plane of the heterostructure interface. The lithographically defined confinement potential in two dimensions creates the equivalent of a two-dimensional harmonic oscillator, as shown in figure 8.2(*a*). A similar quantum dot was discussed in sections 1.3 and 3.3.3, where transport through a quantum dot was used to modulate the Aharonov–Bohm interference (Yacoby *et al* 1994, 1995). In that example, transport through the quantum dot states was used to modulate the tunnelling phase of the particles. In the present case, the tunnelling is used to actually do *spectroscopy* of the quantum dot states. The so-called edge states are the drift of the electrons in the strong magnetic field around the periphery of the quantum dot. The guiding centre orbit is formed by the cyclotron motion of the electron being interrupted by the edge of the dot; that is, they specularly reflect from the confining potential, so they bounce along the interface. These guiding centre orbits are called *edge states* (section 4.7.2).

The spectroscopy of the quantum dot arises from the motion of the Fermi energy in a magnetic field. The Fermi energy in a semiconductor, at low temperature (the experiments discussed here were performed with the sample at approximately 0.3 K), is the energy level at which all states below this energy are full and all states above this energy are empty. In figure 8.2(*b*), an enlarged section of figure 8.1 is shown. The heavy black line is the Fermi energy. The density of electrons is pre-set in the sample, so this defines precisely the number of filled states in the quantum dot, which is two (for spin degeneracy) times the number of energy levels below the Fermi energy. As the magnetic field is increased, and a state crosses the Fermi level from above—for example just above the point $\omega_c/\omega_0 = 7.4$ (our $\omega$ is $\omega_0$ in this figure)—the Fermi energy must follow this state downward allowing the upward-moving state to pass above the Fermi level so as to keep the number of levels below the Fermi energy fixed.

As the gate voltage applied to the confinement potential, or more exactly to a uniform back metal contact, is changed the number of electrons in the quantum dot can be changed. As an electron enters or leaves the dot through the end couplings, a conductance peak along the channel is observed, which can be followed while varying the magnetic field. This is shown in figure 8.3(*a*). The nearly constant Coulomb energy of the electrons in the dot can be subtracted out, and the gate voltage converted to energy to unfold the experimental energy spectrum, which is shown in figure 8.3(*b*). This should be compared with the theoretical spectrum expected for a two-dimensional harmonic oscillator in a magnetic field shown in figure 8.1 and figure 8.2(*b*). The agreement as regards the shape of the spectrum is remarkable.

This experiment samples the spectra for rather large energy and magnetic field. More recent studies have actually looked at the spectra for small numbers

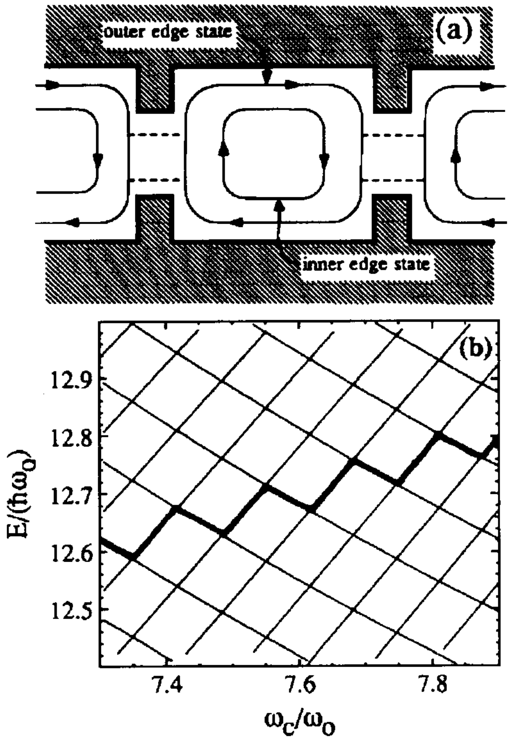


**Figure 8.2.** (*a*) Schematic view of the device, showing the edge states traversing the periphery of the quantum dot of dimensions approximately 0.5 mm × 0.7 mm. (*b*) Energy levels and motion of the Fermi level (dark line) for a fixed density of carriers in the dot. (After McEuen *et al* (1991), with permission.)

of electrons, which is the case that was shown in figure 8.1 (Tarucha *et al* 1996). Here, a double-barrier resonant-tunnelling diode, described in section 3.5.3, was utilized. This structure used a 12 nm $\mathrm{In_{0.05}Ga_{0.95}As}$ quantum well, $\mathrm{Al_{0.22}Ga_{0.78}As}$ barriers of 9.0 and 7.5 nm thickness, and GaAs cladding layers outside the barriers. The structure is shown in figure 8.4, where a pillar region has been etched, so that the resulting quantum dot has a diameter of 0.5 $\mu$m. The sidegate is used to deplete fully the quantum well, and then the bias is slowly backed off to allow single electrons to tunnel into the well. Since the dot is small, this charging proceeds by single-electron tunnelling, as described in section 3.8.3. In this case, the tunnelling comes from the bulk region through one of the AlGaAs barriers, as the resonant levels of the dot are swept past the source–drain bias by the sidegate bias. The resulting current is shown in figure 8.5. Clear Coulomb oscillations are shown in the figure, and the numbers of electrons in

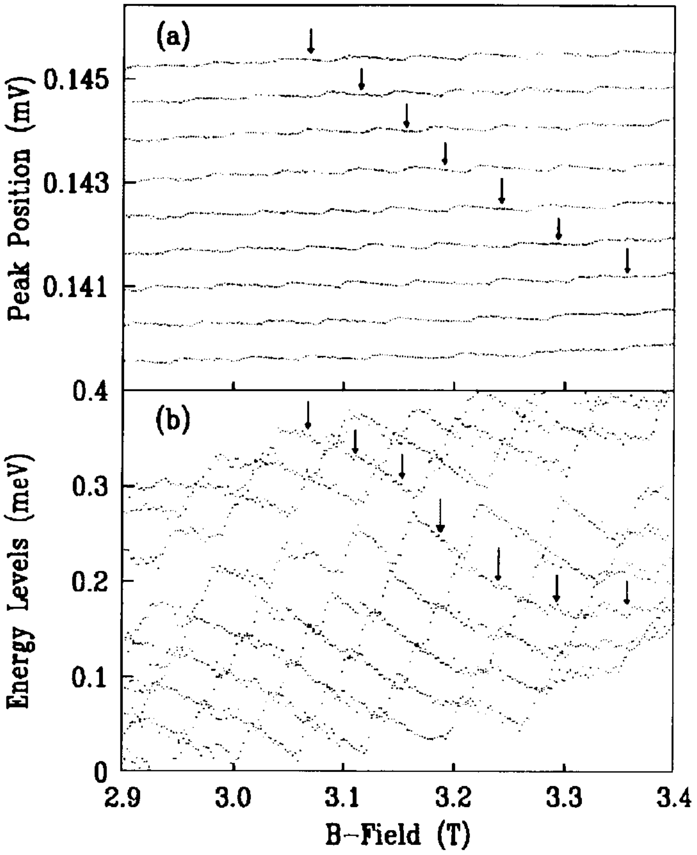


**Figure 8.3.** (*a*) Peak position of the longitudinal conductance peak as a function of the magnetic field for a series of conductance peaks. The arrow follows a particular state in the first Landau level. (*b*) Energy spectrum inferred from (*a*), with an arbitrary zero of energy. (After McEuen *et al* (1991), with permission.)

the dot are indicated. Rather larger gaps are seen for $n = 2, 6, 12, \ldots$, which is in keeping with the states described in (8.21), although the magnetic field has an effect, which will be described below. The spectrum of figure 8.5 can be understood by realizing that the basic single-electron charging energy $e^2/2C$ is being modified by the discrete energy levels of the dot itself. The gaps that appear at $n = 2, 6, 12, \ldots$ correspond to complete filling of the first, second, third, ... shells for the single-particle states. These atomic-like properties of the states can be further elucidated by the magnetic field behaviour. One would expect in this approach that the minimum gap between charging states, for example between $n = 6$ and $n = 12$, corresponds to the Coulomb energy $e^2/2C$, and the additional

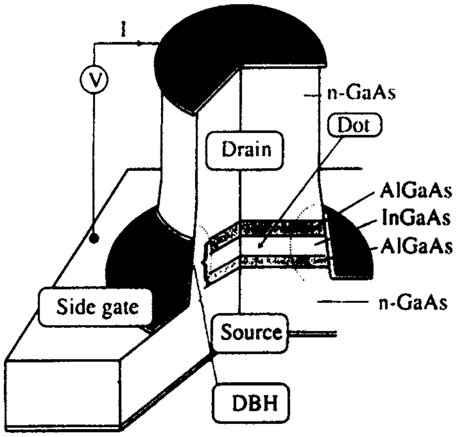


**Figure 8.4.** A schematic diagram of the resonant tunnelling diode used in the study of the energy levels of the harmonic oscillator. (After Tarucha *et al* (1997), by permission.)

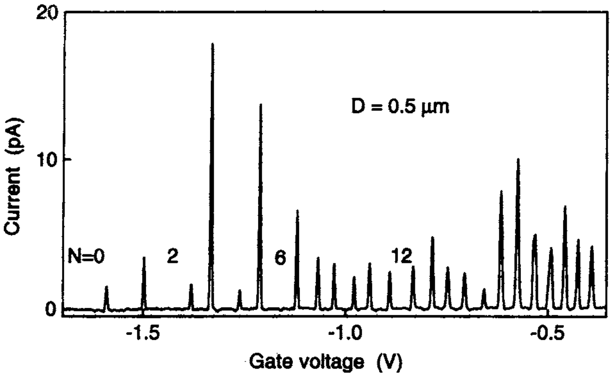


**Figure 8.5.** Coulomb oscillations in the current as a function of the gate voltage at zero magnetic field for a 0.5 $\mu$m diameter dot. (After Tarucha *et al* (1997), by permission.)

amounts of energy correspond to complicated many-body effects in the dot as well as the discrete energy level spacing. For example, the splitting between the $n = 1$ and $n = 2$ levels is larger than the Coulomb energy, but the simple theory of (8.21) would suggest that there was no additional dot-induced separation of these two levels—they should be the two spin-degenerate levels of the first shell. Yet, there is a rather large splitting between these two levels.

In figure 8.6, the effect of a magnetic field is illustrated. In panel (*a*), the energy level scheme of figure 8.1 is repeated with parameters appropriate for this quantum dot. In panel (*b*), the charging current is shown in a parameter space that

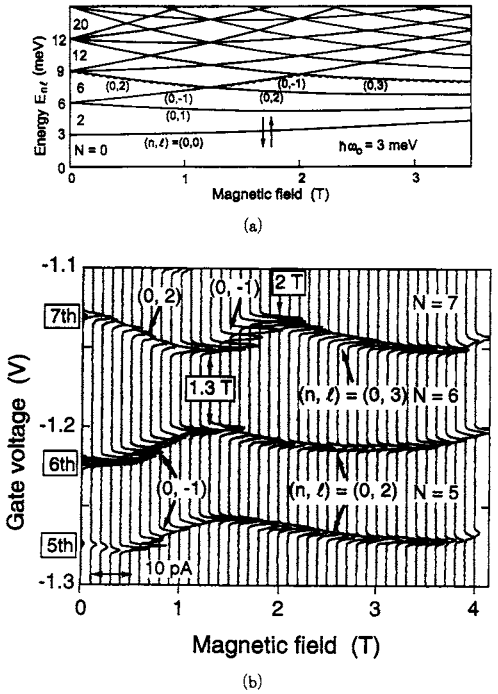


**Figure 8.6.** (*a*) Calculated single-particle energy levels in the Darwin–Fock spectrum as a function of the magnetic field, with $\hbar\omega_0 = 3$ meV. Each state is twofold degenerate, and the broken line is discussed in the text. (*b*) Experimentally determined evolution of the fifth, sixth, and seventh tunnelling current peaks as a function of the magnetic field. (After Tarucha *et al* (1997), by permission.)

plots gate voltage versus magnetic field (Tarucha *et al* 1997). Here, the magnetic field is held fixed at a value and the gate voltage swept, and there are some 41 of these sweeps in the figure. The Coulomb charging energy has not been subtracted from the gate voltage, so the movements of the peaks are kept some distance apart from one another (in gate voltage). The peaks for the fifth, sixth, and seventh electrons are shown in these plots. The seventh electron is characterized by the broken curve in panel (*a*), and it can be seen that this charging peak changes character as one raises the magnetic field. That is, the seventh electron goes into the $(n, l) = (0, 2)$ state at low magnetic field, then begins to occupy the $(0, -1)$ state for a field larger than the level crossing at 1.3 T. Finally, a second level crossing occurs at about 2 T, whereupon the seventh electron begins to fill the $(0, 3)$ level arising from the fourth shell. At the same time, the spin-degenerate fifth and sixth electrons begin in the $(0, -1)$ level and transition to the $(0, 2)$ level for magnetic fields beyond the 1.3 T crossing. The splitting between these latter two levels, in gate voltage, is the Coulomb charging energy. In essence, this same behaviour was occurring in the McEuen *et al* (1991) dots, but was much harder to ascertain due to the much larger number of electrons within the dot. The spectra of Tarucha *et al* (1996, 1997) are among the nicest examples of spectra demonstrating the Darwin–Fock energy levels in the two-dimensional harmonic oscillator. Using vertical tunnelling, these experimenters have been able to clearly probe the single-electron spectra and electronic states of a few-electron quantum dot. This yields the typical shell spectrum expected.

## 8.2 The hydrogen atom

In three dimensions, we could extend the above treatment to a three-dimensional harmonic oscillator. It is of more interest, however, to turn our attention to the motion of an electron about the nucleus of an atom. The case of interest is that of the hydrogen atom, really the only problem that can be solved exactly (the motion of two electrons about a nucleus includes the interaction between the two electrons and produces a three-body problem, which is well beyond the scope of this text). For simplicity, we will deal only with the *relative* motion of the electron about the nucleus, taking the latter as fixed in position. The atom is centrally symmetric in three dimensions and the natural coordinate system is spherical coordinates.

The electron is attracted to the nucleus (which will be assumed to be immensely massive in comparison with the electron) through a Coulomb potential, so the Schrödinger equation is simply

$$H\Psi = \left[-\frac{\hbar^2}{2m}\nabla^2 - \frac{e^2}{4\pi\varepsilon r}\right]\Psi = \mathcal{E}\Psi \tag{8.43}$$

where $m$ is the electron mass. To avoid confusion with one of the integers describing the angular momentum in the subsequent material, we will take the mass as the reduced mass $\mu = mM/(m+M)$, where $M$ is the nuclear mass. In spherical coordinates, (8.43) can be rewritten as

$$\begin{aligned}\mathcal{E}\Psi = & -\frac{\hbar^2}{2\mu r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Psi}{\partial r}\right) - \frac{e^2}{4\pi\varepsilon r}\Psi \\ & -\frac{\hbar^2}{2\mu r^2}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Psi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2\Psi}{\partial\theta^2}\right].\end{aligned} \tag{8.44}$$

The traditional method of solving this is to assume that $\Psi(r,\theta,\phi)$ can be split into two product terms as $R(r)\psi(\theta,\phi)$. By inserting this wave function, and then dividing by the wave function, (8.44) can be split into two terms, one a function of $r$ alone and the other a function of $\theta$ and $\phi$ alone. If this is to be true for all values of the position and angles, each term must be equal to a constant. Thus, the term in the square brackets is a constant, say $\lambda$, times the wave function. While this constant is arbitrary, it is known from treatments of spherical harmonics that it is more appropriate to set this separation constant equal to $\ell(\ell+1)$. It is easy to show that there is a one-to-one correspondence to any value of $\lambda$ and $\ell$, so long as $\lambda > 0$, so no generality is lost by making this latter substitution.

### 8.2.1 The radial equation

By eliminating the angular variables with the above constant interpretation, which still must be shown to be valid (and we do this below), the radial equation now becomes

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial R}{\partial r}\right) - \frac{\ell(\ell+1)}{r^2}R + \frac{2\mu}{\hbar^2}\left[\mathcal{E} + \frac{e^2}{4\pi\varepsilon r}\right]R = 0. \qquad (8.45)$$

To simplify the equation, we introduce the reduced units for energy and length

$$\alpha^2 = \frac{8\mu\mathcal{E}}{\hbar^2} \qquad (8.46a)$$

and

$$\rho = \alpha r \qquad (8.46b)$$

respectively, and set

$$\lambda = \frac{\mu e^2}{2\pi\varepsilon\alpha\hbar^2} = \frac{e^2}{4\pi\varepsilon\hbar}\sqrt{\frac{\mu}{2|\mathcal{E}|}}. \qquad (8.46c)$$

Then, we can rewrite (8.45) as

$$\frac{1}{\rho^2}\frac{\partial}{\partial\rho}\left(\rho^2\frac{\partial R}{\partial\rho}\right) + \left[\frac{\lambda}{\rho} - \frac{1}{4} - \frac{\ell(\ell+1)}{\rho^2}\right]R = 0. \qquad (8.47)$$

The choice of the number $\frac{1}{4}$ is arbitrary, but quite useful for the following development.

As in the treatment of the harmonic oscillator, we first seek the behaviour for large values of $\rho$. For sufficiently large values of the normalized radius, it is clear that the behaviour of $R$ is as $\rho^n \mathrm{e}^{-\rho/2}$. This follows by retaining just the second-order derivative and the term with $\frac{1}{4}$ as the pre-factor. The factor $n$ will have to be determined, though, but this suggests that we seek solutions of the form $R(\rho) = F(\rho)\mathrm{e}^{-\rho/2}$, where $F(\rho)$ will be a polynomial of finite order in $\rho$. Using this substitution leads to

$$\frac{\partial^2 F}{\partial\rho^2} + \left(\frac{2}{\rho} - 1\right)\frac{\partial F}{\partial\rho} + \left[\frac{\lambda-1}{\rho} - \frac{\ell(\ell+1)}{\rho^2}\right]F = 0. \qquad (8.48)$$

We now will assume that $F(\rho)$ varies as $\rho^s L(\rho)$, so (8.48) becomes

$$\rho^2\frac{\partial^2 L}{\partial\rho^2} + \rho[2(s+1)-\rho]\frac{\partial L}{\partial\rho} + [\rho(\lambda-s-1)+s(s+1)-\ell(\ell+1)]L = 0. \qquad (8.49)$$

If we evaluate this at $\rho = 0$ (we assume the derivatives are well behaved and are finite at this point), we see clearly that $s = \ell$ or $-(\ell + 1)$. Since we require the functions to be finite at $\rho = 0$, only the former value is allowed. Equation (8.49) then becomes

$$\rho \frac{\partial^2 L}{\partial \rho^2} + [2(\ell + 1) - \rho] \frac{\partial L}{\partial \rho} + (\lambda - \ell - 1)L = 0. \tag{8.50}$$

Finally, to proceed, we will assume that $L(\rho)$ is a power series in $\rho$, but one that is finite and terminates at some order. This leads to the ratio of coefficients of successive orders $\nu$ and $\nu + 1$:

$$a_{\nu+1} = a_\nu \frac{\nu + \ell + 1 - \lambda}{(\nu + 1)(\nu + 2\ell + 2)}. \tag{8.51}$$

For this to terminate, there must be a maximum value of $n = n'$ such that

$$\lambda = n = n' + \ell + 1. \tag{8.52}$$

Here, $n$ is the total quantum number (this and thereby $\lambda$ determines the total energy of the level), $n'$ is the radial quantum number and $l$ is the angular momentum quantum number; all are integers. With this total quantum number, the energy levels are given by (8.46*c*) as

$$\mathcal{E}_n = -\frac{\mu e^4}{32\pi^2 \varepsilon^2 \hbar^2 n^2}. \tag{8.53}$$

Thus, as in the two-dimensional harmonic oscillator, the energy levels are specified by the total quantum number, except that there is no $n = 0$ level here. This may be used in (8.46*a*) to show that the effective radius of the ground state is given by $2/\alpha = a_0 = 4\pi\varepsilon\hbar^2/\mu e^2 = 5.3 \times 10^{-9}$ cm. This quantity is called the *Bohr radius*. Note, however, that the actual normalization radius, and the factor $\alpha$, depend upon the index $n$.

The coupling between the radial quantum number and the angular momentum quantum number means that as the angular momentum increases, the radial variations decrease (the order of the polynomial decreases). Further, for the lowest energy level ($n = 1$), both the radial and angular momentum quantum numbers are zero, so the only variation is in the exponential term. In this lowest level, the wave function is spherically symmetric, decaying exponentially away from the centre of the atom. For the second level ($n = 2$), $\ell$ can be either $0$ ($n' = 1$) or $1$ ($n' = 0$), so there is a spherically symmetric state and one that has preferential directions due to the angular momentum. This continues, with $\ell$ taking on values $0, 1, \ldots, n - 1$, so there are $n$ values for this variable. Let us now turn to the angular momentum.

We can write out a few of the lower-level wave functions in terms of the components written as $R_{n,\lambda}(\rho)$, where $n$ and $\lambda$ are the two eigenvalue integers that we have discussed above. Thus, the lowest three energy levels have the unnormalized radial wave functions ($\rho = \alpha r$)

$$\begin{aligned}
&R_{1,0}(\rho) \propto \mathrm{e}^{-\rho/2}\\
&R_{2,0}(\rho) \propto (\rho - 2)\mathrm{e}^{-\rho/2}\\
&R_{2,1}(\rho) \propto \rho\mathrm{e}^{-\rho/2}\\
&R_{3,0}(\rho) \propto (\rho^2 - 6(\rho - 1))\mathrm{e}^{-\rho/2}\\
&R_{3,1}(\rho) \propto \rho(\rho - 4)\mathrm{e}^{-\rho/2}\\
&R_{3,2}(\rho) \propto \rho^2\mathrm{e}^{-\rho/2}.
\end{aligned} \tag{8.54}$$

We see that the highest-order term is determined by the radial quantum number $n$, but that the number of lower-order terms is related to the angular momentum quantum number $\lambda$. Note, again, that this is the normalized radius, and this normalization depends upon the level number $n$.

### 8.2.2 Angular solutions

The angular equation that arises from (8.44) is now the focus of our attention. This equation, separated for the angular parts, becomes

$$\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2} + \ell(\ell+1)\psi = 0. \tag{8.55}$$

This can be further separated by letting $\psi = \Theta(\theta)F(\phi)$. Using this form for the angular wave function allows us to separate (8.55) into two equations for the two variables:

$$\frac{\partial^2\Phi}{\partial\phi^2} + m^2\Phi = 0 \tag{8.56$a$}$$

$$\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\frac{\partial\Theta}{\partial\theta}\right) + \left[\ell(\ell+1) - \frac{m^2}{\sin^2\theta}\right]\Theta = 0 \tag{8.56$b$}$$

where we have let the separation constant be the square of an integer, so the variation in the azimuthal angle $\phi$ is periodic in $2\pi$:

$$\Phi(\phi) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{\mathrm{i}m\phi} \qquad |m| > 0 \tag{8.57}$$

and is a constant for $m = 0$. We note that $m$ can take on both positive and negative values, but that $\Phi$ must be continuous and have a continuous derivative for any value of $\phi$.

In addressing the last angular equation, it will be convenient to let $w = \cos\theta$. With this substitution, (8.56$b$) becomes

$$\frac{\mathrm{d}}{\mathrm{d}w}\left[(1 - w^2)\frac{\mathrm{d}\Theta(w)}{\mathrm{d}w}\right] + \left(\ell(\ell+1) - \frac{m^2}{1 - w^2}\right)\Theta(w) = 0. \tag{8.58}$$

Since the angle varies from 0 to $\pi$, the domain of $w$ is 0 to 1. To begin, we let the wave function be assumed to vary as $(1 - w^2)^s F(w)$. With this substitution, we find that

$$(1 - w^2)\frac{\mathrm{d}^2 F}{\mathrm{d}w^2} - 2w(2s + 1)\frac{\mathrm{d}F}{\mathrm{d}w} + (\ell(\ell + 1) - 2s(2s + 1))F = 0 \qquad (8.59)$$

with $4s^2 = m^2$. This leads us to say that $2s = |m|$. $F$ is now a polynomial of finite order, and for this to be the case, we must have the order of the polynomial equal to $\ell - 2s$. The parity of the polynomial (even or odd) is that of $\ell - 2s = \ell - |m|$. These polynomials are the associated Legendre polynomials, and the $\Theta(w)$ are the associated Legendre polynomials.

### 8.2.3 Angular momentum

In (8.6) and (8.7), we introduced the angular momentum, and particularly the $z$-component of this quantity. We want to consider now the overall angular momentum of the particle orbiting the central potential in the hydrogen atom. By continuation of (8.7), we can immediately write the three components of angular momentum as

$$L_x = -\mathrm{i}\hbar \left[y\frac{\partial}{\partial z} - z\frac{\partial}{\partial y}\right] \qquad (8.60a)$$

$$L_y = -\mathrm{i}\hbar \left[z\frac{\partial}{\partial x} - x\frac{\partial}{\partial z}\right] \qquad (8.60b)$$

$$L_z = -\mathrm{i}\hbar \left[x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right]. \qquad (8.60c)$$

It is easily then shown using the normal commutation relations between position and momentum that

$$[L_x, L_y] = \mathrm{i}\hbar L_z \qquad [L_y, L_z] = \mathrm{i}\hbar L_x \qquad [L_z, L_x] = \mathrm{i}\hbar L_y. \qquad (8.61)$$

Finally, the square of the total angular momentum can be written as

$$L^2 = L_x^2 + L_y^2 + L_z^2 \qquad (8.62)$$

and, moreover, this total angular momentum commutes with each of the components individually. Thus, we have four relatively independent quantities—the three components of the angular momentum and the total angular momentum—that can all be measured simultaneously. Normally, we can only specify two of these independently (we have only two parameters, $l$ and $m$, that arise in the treatment of the orbital motion).

Let us first examine $L_z$, since the second of the parameters, $m$, is related to the motion about the polar axis, the $z$-axis as it were. If we convert (8.60$c$) to spherical coordinates, with $x = \rho \sin\theta \cos\phi$, $y = \rho \sin\theta \sin\phi$, $z = \cos\theta$, then we find that

$$L_z = -\mathrm{i}\hbar \frac{\partial}{\partial \phi}. \tag{8.63}$$

Obviously, this involves only the quantum number $m$, so $L_z$ is a good component of the angular momentum to take as one of the independent quantities in the problem. The eigenvalues of the $z$-component of the angular momentum are then $\pm \mathrm{i} m\hbar$, with $m$ an integer. From the discussion of (8.59), it is clear that $|m| \leq l$, so the values of $m$ are

$$-\ell, -\ell + 1, \ldots, -1, 0, 1, \ldots, \ell - 1, \ell. \tag{8.64}$$

Thus, there are $2\ell + 1$ values for the $z$-component of angular momentum, for a given value of $\ell$.

We expect that the final independent quantity will be the total angular momentum. Indeed, this will be the case. To see this, we begin by taking combinations of the $x$- and $y$-components:

$$L_\pm = L_x \pm \mathrm{i} L_y = \hbar \mathrm{e}^{\pm \mathrm{i}\phi} \left[ \pm \frac{\partial}{\partial \theta} + \mathrm{i} \cot\theta \frac{\partial}{\partial \phi} \right]. \tag{8.65}$$

Now,

$$L_+ L_- = (L_x + \mathrm{i} L_y)(L_x - \mathrm{i} L_y) = L_x^2 + L_y^2 + \hbar L_z. \tag{8.66}$$

Thus, we can write the total angular momentum as

$$\begin{aligned} L^2 &= (L_x + \mathrm{i} L_y)(L_x - \mathrm{i} L_y) + L_z^2 - \hbar L_z \\ &= -\hbar^2 \left[ \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left( \sin\theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \phi^2} \right] \end{aligned} \tag{8.67}$$

which is the angular part of the differential equation for the total wave function! We need only show now that the eigenvalue is indeed given by $\ell(\ell + 1)$. For this, we will suppress the factors of $\hbar$.

For the first step, we note that we can rewrite the total angular momentum in the form

$$L^2 - L_z^2 = L_x^2 + L_y^2 \geq 0 \tag{8.68}$$

which means that

$$-\sqrt{L^2} \leq L_z \leq \sqrt{L^2}. \tag{8.69}$$

Thus, the possible limits for $m$ are set by the total angular momentum (this is why the polynomial can be made to be finite in the associated Legendre polynomials). Now, we note that

$$L_z(L_x \pm \mathrm{i} L_y) = (L_x \pm \mathrm{i} L_y)(L_z \pm 1). \tag{8.70}$$

If we now operate on an arbitrary angular wave function, the last term produces $m \pm 1$, or

$$L_z(L_x \pm \mathrm{i}L_y)\psi_m = (m \pm 1)(L_x \pm \mathrm{i}L_y)\psi_m. \tag{8.71}$$

This tells us that the operator $L_-$ reduces the angular momentum by one unit while $L_+$ raises it by one unit. If $m = \ell$, the raising operator must give a zero result, or

$$(L_x + \mathrm{i}L_y)\psi_\ell = 0. \tag{8.72}$$

If we now operate with the lowering operator $L_-$, the same result is obtained, and

$$(L_x \pm \mathrm{i}L_y)(L_x + \mathrm{i}L_y)\psi_\ell = (L^2 - L_z^2 - L_z)\psi_\ell = 0 \tag{8.73}$$

where we have used (8.67). Introducing the eigenvalues for the $z$-component of angular momentum, with $m = l$ in these results, we find

$$L^2 = \ell^2 + \ell. \tag{8.74}$$

Thus, we find, upon re-inserting the Planck's constant terms, that the eigenvalues for the total angular momentum are

$$L^2 = \ell(\ell + 1)\hbar^2 \tag{8.75}$$

as we assumed in section 8.2.1 to solve the radial equation.

## 8.3 Atomic energy levels

In treating the hydrogen atom, the central potential was taken to be the Coulomb potential that existed between the atomic core and the single electron. It would be desirable to continue to do this for the general atom, but there are problems. Certainly, we can continue to use the central-field approximation. However, the simple Coulomb potential of (8.43) is only valid for the outermost electron when the core is shielded by the remaining electrons. For the innermost level, the charge is $Ze$, rather than $e$, where $Z$ is the atomic number of the atom. This means that the potential is not simply a $1/r$ potential, but rather one whose amplitude varies with the electron under consideration. The simple energy levels obtained in (8.53) are no longer valid, in that the angular momentum states are no longer degenerate with the states with no angular momentum—in essence, the energy levels now depend upon both $n$ and $\ell$. This change can be found by treating the difference between the actual potential (for a given electron) and the Coulomb potential as a perturbation, and calculating the shift in energy for various values of $n$ and $\ell$. In fact, this is not particularly accurate, because the inner-shell electrons also interact with the outer electrons (and, in fact, each electron interacts with each other electron) through an additional Coulomb (repulsive) potential. Many schemes have been proposed for calculating the exact eigenvalues through various approximations, but usually some form of variational approach yields the best method. These approaches are, however, well beyond the level we want to treat here. The most important result is that, for a given level index $n$, the states of lowest $\ell$ lie at a lower energy (are more tightly bound to the nucleus). This is sufficient information to begin to construct the periodic table of the elements.

The lowest, and only exact, energy level is that for hydrogen, where there is a single electron and the atomic core of unit charge. The lowest energy level from (8.53) is the one for $n = 1$. The allowed values of radial and angular momentum indices must satisfy $n' + \ell = 0$ from (8.52). Thus, the only occupied state has no angular momentum, and the radial wave function varies (in the asymptotic limit of a simple Coulomb potential) simply as $\mathrm{e}^{-\rho/2}$, in reduced units. Because of electron spin, this energy level can hold two electrons, so this behaviour actually holds also for helium (although the energy level is shifted to lower energy due to the non-Coulomb nature of the potential in helium). Because these wave functions are spherically symmetric, they have come to be termed the 1s levels. The 1s $(n = 1, \ell = 0)$ level can hold two electrons, which then fill the complete shell for $n = 1$. This *shell* comprises the first row of the periodic table.

The next energy level, according to (8.53) has $n = 2$. For this level, (8.52) would tell us that $n' + \ell = 1$, so the lowest angular momentum state has $(n', \ell) = (1, 0)$. This lowest state gives rise to the pair of electrons in the 2s state. For this state, the wave function is again spherically symmetric and has an asymptotic variation (for a simple Coulomb potential) proportional to $(\rho - 2)\mathrm{e}^{-\rho/2}$. This wave function has two nodes (in the amplitude-squared value), and this behaviour continues as the order of the s levels increases. The two elements that are added via these two energy states are Li and Be. For more electrons, the angular momentum states begin to be filled. These levels arise for $(n', \ell) = (0, 1)$. There are $2\ell + 1 = 3$ values of $m$ for these levels, and each will hold two electrons because of the electron spin angular momentum. This means that there can be an additional six electrons accommodated by these levels. From the discussion of section 8.2.2, we know that the angular part of the wave function for $m = 0$ varies as $\cos\theta$ only. This gives amplitude peaks along the $z$-axis, so this state is the $\mathrm{p}_z$ state. Similarly, the $m = \pm 1$ states give rise to the $L_\pm$ states, which are combinations of the $\mathrm{p}_x$ and $\mathrm{p}_y$ states. These states have an angular variation as $\sin\theta \mathrm{e}^{\pm \mathrm{i}\phi}$. In figure 8.7, we indicate the orientation of these three angular momentum wave functions. The filling of these six states, which are termed the 2p states (we use the symbol p to signify the $\ell = 1$ states), is carried out by progressing through the elements B, C, N, O, F, and Ne ($Z = 5$–$10$). Thus, the $n = 2$ level can hold eight electrons, and this shell is completely filled by the element of atomic number ten (Ne). The elements Li through to Ne comprise row two of the periodic table.

The states for $n = 3$ arise in a similar manner. Here, $n' + \ell = 2$, so $\ell$ can take the values 0, 1, and 2. These give rise to the two 3s states and the six 3p states, which do not differ in principle from those described above. However, the ten possible states for $\ell = 2$ raise additional complications, and are described by quite complicated wave functions. These are the d levels, as we use the symbol

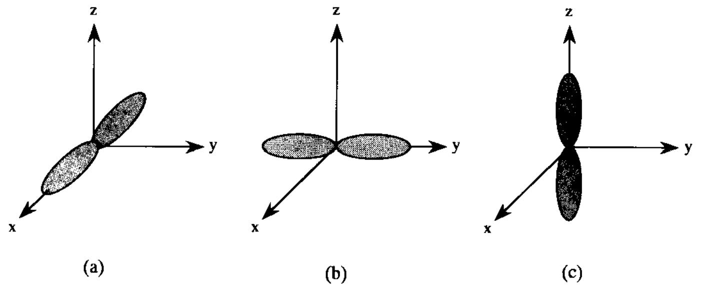


**Figure 8.7.** The orientation of the three p states for the 2p wave functions. The general convention is that the wave function is positive along the positive direction of the axis and negative along the opposite direction.

d to describe the $\ell = 2$ levels. First, however, the third row of the periodic table is formed by filling the two 3s states and the six 3p states, in that order, by the elements Na, Mg, Al, Si, P, S, Cl, and Group A (elements number 11 through to 18). The 3d levels lie rather high in energy, and actually *lie above* the 4s levels! Thus, the next two elements, K and Ca, actually have their electrons in the 4s levels, with the 3d levels completely empty. This is where the complications begin, as the 3d levels are now filled by the next elements, and the ten elements required to fill these levels are the first set of *transition metals*. In the periodic table, these are the elements that fill the Group B columns, as the original series is designated the Group A columns. The transition metals formed from the 3d levels are known as the Pb series, those formed by the 4d levels are known as the Pd series, and so on.

After the first of the transition metals rows is completed, the 4p levels are filled by Ga, Ge, As, Se, Br, and Kr. Thus, in going across row 4 of the periodic table, there is a difference between columns 2A and 3A, in that the inner d levels are filled for the latter column and are empty for the former column, and it is between these two columns that the entire Group B set of columns is inserted. This same behaviour continues in row 5, where the second transition metal series is encountered. In the outer shells, further complications arise from the filling of the $\ell = 3$ (designated the f levels) states, and this gives two new series of elements of elements. The first of these, the rare-earth series, involves elements 58–71, and fills their f levels in such a manner that they fit into the periodic table between the first two transition metals in the 5d shells. The second series of f-shell elements follows Ac, and is known as the actinides. In table 8.1, the energy levels of the s and p levels are given for a variety of atoms that find use in semiconductor technology (Herman and Skillman 1963).

**Table 8.1.** Outermost atomic energy levels for selected atoms.

| Atom | $n$ | $-\mathcal{E}_\mathrm{s}$ | $-\mathcal{E}_\mathrm{p}$ | Atom | $n$ | $-\mathcal{E}_\mathrm{s}$ | $-\mathcal{E}_\mathrm{p}$ |
|---|---|---|---|---|---|---|---|
| H | 1 | 13.6 | | Ga | 4 | 11.37 | 4.9 |
| B | 2 | 12.54 | 6.64 | Ge | 4 | 14.38 | 6.36 |
| C | 2 | 17.52 | 8.97 | As | 4 | 17.33 | 7.91 |
| N | 2 | 23.04 | 11.47 | Se | 4 | 20.32 | 9.53 |
| O | 2 | 29.14 | 14.13 | Cd | 5 | 7.7 | 3.38 |
| Al | 3 | 10.11 | 4.86 | In | 5 | 10.12 | 4.69 |
| Si | 3 | 13.55 | 6.52 | Sn | 5 | 12.5 | 5.94 |
| P | 3 | 17.1 | 8.32 | Sb | 5 | 14.8 | 7.24 |
| S | 3 | 20.8 | 10.27 | Te | 5 | 17.11 | 8.59 |
| Zn | 4 | 8.4 | 3.38 | Hg | 6 | 7.68 | 3.48 |

### 8.3.1 The Fermi–Thomas model

We now turn our attention to the first of the problems listed above that accompany the central-field approximation—the deviation from a simple Coulomb potential. When the atom is completely ionized, the potential provided by the atomic core is indeed a Coulomb potential, but with a total charge $Ze$, where $Z$ is the atomic number and, hence, the number of electrons. As we add electrons to the core, the strength of the potential is reduced to a level below $Ze$; in essence the added inner-shell electrons *screen* the potential seen by the outer-shell electrons. In the Fermi–Thomas model (Fermi 1928, Thomas 1927), a statistical model is used in which it is assumed that the potential varies slowly over a radial distance in which several electrons can be localized; that is, the potential varies slowly over the radial thickness of a particular fraction of an energy shell occupied by, for example, the 3s electrons. It may then be assumed that the electrons will obey Fermi–Dirac statistics. Here, we follow the treatment of Schiff (1955).

The approach is to compute the density of states function, and then to use this to compute the number of electrons that can occupy a volume of $\boldsymbol{k}$ space with radius less than $k$ (here, we assume that $\boldsymbol{p} = \hbar\boldsymbol{k}$). It is assumed further that the kinetic energy of the carriers within this volume will be more or less equal to the potential energy at the outer radius of the volume, and this allows us to connect a density to the actual potential around the atom. This density is then used in Poisson's equation to find the actual radial potential variation. To begin, we note that the density of states in $\boldsymbol{k}$ space is simply

$$2\left(\frac{L}{2\pi}\right)^3 \tag{8.76}$$

where the factor of 2 is for spin degeneracy, and the number of electrons lying within a volume of $\boldsymbol{k}$ space with radius $k$ is just

$$N(k) = 2\left(\frac{L}{2\pi}\right)^3 \int_0^k k^2\,\mathrm{d}k \int_0^\pi \sin\theta\,\mathrm{d}\theta \int_0^{2\pi} \mathrm{d}\phi = \frac{k^3L^3}{3\pi^2}. \tag{8.77}$$

The electron density is just $N/L^3$, and the kinetic energy is $\hbar^2k^2/2m$, so we can write the electron density as (recall that we are going to let the kinetic energy be equal to the potential energy by assuming that the potential energy describes the outer limit of the electron orbit as in a classical motion, and we note that the potential is negative)

$$n(r) = -\frac{(2mV(r))^{3/2}}{3\pi^2\hbar^3}. \tag{8.78}$$

We can now write Poisson's equation (which is for the voltage and not the energy, so an additional $e$ is introduced) as

$$\nabla^2 V(r) = \frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}V}{\mathrm{d}r}\right) = -\frac{e^2n(r)}{4\pi\varepsilon}. \tag{8.79}$$

The boundary conditions on the potential must be set by the physical constraints of the system. As $r \to 0$, there is no charge due to the electrons, and the only charge is that on the atomic core. Thus, $V(r \to 0) \to -Ze/4\pi\varepsilon r$. On the other hand, as $r \to \infty$, there is no net charge within the sphere that we are considering. In this limit the potential must fall off faster than $1/r$, so $rV(r) \to 0$. In this sense, the potential found from the Fermi–Thomas approach is that experienced by a test charge, and is not the one experienced by the electrons themselves. In this view, the potential is screened by the charge, so the actual potential in this large-radius limit goes as follows:

$$V(r) \sim -\frac{1}{r}\mathrm{e}^{-\lambda r}. \tag{8.80}$$

If we now combine the above equations, it is possible to compute an estimate for the radial dependence of the actual potential around the atom. Combining (8.79) and (8.78) leads to

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}V}{\mathrm{d}r}\right) = \frac{e^2(2mV(r))^{3/2}}{3\pi^3\varepsilon\hbar^3}. \tag{8.81}$$

The potential, along with the boundary conditions discussed above, can be conveniently incorporated into a set of dimensionless constants. To achieve this, we define

$$\begin{aligned} V(r) &= \frac{Ze^2}{4\pi\varepsilon r}\chi \qquad r = bx \\ b &= \frac{1}{2}\left(\frac{3\pi}{4}\right)^{3/2}\frac{4\pi\varepsilon\hbar^2}{me^2Z^{1/3}}. \end{aligned} \tag{8.82}$$

With these substitutions, (8.81) becomes

$$x^{1/2}\frac{\mathrm{d}^2\chi}{\mathrm{d}x^2} = \chi^{3/2} \tag{8.83}$$

with

$$\chi = 1 \quad \text{at} \quad x = 0 \quad \text{and} \quad \chi = 0 \quad \text{as} \quad x \to \infty. \tag{8.84}$$

This is still a complicated equation, and the solution is usually obtained numerically. One result of this is that the atomic radius of an atom is inversely proportional to the cube root of its atomic number $Z$. While this is not particularly accurate for small $Z$, the accuracy of the Fermi–Thomas approximation increases as the atomic number increases. If for no other reason, this is true since the increasing number of electrons makes the statistical method more accurate.

### 8.3.2 The Hartree self-consistent potential

A second method for computing the potential for the atom is due to Hartree (1928). In this method, it is assumed that each electron moves in a central field that can be calculated from the nuclear potential and the wave functions of all of the other electrons, where the charge density of each electron is given by the squared magnitude of its wave function. The Schrödinger equation is solved for each electron moving in its own central field in a self-consistent fashion. We will see in the next chapter that the Hartree potential is created by calculating the Coulomb interaction between the chosen electron and all other electrons, but that the coordinates of the other electrons are integrated out of the problem, thus providing an average potential seen by the electron due to the average distance away of the other electrons. Since the position of each electron changes as the other electrons move, iteration to self-consistency is required. In essence, we want to solve the $N$ equations for the $N$ electrons ($i = 1, 2, \ldots, N$)

$$\left[-\frac{\hbar^2}{2m}\nabla_i^2 - \frac{Ze^2}{4\pi\varepsilon r} + \sum_{j\neq i}\int |\psi(\boldsymbol{r}_j)|^2 \frac{e^2}{4\pi\varepsilon|\boldsymbol{r}_j - \boldsymbol{r}_i|}\,\mathrm{d}^3\boldsymbol{r}_j\right]\psi(\boldsymbol{r}_i) = \mathcal{E}_i\psi(\boldsymbol{r}_i). \tag{8.85}$$

To begin, a potential that approximates that represented by the second and third terms in the square brackets is assumed and the charge distributions calculated. These are then inserted into Poisson's equation to calculate the next iteration for the potential. This is then used to calculate a new charge distribution and the process is repeated until consistency is achieved. In a sense, this is an iterative approach to the variational method introduced in section 6.3. There is one more assumption that has been suppressed until now, and this is that the angular effects in the third term are averaged to give a spherically symmetric potential, and it is often assumed that the electrons in a given shell (all s, p, and so on) all move in the same potential.

A further modification was introduced to include the exchange interaction as a modification to the third term within the square brackets, a form that is called the Hartree–Fock approximation. We will discuss the exchange energy term in the next chapter as well. This gives a more complete potential variation, and ensures that the electrons satisfy the Pauli exclusion principle. The energy levels listed in table 8.1 were calculated within the Hartree–Fock approximation.

### 8.3.3 Corrections to the centrally symmetric potential

In both the Fermi–Thomas approximation and the Hartree self-consistent potential, one must still deal with the difference between the actual potential and the average potential that has been used to solve for the energy levels. Generally, this is small and does not make a big correction to the results. However, the interaction between the orbital angular momentum and the spin angular momentum of the electron must also be included, and this effect leads to an observable fine-structure splitting of the energy levels. If we take $\boldsymbol{L}_i$ as the angular momentum of the $i$th electron in its orbit and $\boldsymbol{S}_i$ as its spin angular momentum about its own axis, then the spin–orbit interaction can be written in terms of an additional energy

$$\sum_i \xi(r)\boldsymbol{L}_i \cdot \boldsymbol{S}_i \tag{8.86}$$

where $\xi(r)$ is a function that depends upon the radial variation of the potential through

$$\xi(r) = \frac{1}{2mc^2}\frac{1}{r}\frac{\mathrm{d}V}{\mathrm{d}r}. \tag{8.87}$$

The presence of the factor $c^2$ clearly signals that this correction is a relativistic one. This is why this term really does not appear in the Hamiltonians that have been treated until now. The semi-classical treatment that has been used to gather the appropriate terms for the Hamiltonian is a non-relativistic one, and this term must be put in if it is to be a factor. That is, we add the term in an *ad hoc* manner, knowing that to derive it requires a complicated approach well beyond the interest of the present treatment. Nevertheless, the spin–orbit coupling produces a splitting of those levels in which there exists an orbital angular momentum; that is, it is important in the d, f, ... levels, where $\ell \neq 0$. It is not too difficult to recognize that the contribution of the terms in (8.86) that arise from full shells of electrons vanishes through the summation over completely opposite angular momentum eigenvalues. The summation can therefore be limited to being only over those electrons that are in incomplete outer shells.

#### *8.3.3.1 LS-coupling*

In a given energy level, say a p level for example, there are a number of states that are degenerate in that they have the same energy (six of them for the example used here). These states differ in their values of $m_L$ $(-1, 0, 1)$ and $m_S$ $(\pm\frac{1}{2})$. The theory of the actual spectra consists in introducing an interaction, such as (8.86), that can be used to diagonalize the portion of the Hamiltonian dealing with these (six) degenerate states, which will give a contribution to the diagonal energies that is different (in principle) for each of these states. This splits the degenerate energy level into a set of (six) states.

The case usually treated is one for which the electrostatic energy adjustment discussed above (either the Fermi–Thomas one or the Hartree one) is larger than the spin–orbit interaction, which is termed the *Russell–Saunders* (Russell and Saunders 1925) perturbation scheme. Thus, the spin–orbit interaction is a small effect, and we can say that the states of the total Hamiltonian should be eigenfunctions of any dynamical variables that commute with the Hamiltonian. Then, we seek to put the spin–orbit interaction into a form in which we can utilize these properties. Once these states are considered, with the total Hamiltonian incorporating all of the corrections, the only constants of the motion (for the levels that are of interest here, to raise the degeneracy) are the parity (even or odd, etc) and the total angular momentum. Now, we define the total angular momentum of the electrons as

$$\boldsymbol{J} = \boldsymbol{L} + \boldsymbol{S} = \sum_i (\boldsymbol{L}_i + \boldsymbol{S}_i). \tag{8.88}$$

This total angular momentum is conserved, because the relative angles of any motion with an arbitrary axis are not observable. If one were to neglect the spin–orbit interaction, and treat only the electrostatic corrections, then $\boldsymbol{L}$ and $\boldsymbol{S}$ would be separate constants of the motion, as it is the spin–orbit interaction that produces the coupling.

As previously, the total angular momentum and the $z$-component of the angular moment, both for the orbital component and for the spin component, may be specified. This leads to the expressions

$$\begin{array}{ll} |\boldsymbol{J}|^2 = J(J+1)\hbar^2 & J_z = m_J\hbar \\ |\boldsymbol{L}|^2 = L(L+1)\hbar^2 & L_z = m_L\hbar \\ |\boldsymbol{S}|^2 = S(S+1)\hbar^2 & S_z = m_S\hbar. \end{array} \tag{8.89}$$

When the spin–orbit interaction is neglected, the electrostatic potential will separate states of different $L$ ($= \ell$ for each electron). In some cases, only particular values of $S$ are permitted; for example, if $\ell = 0$ as in the s levels, $S = 0$ as it sums over two electrons of opposite spin. Because of the spherical symmetry of the Hamiltonian, the energy is independent of the $z$-components of the angular momentum, and there are $(2\ell + 1)(2s + 1)$ (=6 for our p levels, as $s = \frac{1}{2}$) degenerate states. The degenerate states can be composed of linear combinations of the individual states specified by the pair $(m_L, m_S)$ of indices. This approach is called the $LS$-coupling scheme since the individual orbital angular momenta are grouped to form the total $L$, and the individual spin angular momenta are grouped to form the total $S$.

When the spin–orbit interaction is now included in order to raise the degeneracy, the individual $L$ and $S$ are no longer good quantum numbers, but

**Table 8.2.** Allowed values for $L$, $S$ for the p level.

| State | $L$ | $S$ |
|---|---|---|
| $\mathrm{p}^1$ | 1 | $\frac{1}{2}, -\frac{1}{2}$ |
| $\mathrm{p}^2$ | 2 | $1, 0$ |
| $\mathrm{p}^3$ | 3 | $\frac{3}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}$ |
| $\mathrm{p}^4$ | 4 | $1, 0, -1$ |
| $\mathrm{p}^5$ | 5 | $\frac{1}{2}, -\frac{1}{2}$ |

the total $J$ and $m_J$ are. It must be assumed that the levels of different $L$ are sufficiently far apart (i.e., the 2p and 2s are far apart) that degenerate perturbation theory can be applied only to the degenerate set for each level. Now, let us consider in more detail the p levels. A full p level has $L = 0$ and $S = 0$. If we write the number $n$ of electrons in the p level as $\mathrm{p}^n$, then the various combinations are given in table 8.2. The entries for $S$ are understandable if we realize that only three electrons may possess the same spin.

The various states in the p levels are named according to a convention that specifies a Russell–Saunders state. We write this as ${}^w\mathrm{X}_J$, where $w$ is the multiplicity $(2|S| + 1)$, X is S, P, D, F, G, H, ... according to $L = 0, 1, 2, 3, 4, 5, \ldots$, respectively, and the subscript is the total-angular-momentum quantum number. The notation for $L$ follows that for $l$, except that capital letters are used instead of the lower-case ones. Thus, for example, the three states for $\mathrm{p}^4$ are ${}^3\mathrm{G}_3$, ${}^3\mathrm{G}_5$, ${}^1\mathrm{G}_4$. The first two are triplets (three degenerate levels) while the last is a singlet (one level). Similarly, the three states for $\mathrm{p}^2$ are ${}^3\mathrm{D}_3$, ${}^3\mathrm{D}_1$, ${}^1\mathrm{D}_2$. The three possible states are that $J$ can take on the values $L + S$, $|L - S|$. If we take the maximum value of $S$ for each row, then $J$ can take values that lie between these two limits and differ by integers. These various notations are important in atomic spectroscopy, where the absorption (or emission) spectra of atoms are studied.

### *8.3.3.2 JJ-coupling*

In heavy atoms, it is often the case that the spin–orbit coupling is quite strong, and may even be stronger than the actual atomic potential. The orbital and spin angular momenta of all the various states are coupled together to determine the states. Here, the problem is solved for the angular momentum states and the electrostatic potential is used to split states of the same $J$ from one another. The technique is mainly of interest for the very heavy atoms, and is mentioned here only for completeness, as it is a very different approach to that of solving for the atomic structure.

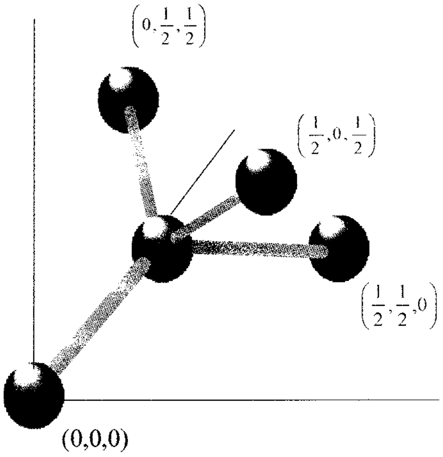

$\left(0,\frac{1}{2},\frac{1}{2}\right)$
$\left(\frac{1}{2},0,\frac{1}{2}\right)$
$\left(\frac{1}{2},\frac{1}{2},0\right)$
(0,0,0)

**Figure 8.8.** The bonding configuration of the tetrahedral coordination in the diamond and zinc-blende lattice common to most semiconductors. The inner atom of the two on the basis located at $(0, 0, 0)$ is bonded to its four nearest neighbours (on the adjacent faces of the cube) by highly directional *orbitals*. These sp$^3$ orbitals give highly directional bonds.

### 8.3.4 The covalent bond in semiconductors

The basic structure of the energy bands can be inferred from a knowledge of the atomic lattice and its periodicity. The semiconductors in which we are interested are *tetrahedrally* coordinated, by which we mean that there are four electrons (on average) from each of two atoms in the basis of the diamond structure. (This is the face-centred cubic structure with a basis of two atoms per lattice site.) These four electrons are in the s and p levels of the outer shell. For example, the Si bonds are composed of 3s and 3p levels, while GaAs has bonds composed of 4s and 4p levels, as does germanium. In each case, the inner shells are not expected to contribute anything at all to the bonding of the solid. This is not strictly true, as the inner d levels often lie quite close to the outer s and p levels when the former are occupied. This would imply that there is some modification of the energy levels in GaAs due to the filled 3d levels. This correction is small, and will not be considered further. The p wave functions are quite directional in nature, and this leads to a very directional nature of the bonding electrons. Thus, the electrons are not diffusely spread, as in a metal, but are quite localized into a set of hybrids which join nearest-neighbour atoms together. This hybrid orbital bonding is shown in figure 8.8.

The bonds are composed of *hybrids* that are formed by composition of the various possible arrangements of the s and p wave functions. There are, of course, four hybrids for the tetrahedrally coordinated semiconductors. These four hybrids may be written as

$$\begin{aligned}|h_1\rangle &= \tfrac{1}{2}[|s\rangle + |p_x\rangle + |p_y\rangle + |p_z\rangle]\\ |h_2\rangle &= \tfrac{1}{2}[|s\rangle + |p_x\rangle - |p_y\rangle - |p_z\rangle]\\ |h_3\rangle &= \tfrac{1}{2}[|s\rangle - |p_x\rangle + |p_y\rangle - |p_z\rangle]\\ |h_1\rangle &= \tfrac{1}{2}[|s\rangle - |p_x\rangle - |p_y\rangle + |p_z\rangle].\end{aligned} \tag{8.90}$$

The first of these hybrids points in the (111) direction, while the other three point in the $(1\bar{1}\bar{1})$, $(\bar{1}1\bar{1})$, and $(\bar{1}\bar{1}1)$ directions, respectively (the bar over the top indicates a negative coefficient). The factor of $\frac{1}{2}$ is included to normalize the hybrids properly so that $\langle h_i/h_j\rangle = \delta_{ij}$. These hybrids are now directional and point in the proper directions for the tetrahedral bonding coordination of these semiconductors. Thus the bonds are directed at the nearest neighbours in the lattice. For Ga in GaAs, the four hybrids point directly at the four As neighbours, which lie at the points of the tetrahedron. The locations of the various atoms for the tetrahedral bond are shown in figure 8.8.

Each of the atomic levels possesses a distinct energy level that describes the atomic energy in the isolated atom, and these were shown in table 8.1. Thus the s levels possess an energy given by $E_\mathrm{s}$ and the p levels have the energy $E_\mathrm{p}$. In general, these levels are properties of the atoms, so that the levels are different in the heteropolar compounds like GaAs. The levels will be marked with a superscript A or B, corresponding to the A–B compound that forms the basis of the lattice. In the following, the compound semiconductors will be treated, as they form a more general case, and the single component semiconductors, such as Si or Ge, are a special case that is easily obtained in a limiting process of setting $\mathrm{A} \equiv \mathrm{B}$.

The s and p energy levels are separated by an energy that has been termed the *metallic* energy (Harrison 1980). In general, this energy may be defined from the basic atomic energy levels through

$$4V_1^\mathrm{A} = E_\mathrm{p}^\mathrm{A} - E_\mathrm{s}^\mathrm{A} \qquad 4V_1^\mathrm{B} = E_\mathrm{p}^\mathrm{B} - E_\mathrm{s}^\mathrm{B}. \tag{8.91}$$

Here the A atom is the cation and the B atom is the anion in chemical terms. The hybrids themselves possess an energy that arises from the nature of the way in which they are formed. Thus the *hybrid* energy is

$$\begin{aligned}E_\mathrm{h} &= \frac{\langle h_i|H|h_i\rangle}{\langle h_i|h_i\rangle} = \frac{1}{4}\frac{\langle s|H|s\rangle + \langle p_x|H|p_x\rangle + \langle p_y|H|p_y\rangle + \langle p_z|H|p_z\rangle}{1}\\ &= \tfrac{1}{4}(E_\mathrm{s} + 3E_\mathrm{p})\end{aligned} \tag{8.92}$$

for all values of the index $i$, where $H$ is the Hamiltonian operator for the Schrödinger equation representing the crystal but neglecting any interaction between the atoms (without these interaction terms, the cross terms that would arise in (8.92) will vanish by orthonormality). In figure 8.9, the hybrid energy is derived from the atomic energies in a compound semiconductor.

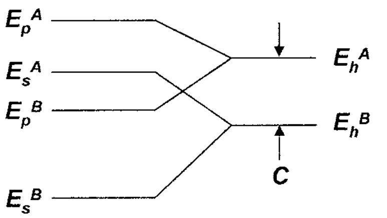


**Figure 8.9.** The atomic energies form hybrids, as indicated here, for each of the two atoms in the basis. These hybrids are separated by the *heteropolar* energy $C$.

The two hybrid energies are separated by the *hybrid polar energy* or the *heteropolar energy*, depending on whose definitions one wants to use. The notation $C$ has been used here for this energy. The heteropolar energy is a product of the ionic transfer of charge in the compound semiconductor (since Ga has only three electrons and As has five electrons, there is a charge transfer in order to get the average four electrons of the tetrahedral bond). Of course, this energy vanishes in a pure single compound such as Si or Ge, which are referred to as homopolar materials. That is, they are composed of a homogeneous set of atoms, while the general compound semiconductor is heterogeneous in that it contains two types of atom. The heteropolar energy $C$ may be easily evaluated using (8.92) as follows:

$$C = E_{\rm h}^{\rm A} - E_{\rm h}^{\rm B} = \tfrac{1}{4}(E_{\rm s}^{\rm A} - E_{\rm s}^{\rm B}) + \tfrac{3}{4}(E_{\rm p}^{\rm A} - E_{\rm p}^{\rm B}). \tag{8.93}$$

It is important to note that the hybrids are not eigenstates of either the isolated atom or of the crystal. Rather, they are constructed under the premise that they are the natural wave function for the tetrahedral bonds. But, we have created them as what seems a natural form (at least to us). In principle, they will be stable in the crystal once the interactions between the various atoms are included. In fact, they are not orthogonal under action of the Hamiltonian, since

$$\langle h_i|H|h_j\rangle = \tfrac{1}{4}(E_{\rm s} - E_{\rm p}) = -V_1 \tag{8.94}$$

so that the metallic energy measures the interaction between the various hybrids. In this sense, the metallic energy describes the contribution to the energy of the itinerant nature of the electrons.

Now we have to turn to the interaction between hybrids localized on neighbouring atoms. Only standing-wave interactions will be considered here $(k = 0)$ and the propagating wave-function-dependent changes are calculated by techniques such as the Kronig–Penney model discussed in section 3.7. The interaction energy between two atoms on different sites can arise from, for example, one atom's hybrid pointed in the (111) direction and the nearest neighbour in that direction, displaced $(\frac{a}{4}, \frac{a}{4}, \frac{a}{4})$, whose hybrid points in the opposite direction (there is a complete flip of the hybrid directions of the atoms as one moves along the body diagonal direction). Including the angular integration, the interaction energy between these nearest neighbour hybrids is

$$-V_2 = \frac{1}{4}\langle s^{\mathrm{A}}|H|s^{\mathrm{B}}\rangle + \frac{\sqrt{3}}{4}[\langle s^{\mathrm{A}}|H|p^{\mathrm{B}}\rangle + \langle p^{\mathrm{A}}|H|s^{\mathrm{B}}\rangle] + \frac{3}{4}\langle p^{\mathrm{A}}|H|p^{\mathrm{B}}\rangle. \quad (8.95)$$

Harrison (1980) has argued that these energies should depend only on the interatomic spacing $d$ $(=\sqrt{3}a/4)$, and that they should have the general form

$$V_2 \cong 4.37\frac{\hbar^2}{md^2}. \quad (8.96)$$

Phillips (1973) also argues that $V_2$ should be a function of the interatomic spacing, but that it should also satisfy another scaling rule. Since the atomic radii are the same for each row of the periodic table, the value of $V_2$ should be the same in AlP as in Si, the same in GaAs and ZnTe as in Ge, and so on. In other words, the value for this quantity is set by the distance between the atoms, and this really does not change as one moves across a row of the table. Thus this value is the same in heteropolar compounds as in homopolar compounds and Phillips has termed $V_2$ the homopolar energy, $E_{\mathrm{ho}} = 2V_2$ (this should not be confused with the hybrid energy $E_{\mathrm{h}}$, which differs for the two atoms). The *average energy gap* between bonding and anti-bonding orbitals is thus composed of a contribution from the homopolar energy $E_{\mathrm{ho}}$ and a contribution from the heteropolar energy $C$.

The bonding orbital will be composed of hybrids based on the atoms at each end of the bond, as suggested above. These bonding orbitals can be written as a linear combination of the two hybrids at each atom, as

$$|\psi_{\mathrm{bo}}\rangle = u_1|h^{\mathrm{A}}\rangle + u_2|h^{\mathrm{B}}\rangle \quad (8.97)$$

where the $u_i$ are coefficients to be determined. This is achieved by minimizing the expectation value of the energy determined by the Hamiltonian, which is given by

$$E = \frac{\langle\psi_{\mathrm{bo}}|H|\psi_{\mathrm{bo}}\rangle}{\langle\psi_{\mathrm{bo}}|\psi_{\mathrm{bo}}\rangle} = \frac{u_1^2E_{\mathrm{h}}^{\mathrm{A}} - 2u_1u_2V_2 + u_2^2E_{\mathrm{h}}^{\mathrm{B}}}{u_1^2 + u_2^2} \quad (8.98)$$

with respect to both $u_l$ and $u_2$ separately. This leads to two equations (from the partial derivatives)

$$2u_1E = 2u_1E_{\mathrm{h}}^{\mathrm{A}} - 2u_2V_2 \qquad 2u_2E = 2u_2E_{\mathrm{h}}^{\mathrm{B}} - 2u_1V_2. \quad (8.99)$$

Introducing the average energy $E_0 = (E_{\mathrm{h}}^{\mathrm{A}} + E_{\mathrm{h}}^{\mathrm{B}})/2$, the determinant of coefficients for (8.99) may readily be solved to give the energies

$$E = E_0 \pm \tfrac{1}{2}\sqrt{(2V_2)^2 + C^2} = E_0 \pm \tfrac{1}{2}\Delta. \quad (8.100)$$

The bonding (lower sign) and the anti-bonding (upper sign) energy levels are symmetrically spaced about the average hybrid energy of the two atoms. Of course, in a homopolar material, these two hybrid energies are the same and are

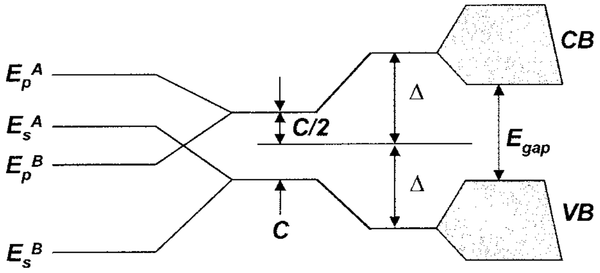


**Figure 8.10.** The hybridization that forms between adjacent atoms leads to the average energy levels (shown separated by $2\Delta$) around which the energy bands now form when the periodicity is introduced.

equal to $E_0$. The significance of the average energy is unexpected in heteropolar materials but is important for getting the positions of the average bonding and anti-bonding energies correct. The positions of these energy levels are shown in figure 8.10. Note that the square root appearing in (8.100) is denoted as the quantity $\Delta$ in the figure.

We can use the two eigenvalues in either of the two equations of (8.99) to determine the quantities $u_i$. These lead to the coefficients

$$u_1 = \sqrt{\frac{1+\alpha}{2}} \qquad u_2 = \sqrt{\frac{1-\alpha}{2}} \qquad \alpha = \frac{C}{\sqrt{C^2 + 4V_2^2}}. \tag{8.101}$$

Here, $\alpha$ is the bond polarization fraction. In homopolar materials, $C = 0$, so that the polarization is zero and the two contributions are equal, as one would expect. This polarization is a result of the charge transfer and we will see it discussed again later in connection with the lattice polar optical mode of vibration. We also note that (8.101) relates the polarization to the fraction of heteropolar nature in the bond itself. Thus one can call $\alpha$ a quantity that is related to the fraction of ionicity in the bond.

It is clear that once we know the atomic energy levels, we can immediately determine the hybrid energies, and the heteropolar and homopolar energies. Using $V_2$ as the constant value given in (8.96), the average energies are now known. However, Phillips (1973) treats the homopolar energy and the heteropolar energy $C$ as adjustable parameters to get better scaling for the average energy gaps, building his results from a need to get the dielectric functions correct. His values are close to the Harrison (1980) values obtained by straightforward application of the atomic energies, but there are differences. In the heteropolar materials, the differences are even greater, and the average gap can differ by more than a volt between the two approaches. To be sure, it is not at all clear that these two methods are comparable or that the two authors are talking about exactly the same quantities, even though it appears to be so. That the numbers are close is perhaps remarkable and points out the basic correctness of the overall picture of the composition of the energy bands in semiconductors.

The values that are obtained, as seen in figure 8.10, are not the actual energy gaps one associates with the conduction and valence bands, but are average gaps around which the energy bands form. That is, these are levels near mid-band that may be associated with the mean energy in the band. These levels are then broadened by the interaction with other atoms, and the periodicity that is invoked in the structure, and this broadening produces the conduction and valence bands, just as in the Kronig–Penney model. The number of states in each of these bands is twice the number of lattice sites for the basis of two atoms per lattice site and another factor of two for spin. Thus, the four electrons per atom are just sufficient to completely fill the valence band, and it is this full valence band, with a gap to the conduction band, that provides most of the properties of the semiconductors.

## 8.4 Hydrogenic impurities in semiconductors

One can achieve in semiconductors a situation in which an impurity atom is placed in the host lattice, with the special case that the impurity may have an additional electron or be short of one electron. When the impurity atom has an additional electron over those (four) required for tetrahedral bonding in the covalent lattice, this extra electron can be ionized rather easily. In this case, the impurity atom has a single positive charge, while the ionized electron has a single negative charge. In many ways this interaction is quite like that in the hydrogen atom. On the other hand, the impurity atom that is short of one electron allows for electrons to move from other atoms to this one in order to complete the tetrahedral bonding requirements. Here, we say that a ‘hole’ moves from the impurity to the other atoms. This hole (the absence of an electron in the valence band) can be ionized from the impurity, leaving a negatively charged impurity atom and a positively charged particle.

The energy required to ionize either the electron or the hole is generally much smaller than that required to lift an electron out of the valence band into the conduction band—the band gap energy. Thus, these impurities introduce defect levels within the band gap region of the semiconductor. Because of the coulombic nature of the interaction between the ionized particle (electron or hole) and the central cell of the ionized impurity, the perturbing potential is simply a Coulomb potential:

$$V(r) = -\frac{e^2}{4\pi\varepsilon r}. \tag{8.102}$$

Here, $\varepsilon$ is the dielectric constant (times the free-space permittivity) of the semiconductor host crystal. It may generally be assumed that the carriers in the semiconductor are characterized by a scalar *effective mass* $m^*$, which accounts for the band structure nature of the electrons or the holes, as the case may be. Thus, we are concerned either with electrons near the minimum of the conduction

**Table 8.3.** Common impurities in Si.

| | $E_\mathrm{d}$ (meV) | $E_\mathrm{a}$ (meV) |
|---|---|---|
| P | 45 | |
| As | 54 | |
| Sb | 39 | |
| B | | 45 |
| Al | | 67 |
| Ga | | 72 |
| In | | 16 |

band or holes near the maximum of the valence band. With this potential, and the effective mass, Schrödinger's equation for this reduced system becomes

$$-\frac{\hbar^2}{2m^*}\nabla^2\psi(\boldsymbol{r}) - \frac{e^2}{4\pi\varepsilon r}\psi(\boldsymbol{r}) = \mathcal{E}\psi(\boldsymbol{r}) \tag{8.103}$$

which is the same equation as for the simple hydrogen atom (with the appropriate changes in the dielectric constant and the mass). This means that the allowed energy levels for the bound electrons are simply given by the hydrogen energy levels suitably adjusted:

$$\mathcal{E}_n = -\frac{e^4 m^*}{8\varepsilon^2\hbar^2 n^2} \tag{8.104}$$

where $n$ is an integer $\geq 1$. The first ionization energy of the hydrogen atom is one Rydberg, or 13.6 eV. Here, this value is reduced by the square of the dielectric constant, which is of the order of 10, and by the effective mass, which is of the order of 0.1. Thus, the ionization energy of the impurity is of the order of 0.0136 eV. This value, of course, varies according to the specific dielectric constant and the effective mass. What is not accounted for in this hydrogenic model is variation according to just which atomic species is providing the impurity atom. In the simple model, the results should be the same for all impurities, but this is not the case. However, the hydrogenic model is a good approximation. In table 8.3, we list a few dopants that are found in silicon, and the values of their ionization energies. The donors $(E_\mathrm{d})$ generally come from group V of the periodic table and have an extra electron (over the four needed to complete the tetrahedral bonding discussed earlier). The acceptors $(E_\mathrm{a})$ generally come from group III of the periodic table, and lead to holes in the valence band.

The wave function for the first ionization state varies just like that for the hydrogen atom, but with an adjusted radius. In fact, we may write the wave function as

$$\psi(r) \sim \mathrm{e}^{-r/a} \tag{8.105}$$

where

$$a = \frac{4\pi\varepsilon\hbar^2}{e^2 m^*} = a_0 \frac{\varepsilon}{\varepsilon_0}\frac{m_0}{m^*} \tag{8.106}$$

where $a$ is the Bohr radius discussed in section 8.2.1. For our hypothetical semiconductor, with a dielectric constant of 10 and an effective mass of 0.1, we find an effective radius of $5.3 \times 10^{-7}$ cm, or 5.3 nm. This is about ten lattice constants, so the size of the orbit of the electron, when it is captured by the impurity, actually samples a great many unit cells of the crystal. This means that the effective-mass approximation is a fairly good approximation for the hydrogenic impurity. Nevertheless, the potential probably deviates in reality from a coulombic one, and this difference is likely to account for the variation in ionization energy seen for different impurities in the same semiconductor.

## References

Cohen-Tannoudji C, Diu B and Laloë F 1977 *Quantum Mechanics* vol 1 (New York: Wiley) p 727

Darwin C G 1931 *Proc. Cambridge Phil. Soc.* **27** 86–90

Fermi E 1928 *Z. Phys.* **48** 73

Fock V 1928 *Z. Phys.* **47** 446

Harrison W A 1980 *Electronic Structure and the Property of Solids* (San Francisco, CA: Freeman)

Hartree D R 1928 *Proc. Cambridge Phil. Soc.* **24** 111

Herman F and Skillman S 1963 *Atomic Structure Calculations* (Englewood Cliffs, NJ: Prentice-Hall)

McEuen P L, Foxman E B, Meirav U, Kastner M A, Meir Y, Wingreen N S and Wind S J 1991 *Phys. Rev. Lett.* **66** 1926–9

Phillips J C 1973 *Bonds and Bands in Semiconductors* (New York: Academic)

Rössler U 1991 *Quantum Coherence in Mesoscopic Systems* ed B Kramer (New York: Plenum) pp 45–62

Russell H N and Saunders F A 1925 *Astrophys. J.* **61** 38

Schiff L I 1955 *Quantum Mechanics* 2nd edn (New York: McGraw-Hill) pp 281–3

Sikorski Ch and Merkt U 1989 *Phys. Rev. Lett.* **62** 2164–6

Tarucha S, Austing D G, Honda T, van der Hage R J and Kouwenhoven L P 1996 *Phys. Rev. Lett.* **77** 3613

Tarucha S, Austing D G, Honda T, van der Hage R J and Kouwenhoven L P 1997 *Japan. J. Appl. Phys.* B **36** 3917

Thomas L H 1927 *Proc. Cambridge Phil. Soc.* **23** 542

Yacoby A, Heiblum M, Umansky V, Shtrikman H and Mahalu D 1994 *Phys. Rev. Lett.* **73** 3149

Yacoby A, Heiblum M, Mahalu D and Shtrikman H 1995 *Phys. Rev. Lett.* **74** 4047

## Problems

1. Find the wave functions for the $n = 2$, $m = 0$ level of the two-dimensional harmonic oscillator. Use the angular momentum operators and show that this level is cylindrically symmetric.

2. Suppose that the $x$-axis harmonic oscillator is characterized by a frequency $\omega_1$, which is slightly larger than that of the $y$-axis ($\omega_1 \simeq \omega + \delta\omega$). Compute the exact energy levels of the two-dimensional harmonic oscillator. Then, using the energy levels and wave functions for the symmetric two-dimensional harmonic oscillator, calculate the perturbation shift of these energy levels with the perturbation $m(\delta\omega)^2x^2/2$.

3. Let us consider the application of an electric field $-E\boldsymbol{a}_x$ to the two-dimensional harmonic oscillator in the presence of a magnetic field. Determine the change in the energy levels introduced by this electric field.

4. What are the energy values of the lowest three energy levels in the hydrogen atom? (That of the lowest energy level, the ionization energy of the hydrogen atom, is termed one rydberg.)

5. Verify that (8.58) is the correct equation for the angular variation of the hydrogen atom.

6. Show that the solution of (8.58) is a polynomial of finite order, and that this order is given by $\ell - |m|$.

7. Develop the full wave functions (the form, neglect normalization) for the 3p states utilizing the angular momentum operators.

8. Show that $L^2$ in (8.62) commutes with each component of the angular momentum.

9. Develop a computer program and solve (8.83) for the potential of an atom in the Fermi–Thomas approximation.

10. (*a*) Using the known lattice constants and atomic energy levels, compute the values of the parameters $C$, $V_2$, and $E_{\mathrm{G}}$ for Si, Ge, GaAs, AlAs, InAs, InSb, and InP. (*b*) The bond polarizability is often related to the ionicity of a particular compound. Using the values found in part (*a*), plot the bond polarizability as a function of the average energy gap for these compounds. Is there a trend to these data?

11. From the known lattice constant and density of Si, compute the number of atoms per cubic metre. Then, use this value of $N$, and the known dielectric constant of Si, to compute the valence plasma frequency from $\omega_{\mathrm{P}}^2 = Ne^2/m_0\varepsilon_{\mathrm{s}}$. What is the average energy gap required to satisfy the Penn dielectric function

$$\varepsilon_{\mathrm{s}} = \varepsilon_0 \left[1 + \left(\frac{\hbar\omega_{\mathrm{P}}}{E_{\mathrm{G}}}\right)^2\right]?$$

How does this compare with the value of $E_{\mathrm{G}}$ found in problem 10?

# Chapter 9

# Electrons and anti-symmetry

The treatment of particles that we have presented in the previous chapters is deficient in a number of respects, mainly because of the many approximations that have been made in the various approaches. One of the most significant of these is constituted by the particular constraints that are introduced by the fact that the electrons are *indistinguishable and identical* particles. In addition, we have only paid lip service to the fact that each electron also possesses a *spin angular momentum* about its own axis, although this became important at the end of the last chapter. When we deal with single isolated electrons, these approximations do not get us into very much trouble. However, most atoms and solids are densely populated with electrons, and the neglected properties can introduce new effects that need to be considered. Even though the effects may be quite small, an understanding of them is necessary, if for no other reason than to be able to ascertain when they may properly be ignored.

In classical mechanics, it is quite easy to follow the individual trajectories of each and every electron. However, the Heisenberg uncertainty principle prevents this from being possible in quantum mechanics. If we completely understand the position of a given particle, we can say nothing about its momentum and most other dynamical variables. This principle of the indistinguishability of identical particles leads to other more complex forms of the wave functions. These more complicated forms can lead to effects in quantum mechanics that have no analogue in classical mechanics. We use the word *identical* to indicate particles that can be freely interchanged with one another with no change in the physical system. While these particles may be distinguished from one another in situations in which their individual wave functions do not overlap, the more usual case is where they have overlapping wave functions because of the high particle density in the system. In this latter case, it is necessary to invoke a *many-particle wave function*. To describe properly the properties of the many-particle wave function, it is necessary to understand the properties of the interactions that occur among and between these single particles. If the one-electron problem were our only interest, this more powerful development that is used for the interacting system would be worthless to us, for its complexity is not worth the extra effort. However, when we move to multi-electron problems, the power of the approach becomes apparent. In this chapter, we want to examine just the properties of these general wave functions for the case of electrons, which also possess their own spin angular momentum, and to introduce the interaction among the electrons.

## 9.1 Symmetric and anti-symmetric wave functions

We begin by considering just two electrons. Because these are indistinguishable particles, the physical state that is obtained by merely interchanging the positions of these two particles must be completely equivalent to the original one. This puts certain constraints upon the wave function. Let us consider the two-electron wave function $\Psi(\xi_1, \xi_2)$. Here, the parameter $\xi_1$ refers to the vector position $\boldsymbol{r}_1$ and the spin orientation $\boldsymbol{\sigma}_1$ of the electron (which will be further explained below). Thus, we can refer to this as a four-vector $\xi_1 = (\boldsymbol{r}_1, \boldsymbol{\sigma}_1)$. Now, under exchange of the positions and spins of the two particles, we must have

$$\Psi(\xi_1, \xi_2) = \mathrm{e}^{\mathrm{i}\phi}\Psi(\xi_2, \xi_1). \tag{9.1}$$

The phase $\phi$ is some real constant. By repeating the interchange, we arrive at

$$\Psi(\xi_1, \xi_2) = \mathrm{e}^{\mathrm{i}\phi}\Psi(\xi_2, \xi_1) = \mathrm{e}^{\mathrm{i}2\phi}\Psi(\xi_1, \xi_2). \tag{9.2}$$

But this requires $\mathrm{e}^{\mathrm{i}2\phi} = 1$, or $\mathrm{e}^{\mathrm{i}\phi} = \pm 1$. Thus, the possible forms for the wave function are

$$\Psi(\xi_1, \xi_2) = \pm\Psi(\xi_2, \xi_1). \tag{9.3}$$

The wave function that is found upon interchange of the two particles is going to be either symmetrical or anti-symmetrical. By symmetrical, we mean that it is unchanged, while anti-symmetrical implies a change of sign. Which of these two are we to choose?

The choice of either a symmetric or an anti-symmetric wave function lies in the imposition of the Pauli exclusion principle (Pauli 1925). We have constantly held that each eigenstate determined by solving the Schrödinger equation could hold two electrons, provided that they had opposite spin. In each situation, the positional wave functions for the two electrons are the same, but the spin wave functions must yield a spin eigenvalue $\boldsymbol{\sigma}$ that is oppositely *directed*. Now, this spin eigenvalue is one of the vectors included above in the description of the two-particle wave function. Thus, the electrons, which obey the Pauli exclusion principle, must have an anti-symmetric wave function. We can summarize this by saying that particles that do not obey the exclusion principle (phonons, photons, etc) are *bosons*, and are found to obey Bose–Einstein statistics. These bosons have symmetric wave functions under interchange of the particles. On the other hand, particles that obey the exclusion principle (electrons and some others, with which we will not be concerned) are *fermions*, and are found to obey Fermi–Dirac statistics. Fermions must have anti-symmetric wave functions under the interchange of particles. We will see below that the use of the anti-symmetric wave function actually ensures that the Pauli exclusion principle is obeyed.

Now, what do we really mean by bosons and fermions? In the previous paragraph, we stated that bosons do not obey the Pauli exclusion principle, and do obey the Bose–Einstein distribution. On the other hand, fermions do obey the Pauli exclusion principle—no more than two fermions, and these are of opposite spin—can be accomodated in any quantum state. Bosons are particles with integer spin, such as phonons (zero spin) and photons (integer spin given by $\pm 1$, corresponding to right- and left-circularly polarized plane waves for example). Fermions are particles with half-integer spin, such as electrons. These two distributions are given by

$$f_{\mathrm{BE}}(E) = \frac{1}{\mathrm{e}^{E/k_{\mathrm{B}}T} - 1} \tag{9.4}$$

and

$$f_{\mathrm{FD}}(E) = \frac{1}{\mathrm{e}^{(E-E_{\mathrm{F}})/k_{\mathrm{B}}T} + 1}. \tag{9.5}$$

There is a significant difference in the behaviour of these two distributions at low energy. The Bose–Einstein distribution diverges at $E \to 0$, as there is no limit to the number of bosons which can occupy the lowest energy state. On the other hand, the Fermi–Dirac distribution approaches unity in this limit, as the state is certainly occupied if it lies well below the Fermi energy $E_{\mathrm{F}}$. As pointed out, one way of achieving the Pauli exclusion principle is to ensure that the wave function for a fermion is anti-symmetric.

Now, what do we mean by anti-symmetric? The fact is that if we have a two-electron system, which satisfies the Schrödinger equation, then we need a proper two-electron wave function $\Psi(\xi_1, \xi_2)$, where the first argument refers to the first electron and the second argument refers to the second electron. By anti-symmetry, we mean that we require the wave function to have the property $\Psi(\xi_1, \xi_2) = -\Psi(\xi_2, \xi_1)$. That is, if we interchange (exchange) the two particles, the wave function is multiplied by a numerical factor of $-1$.

If we insert our many-particle wave function into the Schrödinger equation, it is found that under the circumstances discussed above, it is possible to write this equation as

$$[H(\xi_1, \xi_2) - \mathcal{E}]\Psi(\xi_1, \xi_2) = 0. \tag{9.6}$$

Here, we may assert that the Hamiltonian itself is invariant under the exchange of the particles, and the equivalence of the two physical states implies that the energy has the same invariance. Hence, the imposition of the exclusion principle appears only in the wave function (unless some special spin-dependent interaction, such as the spin–orbit interaction of the last chapter, is introduced in the Hamiltonian to distinguish one spin state from another). Whether the wave function is symmetric or anti-symmetric has no impact upon (9.6). The importance of this latter result is that, for simple Hamiltonians with no interaction among the electrons, it is usually possible to separate (9.6) into two equations, one for each particle, with the total energy being the sum of the single-particle energies. This separation is carried out in exactly the same manner as that in which separation of coordinates is done in a many-dimensional partial differential equation. As a result, it is possible to write the two-particle wave function as a product of the one-particle wave functions $\psi_1(\xi_1)$ and $\psi_2(\xi_2)$. The subscripts on the wave functions themselves (as opposed to those on the variables) refer to the particle 'number'. Carrying this out produces

$$\Psi(\xi_1, \xi_2) \propto \psi_1(\xi_1)\psi_2(\xi_2). \tag{9.7}$$

However, this wave function does not possess the proper symmetry for electrons; for example, interchanging the positions of particles '1' and '2' does not produce the necessary anti-symmetry. However, we can achieve this with a somewhat cleverer summation; that is, we use

$$\Psi(\xi_1, \xi_2) = \frac{1}{\sqrt{2}}[\psi_1(\xi_1)\psi_2(\xi_2) - \psi_2(\xi_1)\psi_1(\xi_2)]. \tag{9.8}$$

This wave function has the desired anti-symmetry under the interchange of the two electrons.

This method of forming a properly anti-symmetric many-electron wave function from the single-electron wave functions has been extended to an arbitrarily large number of electrons by Slater (1929). The resulting wave function for $N$ electrons is given by the *Slater determinant*

$$\Psi(\xi_1, \xi_2, \ldots, \xi_N) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \psi_1(\xi_1) & \psi_1(\xi_2) & \ldots & \psi_1(\xi_N) \\ \psi_2(\xi_1) & \psi_2(\xi_2) & \ldots & \psi_2(\xi_N) \\ \ldots & \ldots & \ldots & \ldots \\ \psi_N(\xi_1) & \psi_N(\xi_2) & \ldots & \psi_N(\xi_N) \end{vmatrix}. \tag{9.9}$$

Equation (9.9) has the added usefulness that it has the property that it vanishes if two of the $\psi_i$ are the same. This implies that no two electrons can have the same state (recall that we have included spin angular momentum explicitly in the variables), and thus the Pauli exclusion principle is automatically satisfied. Thus, when we can separate the Hamiltonian into single-particle parts, and separate the wave function accordingly, the anti-symmetrized product of these wave functions in a Slater determinant ensures that the Pauli exclusion principle is satisfied.

## 9.2 Spin angular momentum

The electrons that were treated in the last section can have two possible spin states, which are generally referred to as 'spin up' and 'spin down'. The fact that there are only two spin states arises from the Pauli exclusion principle, since each state (neglecting spin) can have only two electrons, which possess opposite spin properties or 'directions'. In this section, we would like to examine the properties of the spin states and ascertain the appropriate eigenvalues and wave functions of the operators. In section 8.2.3, we discussed the angular momentum of a particle (there it was the orbital angular momentum of an electron orbiting a nucleus in a centrally symmetric potential). It was clear that $L^2$ and $L_z$ could both be made to commute with the Hamiltonian, and therefore could be simultaneously diagonalized. The $z$-component, $L_z$, had eigenvalues of $m\hbar/2$, while the total angular momentum had eigenvalues of $\ell(\ell+1)\hbar^2$ (here, $m$ is an integer and should not be confused with the mass, which we will write as $m^*$). In the present case, the only angular momentum is the spin of the electron about its own axis, which we shall take as the $z$-axis for convenience (hence, 'up' and 'down' are relative to the normal $z$-axis, which is often used to convey altitude). The range of $m$ is therefore from $-\ell$ to $\ell$, and $m$ thus takes $2\ell+1$ values. Since we have only two values ('up' and 'down'), it is clear that $\ell = \frac{1}{2}$, and $m = \pm\frac{1}{2}$. The eigenvalue of the total momentum $L^2$ is then $3\hbar^2/4$.

In the matrix formulation of quantum mechanics, the total Hilbert space is spanned by a set of eigenfunctions $\phi_i(\boldsymbol{r})$. If we now want to include the spin coordinates, an additional two-component space must be created for the two components of the spin. We can take these as the two unit vectors

$$\phi\begin{pmatrix}\frac{1}{2}\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix} \qquad \phi\begin{pmatrix}-\frac{1}{2}\end{pmatrix} = \begin{pmatrix}0\\1\end{pmatrix} \tag{9.10}$$

where the first row refers to the 'up' state and the second row refers to the 'down' state. Thus, the 'up' state has the $+\frac{1}{2}$ eigenvalue (in units of the reduced Planck's constant), and the 'down' state has the $-\frac{1}{2}$ eigenvalue. Whatever matrix describes the positional variations of the wavefunction must be adjoined by the spin wave functions. This is a general result.

Because the spin angular momentum has been taken to be directed along the $z$-axis, we expect the matrix for the $z$-component of the spin operator to be diagonal, since it must simply produce the eigenvalues when operating on the above spin wave functions. Thus, we have (we will use $S$ now rather than $L$ to indicate the spin angular momentum)

$$\boldsymbol{S}_z = \tfrac{1}{2}\hbar\begin{pmatrix}1 & 0\\0 & -1\end{pmatrix} \tag{9.11}$$

so $\boldsymbol{S}_z\phi(\frac{1}{2}) = (\hbar/2)\phi(+\frac{1}{2})$ and $S_z\phi(-\frac{1}{2}) = -(\hbar/2)\phi(-\frac{1}{2})$. Similarly, the operator for the total angular momentum must produce

$$|\boldsymbol{S}|^2 = \tfrac{3}{4}\hbar^2\begin{pmatrix}1 & 0\\0 & +1\end{pmatrix}. \tag{9.12}$$

How are we to find the other components of the spin angular momentum? For this, we can use the results of (8.66) for the rotating $S_\pm$. We know that $S_-$ is simply $S_+^*$. Now, (8.66) is simply re-expressed as

$$S^2 = S_+S_- + S_z^2 - \hbar S_z. \tag{9.13}$$

Similarly,

$$S^2 = S_- S_+ + S_z^2 + \hbar S_z. \tag{9.14}$$

These can be combined, using (9.11) and (9.12), to give

$$S^2 - S_z^2 = \tfrac{1}{2}[S_+ S_- + S_- S_+] = \tfrac{1}{2}\hbar^2. \tag{9.15}$$

The individual matrices for the rotating components can easily be found, up to a normalizing constant. To begin, we recall from chapter 8 that the $S_+$-component raises the angular momentum by one unit. This must take the 'down' state into the 'up' state and produce a zero from the 'up' state. This can be satisfied by making the definition

$$S_+ = \hbar \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{9.16}$$

Similarly, the operator $S_-$ must reduce the angular momentum by one unit, and hence it must take the 'up' state into the 'down' state and produce zero from the 'down' state. Thus, we can now define

$$S_- = \hbar \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \tag{9.17}$$

and equation (9.15) is satisfied quite easily. From these, we can now use (8.65) to define the $x$- and $y$-components:

$$\boldsymbol{S}_x = \frac{1}{2}(\boldsymbol{S}_+ + \boldsymbol{S}_-) = \frac{\hbar}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{9.18$a$}$$

$$\boldsymbol{S}_y = \frac{1}{2\mathrm{i}}(\boldsymbol{S}_+ - \boldsymbol{S}_-) = \frac{\hbar}{2}\begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}. \tag{9.18$b$}$$

It is normal to define the three components in terms of Pauli spin matrices (Pauli 1927) as $\boldsymbol{S}_i = \hbar^2 \boldsymbol{\sigma}_i$. The various $\boldsymbol{\sigma}_i$ are also called spinors.

## 9.3 Systems of identical particles

In section 9.1, the multi-particle wave function was introduced, and in particular was connected with products of single-particle wave functions through the Slater determinant to ensure the anti-symmetry of the wave function. In this section, we now want to explore some of the properties of the many-electron wave functions in further detail. Quite generally, we may introduce the $N$-particle wave function $\Psi_N(\xi_1, \xi_2, \ldots, \xi_N)$, which represents the probability (when the magnitude squared is computed) of finding particles at positions and spins $\xi_1, \xi_2, \ldots, \xi_N$. This wave function must satisfy the condition

$$\langle \Psi_N | \Psi_N \rangle = \int \mathrm{d}\xi_1 \, \mathrm{d}\xi_2 \ldots \mathrm{d}\xi_N \, |\Psi_N(\xi_1, \xi_2, \ldots, \xi_N)|^2. \tag{9.19}$$

As we have defined it to this point, the Hilbert space for the $N$-particle system is simply the $N$th tensor product of the single-particle Hilbert spaces and the corresponding spin spaces. The wave function of the $N$ fermions is properly anti-symmetric under the exchange of any two particles, and therefore for a large number of permutations must satisfy

$$\Psi_N(\xi_1', \xi_2', \ldots, \xi_N') = (-1)^P \Psi_N(\xi_1, \xi_2, \ldots, \xi_N) \tag{9.20}$$

where $P$ is the total number of permutations required to reach the configuration of the wave function on the left from that on the right of (9.20). Thus, if we have for example $\Psi_4(\xi_1, \xi_2, \xi_3, \xi_4)$, then to get to $\Psi_4(\xi_4, \xi_3, \xi_2, \xi_1)$ requires a total of six permutations (three to get $\xi_4$ moved to the beginning, then two to bring $\xi_3$ adjacent to it, and finally one more to interchange $\xi_2$ and $\xi_1$). Hence, this new example yields a symmetric product after the six interchanges. The general permuted wave function (9.20) then has the inner product with the original of

$$\langle \Psi_N(\xi_1', \xi_2', \ldots, \xi_N') | \Psi_N(\xi_1, \xi_2, \ldots, \xi_N) \rangle = (-1)^P. \tag{9.21}$$

It is important to note that the normalization of the many-electron wave function arises from the normalization of the single-electron wave functions that go into the Slater determinant (9.9). The factor of $(-1)^P$ arises from the need to get these into the 'right' product order for taking the associated inner products of the single-electron wave functions. The normalization factor in (9.9) then goes to cancel the multiplicity of terms that arise from the determinantal form. Let us consider just the two-electron case of (9.8) as an example (remember that the adjoint operator on the left of the inner product reverses the order of terms):

$$\begin{aligned}\langle \Psi_2 | \Psi_2 \rangle &= \tfrac{1}{2} \langle [\psi_1(\xi_1)\psi_2(\xi_2) - \psi_2(\xi_1)\psi_1(\xi_2)] | [\psi_1(\xi_1)\psi_2(\xi_2) - \psi_2(\xi_1)\psi_1(\xi_2)] \rangle \\ &= \tfrac{1}{2} [\langle \psi_1(\xi_1)\psi_2(\xi_2) | \psi_1(\xi_1)\psi_2(\xi_2) \rangle + \langle \psi_2(\xi_1)\psi_1(\xi_2) | \psi_2(\xi_1)\psi_1(\xi_2) \rangle \\ &\quad - \langle \psi_2(\xi_1)\psi_1(\xi_2) | \psi_1(\xi_1)\psi_2(\xi_2) \rangle - \langle \psi_1(\xi_1)\psi_2(\xi_2) | \psi_2(\xi_1)\psi_1(\xi_2) \rangle] \\ &= \tfrac{1}{2} [(-1)^0 + (-1)^0 - (-1)^1 0 - (-1)^1 0] = 1 \end{aligned} \tag{9.22}$$

where the zeros arise because the wave functions at the same variables are different and orthogonal to one another. Thus, one needs to be careful that the one-electron wave functions are orthonormal and that the permutations are properly computed.

Finally, we note that it is usually the case that the coordinates are the same for all of the single-electron wave functions; only the spin coordinates differ. Thus, we can think of the differences between the single-particle states as the particular member of the Hilbert space and the particular spin variable that is excited for each electron. No two combinations can be occupied by more than one electron. With this realization, it is possible to think of a simple set of parameters to characterize each single-electron wave function—its index in the Hilbert space and its spin angular momentum. Thus, we may write the single-particle wave functions as $\psi_{i\sigma}(\boldsymbol{r})$. Here, the index $i$ signifies the particular member of the basis set, and $\sigma$ denotes the spin state. These may be combined for simplicity into the index $\lambda = (i, \sigma)$. We will do this in the following discussion. However, we are making a major change of paradigm with this notational change. In the case treated above, it was assumed that the coordinates (including spin) of a particle defined a particular localized wave function in the position representation, and the index of the wave function described the type of wave function at that position. Thus, at the point $\xi = (\boldsymbol{x}, \sigma)$, there may be many types of wave function available, and the $\psi_i$ correspond to these types. In the latter description, however, it is assumed that there is only a single type of wave function at each site, or that there is no localization and that there are a number of wave function types. An example of the first of these options is a Gaussian wave packet localized at $\boldsymbol{x}$. This is characteristic of, for example, localized delta function wave packets, which satisfy $\langle \boldsymbol{x} | \boldsymbol{x}' \rangle = \delta(\boldsymbol{x} - \boldsymbol{x}')$. The latter formulation is characteristic of, for example, the wave functions of an electron in a quantum well, where the 'space' is the region that lies within the well, and the various wave functions are those that arise from the different energy levels. A second example of the latter, and one that will be heavily used, is that of momentum eigenfunctions, which are plane waves extending over all space. The indices are then the momentum values and the spin indices. It is important in evaluating the anti-symmetrized wave functions to understand fully just which interpretation is being placed on the individual single-electron wave functions.

## 9.4 Fermion creation and annihilation operators

In the treatment of the harmonic oscillator, it was found to be useful to introduce a set of commuting operators which described the creation or annihilation of one unit of energy in the harmonic oscillator. This also changed the wave functions accordingly. Can we use a similar description to enhance our understanding of the many-electron picture? The answer is obviously yes, but we must carefully examine how the rules will be changed by the requirement of anti-symmetry and the corresponding Pauli exclusion principle. New views that arise will be based upon the fact that the energy of the harmonic oscillator could be continuously raised by pumping phonons into the system; the introduction of $n$ phonons could be achieved by using the operator $(a^+)^n$. Here, however, operation with a single creation operator (we use the notation $c$ for fermions) $c_\lambda^+$ creates one fermion in state $\lambda$. This state may be a momentum eigenstate, a state in a quantum well, or any other state in a system in which the wave functions span the entire allowed variable space. If we try to create a second fermion in this state, the result must be forced to vanish because of the exclusion principle; for example,

$$\left(c_\lambda^+\right)^2 |0\rangle = 0 \tag{9.23}$$

where $|0\rangle$ is the so-called vacuum state in which no fermions exist. In fact, however, equation (9.23) must hold for any wave function in which the state $\lambda$

is empty (or even filled). Similarly,

$$c_\lambda^2|\ldots\rangle = 0 \tag{9.24}$$

where $|\ldots\rangle$ is any state of the system (since there can be no more than a single electron in any state, there cannot be more than one annihilation).

This suggests that a different combination of products must be used for these fermion operators. Consider, for example, the product

$$c_\lambda^+ c_\lambda + c_\lambda c_\lambda^+. \tag{9.25}$$

If the state $|\lambda\rangle$ is empty, the first term immediately gives zero. The second term creates a fermion in the state, then destroys it, so the result is, by (4.67), $(0 + 1) = 1$. Similarly, if the state $|\lambda\rangle$ is occupied, the second term gives zero (another fermion cannot be created), while the first term is the number operator and yields 1. Thus, the result in either case is the *anti-commutator* product

$$\{c_\lambda^+, c_\lambda\} = c_\lambda^+ c_\lambda + c_\lambda c_\lambda^+ = 1. \tag{9.26}$$

Here, we use the curly brackets to indicate that the positive sign is used in the anti-commutator, as opposed to the negative sign used in the commutator relation. This may be extended to operators on other states as

$$\{c_\lambda^+, c_\mu\} = \delta_{\lambda\mu} \tag{9.27$a$}$$

$$\{c_\lambda^+, c_\mu^+\} = \{c_\lambda, c_\mu\} = 0. \tag{9.27$b$}$$

This leads to an interesting result for the number of fermions that can exist in, or arise from, any state of the system. Since we can rewrite (9.26) as $c_\lambda^+ c_\lambda = 1 - c_\lambda c_\lambda^+$, we have

$$(c_\lambda^+ c_\lambda)^2 = c_\lambda^+ c_\lambda(1 - c_\lambda c_\lambda^+) = c_\lambda^+ c_\lambda - c_\lambda^+ c_\lambda c_\lambda c_\lambda^+ = c_\lambda^+ c_\lambda \tag{9.28}$$

from (9.24). Thus, $n_\lambda^2 = n_\lambda = 1$ if the state is occupied (it is naturally zero otherwise, a result independent of the choice of whether to use boson operators or fermion operators). The representation in which the number operators are diagonal, along with the energy, is known as the number representation.

Note that the operators $c_\lambda^+$ and $c_\lambda$ used here are fermion operators. The relationship (9.23) ensures that no more than a single fermion can exist in the given state. The attempt to put a second particle in this state must yield zero, as given by (9.23). Similarly, only a single fermion can be removed from an occupied state, as indicated by (9.24). These statistics differ markedly from the operators used in chapters 4 and 8, which were for bosons. While the wave function differed for each state, we referred to a particular harmonic oscillator mode as being described by its occupation. In increasing the occupation, we raised the energy of the mode (given by the number of bosons in that particular harmonic oscillator mode). In the boson harmonic oscillator, the energy of a particular mode was given by

$$E_i = (n_i + \tfrac{1}{2})\hbar\omega_i \qquad n_i = 0, 1, 2, \ldots . \tag{9.29}$$

Thus, there could be a great many bosons occupying the mode, even though the distinct mode wave function would change. In the case of fermions, however, we limit the range of the occupation factor $n_i$ to be only 0 or 1. This is imposed by using the same generation properties of the creation and annihilation operators, but with the limitations for fermions that

$$(c_\lambda)^2 = 0 \tag{9.30$a$}$$

and

$$(c_\lambda^+)^2 = 0. \tag{9.30$b$}$$

These properties are now carried over to the many-particle wave function.

The general many-electron wave function is created by operating on the empty state, or vacuum state, with the operators for positioning the electrons where desired. For example, a three-electron state may be created as

$$|\mu\nu\lambda\rangle = c_\mu^+ c_\nu^+ c_\lambda^+ |0\rangle. \tag{9.31}$$

It should be noted that the order of creation of the particles is important, since changing the order results in a permutation of the indices, and

$$|\mu\lambda\nu\rangle = -|\mu\nu\lambda\rangle. \tag{9.32}$$

This may be generalized to an arbitrary many-electron wave function

$$|n_1 n_2 \ldots n_\infty\rangle = (c_1^+)^{n_1}(c_2^+)^{n_2} \ldots (c_\infty^+)^{n_\infty}|0\rangle. \tag{9.33}$$

Then,

$$c_\lambda^+|\ldots n_\lambda \ldots\rangle = \begin{cases} |\ldots(n_\lambda + 1)\ldots\rangle & \text{if } n_\lambda = 0 \\ 0 & \text{if } n_\lambda = 1 \end{cases} \tag{9.34}$$

and

$$c_\lambda|\ldots n_\lambda \ldots\rangle = \begin{cases} |\ldots(n_\lambda - 1)\ldots\rangle & \text{if } n_\lambda = 1 \\ 0 & \text{if } n_\lambda = 0\,. \end{cases} \tag{9.35}$$

It is possible to recognize that the creation and annihilation operators do not operate in a simple Hilbert space. In general, the creation operator operates in a space of $n$ electrons and moves to a space with $n + 1$ electrons. Similarly, the annihilation operator moves to a space with $n - 1$ electrons. The general space may then be a product space of Hilbert spaces, and elements may be combined with a variety of partially occupied wave functions. This complicated structure is called a *Fock space*, but the details of this structure are beyond the simple treatment that we desire here.

## 9.5 Field operators

We have seen in the above sections that the fermion creation and annihilation operators may be used to put electrons into and take electrons from particular states in a complete Hilbert space. In (2.92) we expanded the arbitrary wave function solution to the time-independent Schrödinger equation in terms of these very same basis functions, which are the eigenfunctions of the equation itself. There, we used a set of expansion coefficients to provide the weights for each of the basis functions. Here, however, we cannot excite a fraction of an electron—a basis state either has an electron in it or it does not. This means that we can use the creation operators and the annihilation operators to define the overall wave function for fermions as

$$\Psi(\xi) = \sum c_\lambda \phi_\lambda(\xi) \tag{9.36}$$

where the wave functions of the basis set include the appropriate spin functions, and the coordinates include both position and spin coordinates. Similarly, we can write

$$\Psi^+(\xi) = \sum c_\lambda^+ \phi_\lambda^*(\xi). \tag{9.37}$$

Here, the expansion functions satisfy

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{x})\right)\phi_\lambda(\xi) = \mathcal{E}_\lambda \phi_\lambda(\xi) \tag{9.38}$$

where there are no specific spin-dependent operators in the normal Hamiltonian. It should be remarked that these are single-electron wave functions even though there are excitations into various states, since there are no products of wave functions for which to require proper anti-symmetry. Two-electron wave functions can only be created by products of creation operators, and similarly for more densely populated systems. This is a problem that we will have to address below, where we will make the connection between this description of a many-electron system with one-electron wave functions and the many-electron wave functions introduced above.

The interpretation of (9.36) and (9.37) as wave functions is complicated as they are now *operators*—the expansion coefficients are creation and annihilation operators. These quantities are called field operators. The concept of field here is precisely the one that is used in electromagnetic *field theory*. Normally, quantization is invoked via non-commuting operators such as position and momentum. Here, however, these operators are buried in the creation and annihilation operators, and the normally simple wave functions have now become operators in their own right. We have quantized the normally $c$-number wave functions. This process is usually referred to as *second quantization*, and we want to examine this concept further here. In this approach, the Hamiltonian operator appears in a totally different light (but does not invoke different physics).

It is easy now to examine the behaviour of the field operators under the anti-commutation operations. For example,

$$\begin{aligned}
\{\Psi^+(\xi), \Psi(\xi')\} &= \Psi^+(\xi)\Psi(\xi') + \Psi(\xi')\Psi^+(\xi) \\
&= \sum_\lambda c_\lambda^+ \phi_\lambda^*(\xi) \sum_\mu c_\mu \phi_\mu(\xi') + \sum_\mu c_\mu \phi_\mu(\xi') \sum_\lambda c_\lambda^+ \phi_\lambda^*(\xi) \\
&= \sum_\lambda \phi_\lambda^*(\xi) \sum_\mu \phi_\mu(\xi') \{c_\lambda^+, c_\mu\} \\
&= \sum_\lambda \phi_\lambda^*(\xi) \sum_\mu \phi_\mu(\xi') \delta_{\lambda\mu} \\
&= \sum_\lambda \phi_\lambda^*(\xi)\phi_\lambda(\xi') = \delta(\xi - \xi'). \qquad (9.39)
\end{aligned}$$

This last line is precisely the principle of closure of a complete set (5.38). In a sense, the anti-commutation of the fermion operators ensures the orthonormality of the basis functions in a way connected with the Pauli principle. Similarly,

$$\{\Psi^+(\xi), \Psi^+(\xi')\} = \{\Psi(\xi), \Psi(\xi')\} = 0. \qquad (9.40)$$

We can now use these field operators to show how the Schrödinger equation appears in second quantization, and we can then relate these operators to the proper anti-symmetrized many-electron wave functions.

### 9.5.1 Connection with the many-electron formulation

Let us now look at how the energy levels and wave functions may be connected between the field operator forms and the many-electron wave function forms. Here, we follow the approach of Haken (1976). We begin by assuming that a proper vacuum state $|0\rangle$ exists such that

$$c_\lambda|0\rangle = 0 \qquad \text{for all } \lambda. \qquad (9.41)$$

We can then write an arbitrary many-electron wave function as (9.33), which we rewrite as

$$|\{n\}\rangle = \prod_\lambda (c_\lambda^+)^{n_\lambda}|0\rangle. \qquad (9.42)$$

Since $(c_\lambda^+)^2 = 0$, the allowed values for $n_\lambda$ are only 0 or 1 (formally, we have set $(c_\lambda^+)^0 = 1$). One possible set of values for the wave function is given by

$$\{n\} = \{1, 0, 0, 1, 1, 0, \ldots\}. \qquad (9.43)$$

The energy for this wave function is given by the energy of the excited states, and may be written as

$$\mathcal{E} = \sum_\lambda \mathcal{E}_\lambda n_\lambda. \qquad (9.44)$$

If the total number of electrons is $N$, then there are exactly this number of 1s in the set (9.43). With this in mind, we can write an ordered representation for (9.42) as

$$|\{n\}\rangle = c^{+}_{\lambda_1} c^{+}_{\lambda_2} c^{+}_{\lambda_3} \ldots c^{+}_{\lambda_n} |0\rangle \tag{9.45}$$

where it is assumed that

$$\lambda_1 < \lambda_2 < \lambda_3 < \cdots < \lambda_n \tag{9.46}$$

orders the creation of the electrons by some measure, which may be arrived at quite arbitrarily. This measure of ordering provides the 'yardstick' via which we enforce anti-symmetry through the exchange of any two creation operators.

In order to see how this formulation now compares with the field operator approach, we will look specifically at a simple one-electron form of this wave function. We shall assume that the correct (in some measure) single-electron wave function may be written as a weighted summation over the possible one-electron states that can occur from (9.42) (there are almost an infinity of these, of course). We write this weighted sum in exactly the same manner as used in previous chapters; for example,

$$|1\rangle = \sum_{\lambda} \alpha_{\lambda} c^{+}_{\lambda} |0\rangle \tag{9.47}$$

where the $\alpha_{\lambda}$ are $c$-number coefficients. The creation operator can be eliminated by multiplying (9.37) by $\phi_{\mu}(\xi)$ and integrating over the coordinates (including a summation over the spin coordinates), so

$$|1\rangle = \int \sum_{\lambda} \alpha_{\lambda} \phi_{\lambda}(\xi) \Psi^{+}(\xi)\, \mathrm{d}\xi\, |0\rangle = \int f(\xi) \Psi^{+}(\xi)\, \mathrm{d}\xi\, |0\rangle \tag{9.48}$$

in which we have defined a spatially varying weight function (which will become a wave function in the manner of previous chapters)

$$f(\xi) = \sum_{\lambda} \alpha_{\lambda} \phi_{\lambda}(\xi). \tag{9.49}$$

It will be shown below that the weight function $f(\xi)$ satisfies the Schrödinger equation (for one particle). Equation (9.48) may now be interpreted as $\Psi^{+}(\xi)$ creating a single particle at the position $\boldsymbol{x}$ and for spin $\sigma$ (recall that $\xi$ is the set of a position vector and a spin index), with the probability of any particular value of $\xi$ given by the weight function.

The above interpretation that $\Psi^{+}(\xi)$ creates a particle at $\xi$ can be verified by operating on this function with the particle density operator $\Psi^{+}(\xi')\Psi(\xi')$:

$$\begin{aligned}\Psi^{+}(\xi')\Psi(\xi')\Psi^{+}(\xi)|0\rangle &= \Psi^{+}(\xi')[\delta(\xi' - \xi) - \Psi^{+}(\xi)\Psi(\xi')]|0\rangle \\ &= \Psi^{+}(\xi')\delta(\xi' - \xi)|0\rangle + \Psi^{+}(\xi)\Psi^{+}(\xi')\Psi(\xi')|0\rangle \\ &= \Psi^{+}(\xi')\delta(\xi' - \xi)|0\rangle. \end{aligned} \tag{9.50}$$

Similarly, the result can also be written (we note that the density operator produces a zero when operating on the vacuum state) as

$$\Psi^+(\xi')\Psi(\xi')\Psi^+(\xi)|0\rangle = \delta(\xi' - \xi)\,\Psi^+(\xi')|0\rangle. \tag{9.50$a$}$$

This shows that the state $\Psi^+(\xi)|0\rangle$ is an eigenstate of the density operator, with the value 0 if the created particle's position and spin $\xi$ do not correspond with that of the density operator.

### 9.5.2 Quantization of the Hamiltonian

The one-electron wave function is now a function of the field operator. To proceed, we will now need to quantize the Schrödinger equation itself, which really means that we want to work out how to write the Hamiltonian in a second-quantized form in terms of the field operators. Quantization of a field is the usual method by which classical (electromagnetic) fields are quantized to produce the quantum mechanical equivalent fields. Here, however, our 'field' is described by the Schrödinger equation. The normal manner in which the quantization proceeds is through the use of Lagrange's equations, which is an approach beyond the level to be maintained here. Therefore, we will proceed with a heuristic approach. The idea is as follows: the field operators are expansions in terms of the basis set of functions in our Hilbert space, so each operator is a 'vector position' in the Hilbert space, which we will term a pseudo-position. Similarly, the motion of this field operator is a pseudo-momentum. We want to generate the second-quantized Hamiltonian by writing it in terms of these dynamic pseudo-variables, remembering that the spatial operators will really operate on the density operator. We reach the operator form by integrating over all space (and summing over all spins), so the second-quantized Hamiltonian operator is expressed as

$$\begin{aligned} H &= \int \left[\frac{1}{2m}\boldsymbol{P}^*(\xi)\cdot\boldsymbol{P}(\xi) + V(\xi)X^*(\xi)X(\xi)\right] \mathrm{d}\xi \\ &\to \int \left[(\hbar^2/2m)\boldsymbol{\nabla}\Psi^+(\xi)\cdot\boldsymbol{\nabla}\Psi(\xi) + V(\xi)\Psi^+(\xi)\Psi(\xi)\right] \mathrm{d}\xi \\ &= \int \Psi^+(\xi)\left[-(\hbar^2/2m)\nabla^2 + V(\xi)\right]\Psi(\xi)\,\mathrm{d}\xi \end{aligned} \tag{9.51}$$

where the first term on the right-hand side has been integrated by parts. Now, the Hamiltonian is an operator written in terms of the field operators, and is said to be second quantized.

Let us now use this Hamiltonian operator to examine the one-electron wave function (9.48). This becomes

$$H|1\rangle = \int \Psi^+(\xi)\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\xi)\right]\Psi(\xi)\,\mathrm{d}\xi \int f(\xi')\Psi^+(\xi')\,\mathrm{d}\xi'\,|0\rangle. \tag{9.52}$$

Using the anti-commutation relations of the field operators, this can be rewritten in the form (the second term arising from the anti-commutator vanishes as usual)

$$H|1\rangle = \iint \Psi^+(\xi)\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\xi)\right]\delta(\xi-\xi')f(\xi')|0\rangle\,\mathrm{d}\xi\,\mathrm{d}\xi'. \tag{9.53}$$

Since the Laplacian operator produces the same result when operating on the delta function, regardless of whether the primed or unprimed coordinates are used, and the Hamiltonian is Hermitian, equation (9.53) can be rewritten as

$$H|1\rangle = \int \Psi^+(\xi)|0\rangle\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\xi)\right]f(\xi)\,\mathrm{d}\xi. \tag{9.54}$$

Since $H|1\rangle = \mathcal{E}|1\rangle$, and $\Psi^+(\xi)|0\rangle$ is linearly independent of the weight function, the conclusion to be drawn is that the weight function must satisfy the Schrödinger equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\xi)\right]f(\xi) = \mathcal{E}f(\xi). \tag{9.55}$$

Thus, we have transferred the need to satisfy the Schrödinger equation from the basis functions, which are assumed to provide the coordinate system in the Hilbert space, to the weight function that describes the one-electron wave function in terms of the field operators.

The fact that the weight function now satisfies the Schrödinger equation tells us that the field operator approach, based upon the use of creation and annihilation operators, produces a proper treatment of at least the one-electron problem in quantum mechanics. The weight function is a proper wave function itself, and satisfies the Schrödinger equation. As stated earlier, if the one-electron problem were our only interest, this more powerful development would be worthless to us, for its complexity is not worth the extra effort. However, when we move to multi-electron problems, the power of the approach becomes apparent. The many-electron wave function is expressed equivalently in terms of the field operators or the creation and annihilation operators that make up the field operators. Consequently, complicated quantum mechanical calculations may be expressed in terms of the simple anti-commutator algebra of the anti-commuting operators for fermions. Let us continue to the two-electron problem, and show that this approach is quite general.

### 9.5.3 The two-electron wave function

We now turn to a general two-particle treatment. In addition to the normal Schrödinger equation with two electrons excited into the system, we will also introduce the Coulomb interaction between the pair of particles. This will now produce an interacting system, since there is now a force between each pair of particles due to this potential term. Once we begin to treat the many-electron system, particularly with interacting particles, the second-quantized approach begins to show its value, as there are many processes that can only be treated adequately once this approach is adopted. Let us begin with the most general two-particle state, which is written as above as a summation over all possible two-particle states, as

$$|2\rangle = \sum_{\lambda\mu} \alpha_{\lambda\mu} c_\lambda^+ c_\mu^+ |0\rangle. \tag{9.56}$$

The procedure to be followed is exactly that used in the previous paragraphs. We will first replace the creation operators and find a general function that will be required to satisfy the Schrödinger equation. Thus, we use (9.48) twice, so we have

$$\begin{aligned} |2\rangle &= \sum_{\lambda\mu} \alpha_{\lambda\mu} \int \phi_\lambda(\xi)\Psi^+(\xi)\,\mathrm{d}\xi \int \phi_\mu(\xi')\Psi^+(\xi')\,\mathrm{d}\xi'|0\rangle \\ &= \iint f(\xi,\xi')\Psi^+(\xi)\Psi^+(\xi')\,\mathrm{d}\xi\,\mathrm{d}\xi'|0\rangle \end{aligned} \tag{9.57}$$

where

$$f(\xi,\xi') = \sum_{\lambda\mu} \alpha_{\lambda\mu}\phi_\lambda(\xi)\phi_\mu(\xi'). \tag{9.58}$$

It must now be shown that the function $f(\xi,\xi')$ is properly anti-symmetric with respect to its two coordinates. To do so, we take the last expression from (9.57) and interchange the two coordinates:

$$\begin{aligned} |2\rangle &= \iint f(\xi,\xi')\Psi^+(\xi')\Psi^+(\xi)\,\mathrm{d}\xi\,\mathrm{d}\xi'\,|0\rangle \\ &= -\iint f(\xi,\xi')\Psi^+(\xi)\Psi^+(\xi')\,\mathrm{d}\xi\,\mathrm{d}\xi'\,|0\rangle. \end{aligned} \tag{9.59}$$

In the last line, the two field operators were interchanged using the anti-commutation relation (9.40). Now, these operations really do not change (9.57), since the coordinates are true dummy variables which are being integrated out of the problem. Thus, expression (9.59) must be equivalent to (9.57) which is only possible if

$$f(\xi',\xi) = -f(\xi,\xi') \tag{9.60}$$

and it is in this function that the anti-symmetry is imposed directly (it is still imposed on the field operators by the appropriate anti-commutation relations).

We can now use the Hamiltonian operator (9.53) to determine the equation for the two-electron wave function $f(\xi',\xi)$. This leads to, for the simple one-particle operators,

$$\begin{aligned} H_1|2\rangle = &\int \Psi^+(\xi)\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\xi)\right]\Psi(\xi)\,\mathrm{d}\xi \\ &\times \iint f(\xi'',\xi')\Psi^+(\xi'')\Psi^+(\xi')\,\mathrm{d}\xi''\,\mathrm{d}\xi'\,|0\rangle. \end{aligned} \tag{9.61}$$

Now,

$$\Psi(\xi)\Psi^{+}(\xi'')\Psi^{+}(\xi') = \delta(\xi-\xi'')\,\Psi^{+}(\xi') - \Psi^{+}(\xi'')\delta(\xi-\xi') \tag{9.62}$$

so, with (9.60), equation (9.61) becomes

$$\begin{aligned} H_1|2\rangle = &\iint \Psi^{+}(\xi'')\Psi^{+}(\xi')\left[-\frac{\hbar^2}{2m}\nabla''^2 + V(\xi'')\right] f(\xi'',\xi')\,\mathrm{d}\xi''\,\mathrm{d}\xi'|0\rangle \\ &- \iint \Psi^{+}(\xi')\Psi^{+}(\xi'')\left[-\frac{\hbar^2}{2m}\nabla'^2 + V(\xi')\right] f(\xi'',\xi')\,\mathrm{d}\xi''\,\mathrm{d}\xi'|0\rangle. \end{aligned} \tag{9.63}$$

Using the anti-commutator relations for the field operators, the sign between the two terms will be changed, and they can be combined. Thus, the one-particle Hamiltonian leads to the Schrödinger equation for the function $f(\xi,\xi')$, if we include the energy term as in the previous section,

$$\left[-\frac{\hbar^2}{2m}(\nabla^2+\nabla'^2) + V(\xi) + V(\xi')\right] f(\xi,\xi') = \mathcal{E} f(\xi,\xi'). \tag{9.64}$$

Let us now turn to the two-electron interaction potential.

The Coulomb interaction arises between two charge densities. This can be expressed as

$$H_2 = \frac{1}{2}\iint \rho(\boldsymbol{x})\frac{1}{4\pi\varepsilon|\boldsymbol{x}-\boldsymbol{x}'|}\rho(\boldsymbol{x}')\,\mathrm{d}\xi\,\mathrm{d}\xi' \tag{9.65}$$

where the spin index is summed over. This can be extended by the introduction of the field operators through the particle density operator used in (9.50) and

$$\begin{aligned} H_2 &= \frac{1}{2}\iint \Psi^{+}(\xi)\Psi(\xi)\frac{e^2}{4\pi\varepsilon|\boldsymbol{x}-\boldsymbol{x}'|}\Psi^{+}(\xi')\Psi(\xi')\,\mathrm{d}\xi\,\mathrm{d}\xi' \\ &= \frac{1}{2}\iint \Psi^{+}(\xi)\Psi^{+}(\xi')\frac{e^2}{4\pi\varepsilon|\boldsymbol{x}-\boldsymbol{x}'|}\Psi(\xi')\Psi(\xi)\,\mathrm{d}\xi\,\mathrm{d}\xi' \end{aligned} \tag{9.66}$$

and, using (9.59) and the commutator relations, we have

$$H_2|2\rangle = \iint \Psi^{+}(\xi)\Psi^{+}(\xi')\frac{e^2}{4\pi\varepsilon|\boldsymbol{x}-\boldsymbol{x}'|} f(\xi,\xi')\,\mathrm{d}\xi\,\mathrm{d}\xi'\,|0\rangle. \tag{9.67}$$

Thus, the full two-electron Schrödinger equation for the two-electron function is given by

$$\left[-\frac{\hbar^2}{2m}(\nabla^2+\nabla'^2) + V(\xi) + V(\xi') + \frac{e^2}{4\pi\varepsilon|\boldsymbol{x}-\boldsymbol{x}'|}\right] f(\xi,\xi') = \mathcal{E} f(\xi,\xi'). \tag{9.68}$$

This is easily extended to an arbitrary $N$-electron wave function, which yields

$$\left\{\sum_j\left[-\frac{\hbar^2}{2m}\nabla_j^2+V(\xi_j)\right]+\sum_{i<j}\frac{e^2}{4\pi\varepsilon|\boldsymbol{x}_j-\boldsymbol{x}_i|}\right\}f(\xi_1,\xi_2,\dots,\xi_N)$$
$$=\mathcal{E}f(\xi_1,\xi_2,\dots,\xi_N). \tag{9.69}$$

Even with this multi-electron wave function, the Hamiltonian is still expressible in terms of the field operators as

$$H=\int\Psi^+(\xi)\left[-\frac{\hbar^2}{2m}\nabla^2+V(\xi)\right]\Psi(\xi)\,\mathrm{d}\xi$$
$$+\frac{1}{2}\iint\Psi^+(\xi)\Psi^+(\xi')\frac{e^2}{4\pi\varepsilon|\boldsymbol{x}-\boldsymbol{x}'|}\Psi(\xi')\Psi(\xi)\,\mathrm{d}\xi\,\mathrm{d}\xi'. \tag{9.70}$$

### 9.5.4 The homogeneous electron gas

In the case of a homogeneous electron gas, the function $f(\xi_1,\xi_2,\dots,\xi_N)$ can be expected to be a combination of plane waves; that is, each of the basis states from which this function is made up is itself a plane wave (in momentum representation). In fact, there are several modifications to this, which we will examine in a subsequent section, but for now we assume that the basis states are simply

$$\phi_\lambda(\xi)=\frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\eta_\sigma \tag{9.71}$$

where $\eta_\sigma$ is the spin function, which can be represented by the matrices of (9.10). These basis functions can now be used in the representations for the field operators to rewrite the Hamiltonian. The first term (we neglect the potential term for the homogeneous gas) can be written as

$$H_1=\frac{1}{V}\sum_{\boldsymbol{k}\sigma}\sum_{\boldsymbol{k}'\sigma'}\int c^+_{\boldsymbol{k}\sigma}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\eta^+_\sigma\left[-\frac{\hbar^2}{2m}\right]\nabla^2c_{\boldsymbol{k}'\sigma'}\mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{x}}\eta_{\sigma'}\,\mathrm{d}\xi$$
$$=\frac{1}{V}\sum_{\boldsymbol{k}\sigma}\sum_{\boldsymbol{k}'\sigma'}\frac{\hbar^2k^2}{2m}c^+_{\boldsymbol{k}\sigma}c_{\boldsymbol{k}'\sigma'}V\delta(\boldsymbol{k}-\boldsymbol{k}')\,\delta_{\sigma\sigma'}$$
$$=\sum_{\boldsymbol{k}\sigma}\frac{\hbar^2k^2}{2m}c^+_{\boldsymbol{k}\sigma}c_{\boldsymbol{k}\sigma} \tag{9.72}$$

where the delta function in position arises from the integration over the position variable and that in the spin index arises from the spin operators. The latter arises from the fact that the adjoint spin operator is a row matrix, and the product of the row matrix and the column matrix is zero unless the spin states are the same. In this latter case, the product is unity.

For the interaction term, we must introduce the Fourier transform of the interaction potential in order to proceed. When this is done, this term becomes

$$\begin{aligned}
H_2 &= \frac{1}{2V^2} \sum_{\boldsymbol{k}\sigma\boldsymbol{k}'\sigma'\boldsymbol{k}''\sigma''\boldsymbol{k}'''\sigma'''} \int c^+_{\boldsymbol{k}\sigma} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \eta^+_\sigma \int c^+_{\boldsymbol{k}'\sigma'} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{x}'} \eta^+_{\sigma'} \\
&\quad \times \frac{e^2}{4\pi\varepsilon|\boldsymbol{x}-\boldsymbol{x}'|} c_{\boldsymbol{k}''\sigma''} \mathrm{e}^{\mathrm{i}\boldsymbol{k}''\cdot\boldsymbol{x}'} \eta_{\sigma''} c_{\boldsymbol{k}'''\sigma'''} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'''\cdot\boldsymbol{x}} \eta_{\sigma'}\, \mathrm{d}\xi'\, \mathrm{d}\xi \\
&= \frac{1}{2V^2} \sum_{\boldsymbol{k}\sigma\boldsymbol{k}'\sigma'\boldsymbol{k}''\sigma''\boldsymbol{k}'''\sigma'''} \int c^+_{\boldsymbol{k}\sigma} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \eta^+_\sigma \int c^+_{\boldsymbol{k}'\sigma'} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{x}'} \eta^+_{\sigma'} \\
&\quad \times \sum_q \frac{e^2}{\varepsilon q^2} \varepsilon^{\mathrm{i}\boldsymbol{q}\cdot(\boldsymbol{x}-\boldsymbol{x}')} c_{\boldsymbol{k}''\sigma''} \mathrm{e}^{\mathrm{i}\boldsymbol{k}''\cdot\boldsymbol{x}'} \eta_{\sigma''} c_{\boldsymbol{k}'''\sigma'''} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'''\cdot\boldsymbol{x}} \eta'_\sigma\, \mathrm{d}\xi'\, \mathrm{d}\xi. \qquad (9.73)
\end{aligned}$$

Now, the integrations can be carried out over $\boldsymbol{x}$ and $\boldsymbol{x}'$ separately. These give delta functions that provide $\boldsymbol{k}''' = \boldsymbol{k} - \boldsymbol{q}$, and $\boldsymbol{k}'' = \boldsymbol{k}' + \boldsymbol{q}$. Moreover, the spin functions are similarly combined, and we can now write

$$\begin{aligned}
H_2 &= \tfrac{1}{2} \sum_{\boldsymbol{k}\sigma\boldsymbol{k}'\sigma'\boldsymbol{q}} c^+_{\boldsymbol{k}\sigma} c^+_{\boldsymbol{k}'\sigma'} \frac{e^2}{\varepsilon q^2} c_{(\boldsymbol{k}'+\boldsymbol{q})\sigma'} c_{(\boldsymbol{k}-\boldsymbol{q})\sigma} \\
&= \tfrac{1}{2} \sum_{\boldsymbol{k}\sigma\boldsymbol{k}'\sigma'\boldsymbol{q}} \frac{e^2}{\varepsilon q^2} c^+_{(\boldsymbol{k}+\boldsymbol{q})\sigma} c^+_{(\boldsymbol{k}'-\boldsymbol{q})\sigma'} c_{\boldsymbol{k}'\sigma'} c_{\boldsymbol{k}\sigma} \qquad (9.74)
\end{aligned}$$

where we have shifted the axes of the momentum vectors in the last line to a more usual notation. The interpretation is that the $\boldsymbol{q}$ momentum component of the Coulomb interaction potential scatters one electron from $\boldsymbol{k}$ to $\boldsymbol{k} + \boldsymbol{q}$ while scattering the second of the interacting pair from $\boldsymbol{k}'$ to $\boldsymbol{k}' - \boldsymbol{q}$, all the while preserving the spin of the two electrons. When $\boldsymbol{q} = \boldsymbol{0}$, a problem arises and one must resort to using a convergence factor, which can be set to zero after the summations are taken. However, this latter term is precisely the component that cancels the uniform background charge (the neutralizing charge that we have said nothing about). Thus, in (9.74), the term for $\boldsymbol{q} = \boldsymbol{0}$ must be omitted from the summation.

The total Hamiltonian can now be written by combining (9.72) and (9.74) into a single term, which becomes

$$H = \sum_{\boldsymbol{k}\sigma} \frac{\hbar^2 k^2}{2m} c^+_{\boldsymbol{k}\sigma} c_{\boldsymbol{k}\sigma} + \tfrac{1}{2} \sum_{\boldsymbol{k}\sigma\boldsymbol{k}'\sigma'\boldsymbol{q}} \frac{e^2}{\varepsilon q^2} c^+_{(\boldsymbol{k}+\boldsymbol{q})\sigma} c^+_{(\boldsymbol{k}'-\boldsymbol{q})\sigma'} c_{\boldsymbol{k}'\sigma'} c_{\boldsymbol{k}\sigma}. \qquad (9.75)$$

This form of the Hamiltonian is often found in introductory texts on quantum effects in solids, and is sometimes referred to as the number representation, since the number operator appears in the first term. However, it should be noted that this representation is only for the homogeneous electron gas, and it has been obtained with single-electron wave functions for the basis functions. In the next section, we will begin to examine how the general many-electron wave function can be expressed in single-electron functions.

Let us now recap on the introduction of the field operators. These functions allow us to talk about the excitation of an electron into the system, with a weight function describing the amount of each basis state included in the description. The advantage of this is that it allows us to talk about a quantized Hamiltonian, and to have a weight function for 1, 2, ... , $N$ electrons that satisfies the one-, two-, ... , $N$-electron Schrödinger equation. Instead of now having to solve this equation, we work on the basic algebra of the anti-commuting field operators. We will see, in the next section, that the need to solve the $N$-electron Schrödinger equation remains with us, and in the end we develop an approximation scheme that is essentially just the time-dependent perturbation theory of chapter 7. What the field operators then give us is a methodology for approaching complicated interactions in a simplified manner, in which the work is going to be in the careful evaluation of the proper terms in the perturbation series. The next section will give us a formal procedure for evaluating these terms without having to work through the multiple integrals and complications that can arise in such involved problems.

## 9.6 The Green's function

In the previous section, the Hamiltonian was written in terms of field operators defined in a Hilbert space of basis functions. Products of the form $\Psi^+(\xi')\Psi(\xi)$ arise in this expression. We can show this in the first term by inserting the expansion of a delta function, and (9.70) becomes

$$H = \int \mathrm{d}\xi \int \mathrm{d}\xi' \left[ -\frac{\hbar^2}{2m}\delta(\xi - \xi')\, \Psi^+(\xi')\, \nabla^2\Psi(\xi) \right.$$
$$\left. + \tfrac{1}{2}\Psi^+(\xi)\Psi^+(\xi')\frac{e^2}{4\pi\varepsilon|\boldsymbol{x} - \boldsymbol{x}'|}\Psi(\xi')\Psi(\xi) \right]. \qquad (9.76)$$

Let us assume that there is a *ground state* $|\mathcal{F}\rangle$ of the system, in which all states up to the Fermi energy are filled and all states above the Fermi energy are empty at $T = 0$ K. We assume that this ground state has the lowest energy for the system, and that

$$H|\mathcal{F}\rangle = \mathcal{E}_0|\mathcal{F}\rangle. \qquad (9.77)$$

Similarly, we can introduce the time variation of the field operators by going over to the time-dependent Schrödinger equation (which introduces the time variation of each of the basis states that may arise through perturbation theory in the interaction representation, as the basis states do not vary with time in the simpler Heisenberg representation). Nevertheless, we may introduce a time variation in the definition of the field operators. Simultaneously, we can also introduce the idea of the ensemble average of the Hamiltonian through the inner product $\langle H\rangle = \langle\mathcal{F}|H|\mathcal{F}\rangle$ (at present, we shall assume that the normalizing denominator $\langle\mathcal{F}|\mathcal{F}\rangle = 1$, but we will reconsider this normalization term later). From these simple ideas, we can define the time-ordered Green's function through the definition

$$\mathrm{i}G_{\sigma\sigma'}(\boldsymbol{x},t;\boldsymbol{x}',t') = \langle\mathcal{F}|T[\Psi(\boldsymbol{x},t)\Psi^{+}(\boldsymbol{x}',t')]|\mathcal{F}\rangle \tag{9.78}$$

where $T$ is the *time-ordering operator*. That is,

$$\mathrm{i}G_{\sigma\sigma'}(\boldsymbol{x},t;\boldsymbol{x}',t') = \langle\Psi_{\sigma}(\boldsymbol{x},t)\Psi^{+}_{\sigma'}(\boldsymbol{x}',t')\rangle \qquad t > t' \tag{9.79$a$}$$
$$\mathrm{i}G_{\sigma\sigma'}(\boldsymbol{x},t;\boldsymbol{x}',t') = -\langle\Psi^{+}_{\sigma'}(\boldsymbol{x}',t')\Psi_{\sigma}(\boldsymbol{x},t)\rangle \qquad t < t'. \tag{9.79$b$}$$

The subscripts in (9.78) and (9.79) are the spin indices for the two field operators, and represent a summation that appears in the Hamiltonian term (9.76). These functions have quite simple interpretations, and describe the response of the degenerate Fermi gas to perturbations. In general, these functions may be written as

$$\mathrm{i}G_{\mathrm{r},\sigma\sigma'}(\boldsymbol{x},t;\boldsymbol{x}',t') = \Theta(t-t')\langle\{\Psi_{\sigma}(\boldsymbol{x},t),\Psi^{+}_{\sigma'}(\boldsymbol{x}',t')\}\rangle \tag{9.80$a$}$$
$$\mathrm{i}G_{\mathrm{a},\sigma\sigma'}(\boldsymbol{x},t;\boldsymbol{x}',t') = -\Theta(t'-t)\langle\{\Psi^{+}_{\sigma'}(\boldsymbol{x}',t'),\Psi_{\sigma}(\boldsymbol{x},t)\}\rangle \tag{9.80$b$}$$

for the *retarded* and *advanced* Green's functions, respectively, in keeping with the definitions of chapter 2. The retarded Green's function defines the response of the Fermi gas when an additional electron is introduced at $(\boldsymbol{x}',t')$, propagates to $(\boldsymbol{x},t)$, and then is annihilated. The advanced function describes the reverse process, or as more commonly described, the destruction of an electron through the creation of a hole in the gas at $(\boldsymbol{x},t)$, the propagation of this hole to $(\boldsymbol{x}',t')$, and consequent re-excitation of an electron (destruction of the hole). We can examine these properties through some simple examples.

Let us first consider the case of a single free electron (for which $G^{\mathrm{a}} = 0$). Here, we will take $\boldsymbol{x}' = 0$, $t' = 0$, and will note that the spin is conserved, so this index can be ignored. Then, we can write the retarded Green's function, using (9.36), (9.37), and (9.71), as

$$\begin{aligned}
\mathrm{i}G_{\mathrm{r}}(\boldsymbol{x},t) &= \langle\{\Psi_{\sigma}(\boldsymbol{x},t),\Psi^{+}_{\sigma'}(\boldsymbol{x}',t')\}\rangle \\
&= \left\langle\left\{\left(\sum_{\boldsymbol{k}\sigma} c_{\boldsymbol{k}\sigma}\phi_{\boldsymbol{k}}(\boldsymbol{x},t)\right),\left(\sum_{\boldsymbol{k}'\sigma'} c^{+}_{\boldsymbol{k}'\sigma'}\phi^{*}_{\boldsymbol{k}'}(0,0)\right)\right\}\right\rangle \\
&= \frac{1}{V}\sum_{\boldsymbol{k}\sigma\boldsymbol{k}'\sigma'} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}-\mathrm{i}\omega_{k}t}\langle\{c_{\boldsymbol{k}\sigma},c^{+}_{\boldsymbol{k}'\sigma'}\}\rangle \\
&= \frac{1}{V}\sum_{\boldsymbol{k}\sigma\boldsymbol{k}'\sigma'} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}-\mathrm{i}\omega_{k}t}\delta_{\boldsymbol{k}\boldsymbol{k}'}\delta_{\sigma\sigma'} \\
&= \frac{1}{V}\sum_{\boldsymbol{k}\sigma} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}-\mathrm{i}\omega_{k}t}
\end{aligned} \tag{9.81}$$

where curly brackets denote the anti-commutator, and the second term in the anti-commutator vanishes as we require the state $\boldsymbol{k}'$ to be empty, and

$$\omega_k = \frac{\mathcal{E}(k)}{\hbar} \tag{9.82}$$

is the frequency (or energy) of the state and this term re-introduces the time variation. The form of (9.81) is highly suggestive of a Fourier transform, and we can actually write this transform as

$$G_{\mathrm{r}}(\boldsymbol{x}, t) = \frac{1}{V} \sum_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} G_{\mathrm{r}}(\boldsymbol{k}, t) \tag{9.83}$$

for which

$$G_{\mathrm{r}}(\boldsymbol{k}, t) = -\mathrm{i}\mathrm{e}^{-\mathrm{i}\omega_{\boldsymbol{k}} t - \gamma t} \qquad t > 0 \tag{9.84}$$

where $\gamma$ is a convergence factor (that will be set to zero after any calculation). This can be further Fourier transformed in frequency to give

$$G_{\mathrm{r}}(\boldsymbol{k}, \omega) = \frac{1}{\omega - \omega_{\boldsymbol{k}} + \mathrm{i}\gamma}. \tag{9.85}$$

While the particular form that we have assumed begins with the vacuum state and a single electron, it should not be surprising that the same result is obtained for the Fermi ground state, provided that the initial electron is created in an empty state. For this case, the advanced function no longer vanishes, but the double Fourier transform for the advanced function differs from (9.85) only in a change of the sign of the convergence factor (moving the closure of the inverse transform integral from the lower complex frequency half-plane to the upper half-plane). The same result for the retarded function could have been obtained, if we had defined $G_{\mathrm{r}}(\boldsymbol{k}, t)$ in terms of the creation and annihilation operators in the interaction representation as

$$G_{\mathrm{r}}(\boldsymbol{k}, t) = -\mathrm{i}\langle \mathcal{F} | \{ c_{\boldsymbol{k}}(t), c_{\boldsymbol{k}}^{+}(0) \} | \mathcal{F} \rangle. \tag{9.86}$$

The Green's function provides a fundamental difference from classical statistics, and the most important aspect of this is given by the *spectral density*. In classical mechanics, the energy is given by a specific function of the momentum; for example, in parabolic energy bands,

$$\hbar\omega = \mathcal{E} = \hbar\omega_{\boldsymbol{k}} = \frac{\hbar^2 k^2}{2m} \tag{9.87}$$

as discussed above. Here, the last term is the $\mathcal{E}(k)$ used above. Quantum mechanically, however, this is no longer the case. Now, the energy $\omega$ and momentum $\boldsymbol{k}$ are separate dynamic variables, which are not given by a simple relationship. In fact, if we take the imaginary part of (9.85), we find that

$$\operatorname{Im} G_{\mathrm{r}}(\boldsymbol{k}, \omega) = -\frac{\gamma}{(\omega - \omega_{\boldsymbol{k}}) + \gamma^2}. \tag{9.88}$$

In the limit of $\gamma \to 0$, we can recover a delta function between the energy and $\mathcal{E}(k)$. However, the spectral density $A(\omega, \boldsymbol{k})$ provides the quantum relationship between the energy $\hbar\omega$ and the momentum function $\mathcal{E}(k)$ through

$$A(\omega, \boldsymbol{k}) = -2\,\mathrm{Im}\,G_{\mathrm{r}}(\omega, \boldsymbol{k}) \tag{9.89$a$}$$

with

$$\int_{-\infty}^{\infty} A(\omega, \boldsymbol{k})\,\mathrm{d}\omega = 1. \tag{9.89$b$}$$

The last expression provides the conservation of area of the spectral density, by providing that the integral is the same as integrating over $\delta(\omega - \mathcal{E}(k)/\hbar)$, the classical form.

Before proceeding, it is worthwhile to consider just why we would want to use Green's functions. The answer to this lies in their use in many other fields of mathematics and physics (including engineering). In electromagnetics, we use the response to (the fields generated by) a single point charge by summing this response over the entire charge distribution. The response to the single point charge is the Green's function for the wave equation that is obtained from Maxwell's equation (whether in the the electrostatic or dynamic limit, depending upon the exact response used for the single point charge). In circuit theory, it is usual to use the impulse response for a circuit to generate the response for a complicated input function. Again, the impulse reponse is the Green's function for the differential equation describing the circuit response. In each of these two examples, the Green's function is used to build up the linear response through the use of the superposition principle. The purpose of Green's functions in quantum mechanics is precisely the same. We use them where they are of benefit to find the more complicated solutions more easily. We have written them in terms of field operators for a similarly easier understanding. The field operators give a wave-function-like form to the operators, while the operators provide us with a simple algebra for analysing complicated interaction terms. The field operators have a series expansion in terms of the basis functions providing the Hilbert space (our coordinate system). Thus, while the approach seems more complicated than solving the Schrödinger equation directly, the result provides an easier methodology for analysing the complicated situations that can arise in real physical systems. Of course, this is true only when the work of solving the Schrödinger equation in the real system would be more than we have expended in setting up our Green's functions for their own ease of use.

### 9.6.1 The equations of motion

To proceed to discuss the manner in which the Green's functions can be used to solve the many-electron problem, we need to obtain the appropriate equations of motion for these functions. Certainly, this can be done directly from the Hamiltonian (9.70), but we need to be careful with the two-electron interaction term $H_2$. If we were to drop the integrations and one of the field operators, $\Psi^+_{\sigma'}(\boldsymbol{x}', t')$, the remainder would be the time-independent terms of the Schrödinger equation for $\Psi(\boldsymbol{x}, t)$, except that the integration within the interaction term is over the variables $\boldsymbol{x}'$ and $\sigma'$ (a summation in the latter case). The remaining terms are related to the interaction of the charge density with the field operator of interest. To expedite achieving an understanding, we will repeat the trick used in the previous chapter; that is, we will insert a $\delta(\boldsymbol{x}' - \boldsymbol{x})$ with an integration over the first variable (using a double prime for the previous primed set of variables). Then ignoring the integrations over $\boldsymbol{x}$ and $\boldsymbol{x}'$ (and their consequent spin summations), we can write the time-independent terms as (again, for the homogeneous case we will neglect the potential terms)

$$\Psi^+(\boldsymbol{x}', t')H\Psi(\boldsymbol{x}, t) = \left[ -\frac{\hbar^2}{2m}\Psi^+(\boldsymbol{x}', t')\nabla^2\Psi(\boldsymbol{x}, t) + \int \Psi^+(\boldsymbol{x}', t')\Psi^+(\boldsymbol{x}'', t) \right.$$
$$\left. \times \frac{e^2}{4\pi\varepsilon|\boldsymbol{x} - \boldsymbol{x}''|}\Psi(\boldsymbol{x}'', t)\Psi(\boldsymbol{x}, t)\right] \mathrm{d}\boldsymbol{x}''. \tag{9.90}$$

If we now construct the ensemble average by taking the inner product with the Fermi ground state $|\mathcal{F}\rangle$, and sum over the spin indices (noting that the spin states are orthogonal), as above, we can construct the Green's function equation. There are two problems that must be faced. The first is in replacing the terms on the left-hand side of (9.90) with the time derivative of the Green's function, and the second in that the averaging of the interaction term produces a two-particle Green's function involving four field operators instead of two. Approximations must be made to this last term. Let us first deal with the time-varying term.

The left-hand side of (9.90) is the term that is normally set equal to the time variation of the wave function and then averaged as follows:

$$\langle \Psi^+(\boldsymbol{x}', t')H\Psi(\boldsymbol{x}, t)\rangle = \mathrm{i}\hbar \left\langle \Psi^+(\boldsymbol{x}', t')\frac{\partial}{\partial t}\Psi(\boldsymbol{x}, t)\right\rangle. \tag{9.91}$$

However, the term on the right-hand side is not the time derivative of the Green's function because of the discontinuity at $t = t'$ that is given by the definition (9.79). To see this, we will take the derivative of (9.79):

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{i}G(\boldsymbol{x}, t, \boldsymbol{x}', t') &= \frac{\partial}{\partial t}\{\Theta(t - t')\langle\Psi_\sigma(\boldsymbol{x}, t)\Psi^+_{\sigma'}(\boldsymbol{x}', t')\rangle \\ &\quad - \Theta(t' - t)\langle\Psi^+_{\sigma'}(\boldsymbol{x}', t')\Psi_\sigma(\boldsymbol{x}, t)\rangle\} \\ &= \left\langle T\left[\frac{\partial}{\partial t}\Psi_\sigma(\boldsymbol{x}, t)\Psi^+_{\sigma'}(\boldsymbol{x}', t')\right]\right\rangle \\ &\quad + \delta(t - t')\,\langle\Psi_\sigma(\boldsymbol{x}, t)\Psi^+_{\sigma'}(\boldsymbol{x}', t')\rangle \\ &\quad + \delta(t - t')\,\langle\Psi^+_{\sigma'}(\boldsymbol{x}', t')\Psi_\sigma(\boldsymbol{x}, t)\rangle \\ &= \left\langle T\left[\Psi^+_{\sigma'}(\boldsymbol{x}', t')\frac{\partial}{\partial t}\Psi_\sigma(\boldsymbol{x}, t)\right]\right\rangle + \delta_{\sigma\sigma'}\delta(t - t')\,\delta(\boldsymbol{x} - \boldsymbol{x}').\end{aligned} \tag{9.92}$$

We recognize that the first term on the right-hand side of this last equation is the term we need in (9.92). Thus, we can now insert these results into (9.91), and write the equation for the Green's function as

$$\mathrm{i}\hbar\frac{\partial}{\partial t}G(\boldsymbol{x},t,\boldsymbol{x}',t') = \hbar\,\delta(t-t')\delta(\boldsymbol{x}-\boldsymbol{x}')\delta_{\sigma\sigma'} - \frac{\hbar^2}{2m}\nabla_x^2 G(\boldsymbol{x},t,\boldsymbol{x}',t') - \mathrm{i}\int \frac{e^2}{4\pi\varepsilon|\boldsymbol{x}-\boldsymbol{x}''|}G(\boldsymbol{x}'',t,\boldsymbol{x}'',t^-,\boldsymbol{x},t,\boldsymbol{x}',t')\,\mathrm{d}\boldsymbol{x}'' \tag{9.93}$$

where the minus sign of the interaction term arises from re-ordering the field operators to fit the definition of the two-particle Green's function, which is given by:

$$G(\boldsymbol{x}_1,t_1,\boldsymbol{x}_2,t_2,\boldsymbol{x}_3,t_3,\boldsymbol{x}_4,t_4) \equiv -\langle T[\Psi^+(\boldsymbol{x}_1,t_1)\Psi(\boldsymbol{x}_2,t_2)\Psi(\boldsymbol{x}_3,t_3)\Psi^+(\boldsymbol{x}_4,t_4)]\rangle \tag{9.94}$$

and the sign arises from the ubiquitous set of factors of i necessary to match to single-particle Green's functions ($\mathrm{i}^2 = -1$). The choice of the time factor for the second set of variables in the interaction term in (9.93) ensures that the interaction occurs prior to the time of measurement $t$. We now turn to approximations for this latter term.

### 9.6.2 The Hartree approximation

The general problem of the two-particle Green's function that arises in (9.90) and (9.94) is that it is part of an infinite hierarchy. For example, if we decided to develop an equation for the two-particle Green's function, the interaction term would couple it to a three-particle Green's function, and so on. It is this interaction term that 'fouls up' any possible separation of the Hamiltonian into a sum of single-electron Hamiltonians, which would allow us to solve the one-electron problem and use the Slater determinant. We cannot do this if the interaction term is present. This is because any projection from the many-particle Green's function onto an equivalent equation for the single-particle function always involves some averaging. The interaction terms always couple any reduced function to the higher-order functions and provide a correlation of the typical one-particle motion with that of all other particles. Thus, if we are to solve for a simple equation of motion for the one-particle Green's function, it is necessary to terminate this set of coupled equations at some order. This termination is made by approximating the two-particle Green's function as a product of single-particle functions. The simplest approximation is to write this product as (the sign is that appropriate for fermions)

$$G(\boldsymbol{x}_1,t_1,\boldsymbol{x}_2,t_2,\boldsymbol{x}_3,t_3,\boldsymbol{x}_4,t_4) = G(\boldsymbol{x}_2,t_2,\boldsymbol{x}_1,t_1)G(\boldsymbol{x}_3,t_3,\boldsymbol{x}_4,t_4). \tag{9.95}$$

This simple approximation is termed the Hartree approximation. We will see that this approximation completely uncouples the interaction term and allows us to write a single simple equation of motion for the Green's function. Indeed, using this approximation in the two-particle Green's function in (9.94) leads to one of the Green's functions being

$$G(\boldsymbol{x}_1, t_1, \boldsymbol{x}_2, t_2) = G(\boldsymbol{x}'', t^-, \boldsymbol{x}'', t) = \mathrm{i}\langle \Psi^+(\boldsymbol{x}'', t)\Psi(\boldsymbol{x}'', t)\rangle = \mathrm{i}\rho(\boldsymbol{x}'', t) \tag{9.96}$$

which is the particle density of the electron gas. This now leads to the following equation for the Green's function:

$$\begin{aligned}\mathrm{i}\,\hbar\frac{\partial}{\partial t}G(\boldsymbol{x}, t, \boldsymbol{x}', t') = \left[-\frac{\hbar^2}{2m}\nabla^2_{\boldsymbol{x}} + \int \frac{\rho(\boldsymbol{x}'', t)e^2}{4\pi\varepsilon|\boldsymbol{x} - \boldsymbol{x}''|}\,\mathrm{d}\boldsymbol{x}''\right] G(\boldsymbol{x}, t, \boldsymbol{x}', t')\\ + \hbar\delta(t - t')\delta(\boldsymbol{x} - \boldsymbol{x}')\delta_{\sigma\sigma'}.\end{aligned} \tag{9.97}$$

The second term in the square brackets is often called the Hartree energy. If we Fourier transform in space and time for our plane-wave homogeneous states, the Green's function can be solved (with the convergence factor) as

$$G(\boldsymbol{k}, \omega) = \frac{1}{\omega - \omega_{\boldsymbol{k}} + V_{\mathrm{H}}(\boldsymbol{k})/\hbar \pm \mathrm{i}\gamma}\delta_{\sigma\sigma'} \tag{9.98}$$

where $V_{\mathrm{H}}(\boldsymbol{k})$ is the Fourier transform of the Hartree energy, and we have assumed that $\boldsymbol{x}' = \boldsymbol{0}$, $t' = 0$, and the upper sign is for the retarded Green's function while the lower sign is for the advanced function. This should be compared with (9.85) for the one-electron case. The correction that the interaction term has introduced is to 'renormalize' the energy with the Hartree energy.

The form of the Green's function has been seen previously, in particular as (7.51) for the decay of the initial state in perturbation theory. There, the extra potential, represented here as the Hartree potential, was termed the self-energy correction to the initial state. Indeed, that is the normal terminology. In the present case, the Hartree potential is usually real so the only effect is a self-energy shift of the initial energy represented by $\omega_{\boldsymbol{k}} = \hbar k^2/2m$. Note that since the Hartree potential has the sign opposite to that of $\omega_{\boldsymbol{k}}$, it serves to lower the energy of the state. In the next subsection, this connection with perturbation theory will be explored further.

Another approximation to the two-particle Green's function is given by the so-called *Hartree–Fock approximation*, where the two-particle function is defined in terms of one-particle functions according to

$$\begin{aligned}G(\boldsymbol{x}_1, t_1, \boldsymbol{x}_2, t_2, \boldsymbol{x}_3, t_3, \boldsymbol{x}_4, t_4) = G(\boldsymbol{x}_2, t_2, \boldsymbol{x}_1, t_1)G(\boldsymbol{x}_3, t_3, \boldsymbol{x}_4, t_4)\\ - G(\boldsymbol{x}_3, t_3, \boldsymbol{x}_1, t_1)G(\boldsymbol{x}_2, t_2, \boldsymbol{x}_4, t_4).\end{aligned} \tag{9.99}$$

This leads, for the interaction term in (9.93), to

$$\begin{aligned}G(\boldsymbol{x}'', t, \boldsymbol{x}'', t^-, \boldsymbol{x}, t, \boldsymbol{x}', t') = G(\boldsymbol{x}'', t^-, \boldsymbol{x}'', t)G(\boldsymbol{x}, t, \boldsymbol{x}', t')\\ - G(\boldsymbol{x}, t, \boldsymbol{x}'', t)G(\boldsymbol{x}'', t^-, \boldsymbol{x}', t').\end{aligned} \tag{9.100}$$

Whereas the Hartree approximation represented an interaction between the background density and the Green's function test particle that occurred at $\boldsymbol{x}, t$, the second term in (9.100) represents a distributed interaction. This cannot be handled so simply, and leads to a more complicated interaction. In fact, this second term is called the *exchange energy*, and arises from the interchange of positions $\boldsymbol{x}''$ and $\boldsymbol{x}$ (interchange of $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$) in the two-particle Green's function. In the Hartree energy, one of the Green's functions reduced to the simple density and a solvable equation was obtained. This is not the case for the exchange energy. Even though the integral over $\boldsymbol{x}''$ can be separated by Fourier transformation (it becomes a convolution integral, so the transform is simpler), the result still involves a product of two Green's functions. Thus, no simple equation for the desired Green's function is obtained, and some form of approximation approach must be used. Here, we now turn to perturbation theory.

### 9.6.3 Connection with perturbation theory

In perturbation theory, it was supposed that we could work with the known solutions to a particular Hamiltonian and treat the more complicated terms as a perturbation that modified these known solutions. While perturbation theory can be used for any small interaction term, we shall pursue it here for the particular case of the electron–electron interaction, which is the process that we have referred to as the interaction term in this section. It could as easily be used for the electron–phonon interaction.

In the present case, we know that in the absence of the interaction term, the solutions would be normal plane waves for the homogeneous electron gas. The Green's function that we are interested in is just

$$\begin{aligned} \mathrm{i}G_{\sigma\sigma'}(\boldsymbol{x}, t; \boldsymbol{x}', t') = {} & \Theta(t - t')\langle\Psi_\sigma(\boldsymbol{x}, t)\Psi^+_{\sigma'}(\boldsymbol{x}', t')\rangle \\ & - \Theta(t' - t)\langle\Psi^+_{\sigma'}(\boldsymbol{x}', t')\Psi_\sigma(\boldsymbol{x}, t)\rangle. \end{aligned} \tag{9.101}$$

The ensemble averages are evaluated over the Fermi gas at some particular time. However, when we revert to perturbation theory, we must be somewhat careful about this time of evaluation. The Green's function is evaluated from $t'$ to $t$, which means that the perturbation has already been initiated into the Fermi gas prior to the onset of the time for the Green's function. This process can be incorporated by noting that the perturbing interaction can be treated using the interaction representation, in which the effect is governed by the unitary operator defined in (7.40*d*):

$$U(t, t_0) = \exp\left[-\frac{\mathrm{i}}{\hbar}\int_{t_0}^{t} V(t')\,\mathrm{d}t'\right] \tag{9.102}$$

where here the perturbing potential is

$$V(t) = \int \mathrm{d}\boldsymbol{x}_1 \int \mathrm{d}\boldsymbol{x}_2\, \Psi^+(\boldsymbol{x}_1, t)\Psi^+(\boldsymbol{x}_2, t)\frac{e^2}{4\pi\varepsilon|\boldsymbol{x}_1 - \boldsymbol{x}_2|}\Psi(\boldsymbol{x}_2, t)\Psi(\boldsymbol{x}_1, t). \tag{9.103}$$

Thus, the Fermi gas state can be described as being derived from the zero-order state in which there is no interaction, through

$$|\mathcal{F}\rangle = U(0,-\infty)|\mathcal{F}_0\rangle \tag{9.104}$$

where the time scale has been set by that desired for the Green's function (9.101). Similarly, the adjoint state is

$$\langle F| = \langle \mathcal{F}_0|U(\infty,0). \tag{9.105}$$

In the following, we will only use the retarded function for which $t > t\,'$, although the method can easily be extended to the general case. We can now write the general Green's function in terms of the interaction propagators as (we have re-introduced the time-ordering operator in the final result)

$$\begin{aligned} \mathrm{i}G(\boldsymbol{x},t;\boldsymbol{x}',t') &= \langle \mathcal{F}_0|U(\infty,0)U^+(t,0)\Psi(\boldsymbol{x},t)U(t,0)U^+(t',0) \\ &\quad \times \Psi^+(\boldsymbol{x}',t')U(t',0)U(0,-\infty)|\mathcal{F}_0\rangle \\ &= \langle \mathcal{F}_0|T[U(\infty,-\infty)\Psi(\boldsymbol{x})\Psi^+(\boldsymbol{x}')]|\mathcal{F}_0\rangle. \end{aligned} \tag{9.106}$$

This suggests that we can call

$$\mathrm{i}G^0(\boldsymbol{x},t;\boldsymbol{x}',t') = \langle \mathcal{F}_0|T[\Psi(\boldsymbol{x},t)\Psi^+(\boldsymbol{x}',t')]|\mathcal{F}_0\rangle \tag{9.107}$$

the unperturbed Green's function, or the Green's function for a single (free) electron. The solutions for these two (the retarded and the advanced functions) are given by (9.85) and (9.86), with the proper sign of the convergence factor, for the Fourier-transformed version.

Let us consider the Fourier-transformed version (9.86), which is a function of the momentum $\boldsymbol{k}$ and the time $t$. We note that this version suggests that the Green's function is properly a function only of the difference $\boldsymbol{x}-\boldsymbol{x}\,'$ and $t-t'$, and this is a general property of a homogeneous system, particularly in equilibrium. This is an important point, but we will return to using the position representation. The Green's function that we must evaluate is then one in which a number of perturbation terms appear, and we can write the first-order term, in the position representation, as

$$\begin{aligned} G(\boldsymbol{x},t;\boldsymbol{x}',t') &= G^0(\boldsymbol{x},t;\boldsymbol{x}',t') \\ &\quad - \frac{1}{\hbar}\int_{-\infty}^{\infty} \langle \mathcal{F}|T[V(t_1)\Psi(\boldsymbol{x},t)\Psi^+(\boldsymbol{x}',t')]|\mathcal{F}\rangle \,\mathrm{d}t_1 \end{aligned} \tag{9.108}$$

where, the interaction potential is just (9.103). In this case, the integration over $t_1$ ensures conservation of energy in the interaction process, but care needs to be exercised. The problem is that there are now six operators in (9.108), three creation operators and three annihilation operators. How are we to organize these operators into Green's functions?

It is a relatively arduous task to organize the operators in (9.108) as there are many possible pairings that can be made. By pairings, we mean the putting together of a creation operator and an annihilation operator in a form that will lead to a Green's function of some type. For example, we can bring together $\Psi(\boldsymbol{x}, t)$ and $\Psi^+(\boldsymbol{x}_2, t)$ to form a Green's function. The second of these two terms puts an electron into a state, while the first operator removes the electron from this state at a later time, so if this is to be a Green's function in the sense described above, the electron must be taken from the state into which it is initially placed. This is also true for the advanced function in which the particle is first removed and then replaced at a later time. Thus, the various pairings are actually the various ways in which three electrons can be created and three electrons annihilated by the six operators, in any order. There are a set of rules for the breaking of the operators into pairs that taken together are termed *Wick's theorem*.

The first rule is that in making the pairings, each pair of operators must be put into their time-ordered positioning, but with the creation operator to the right as is done in the unperturbed Green's function. One must keep track of the sign changes that must be incorporated for the various exchanges of operators. The second rule says that the time ordering of different excitation terms should be maintained. The third rule states that when the two paired operators operate at the 'same' time, the creation operator should be placed to the right. The time-ordered product of interest now is

$$\begin{aligned}
T[\,] &= T[\Psi^+(\boldsymbol{x}_1, t_1)\Psi^+(\boldsymbol{x}_2, t_1)\Psi(\boldsymbol{x}_2, t_1)\Psi(\boldsymbol{x}_1, t_1)\Psi(\boldsymbol{x}, t)\Psi^+(\boldsymbol{x}', t')] \\
&= T[\Psi(\boldsymbol{x}_2, t_1)\Psi^+(\boldsymbol{x}_1, t_1)]T[\Psi(\boldsymbol{x}, t)\Psi^+(\boldsymbol{x}_2, t_1)]T[\Psi(\boldsymbol{x}_1, t_1)\Psi^+(\boldsymbol{x}', t')] \\
&\quad - T[\Psi(\boldsymbol{x}_2, t_1)\Psi^+(\boldsymbol{x}_1, t_1)]T[\Psi(\boldsymbol{x}_1, t_1)\Psi^+(\boldsymbol{x}_2, t_1)]T[\Psi(\boldsymbol{x}, t)\Psi^+(\boldsymbol{x}', t')] \\
&\quad + T[\Psi(\boldsymbol{x}_1, t_1)\Psi^+(\boldsymbol{x}_1, t_1)]T[\Psi(\boldsymbol{x}_2, t_1)\Psi^+(\boldsymbol{x}_2, t_1)]T[\Psi(\boldsymbol{x}, t)\Psi^+(\boldsymbol{x}', t')] \\
&\quad - T[\Psi(\boldsymbol{x}_1, t_1)\Psi^+(\boldsymbol{x}_1, t_1)]T[\Psi(\boldsymbol{x}, t)\Psi^+(\boldsymbol{x}_2, t_1)]T[\Psi(\boldsymbol{x}_2, t_1)\Psi^+(\boldsymbol{x}', t')] \\
&\quad - T[\Psi(\boldsymbol{x}, t)\Psi^+(\boldsymbol{x}_1, t_1)]T[\Psi(\boldsymbol{x}_2, t_1)\Psi^+(\boldsymbol{x}_2, t_1)]T[\Psi(\boldsymbol{x}_1, t_1)\Psi^+(\boldsymbol{x}', t')] \\
&\quad + T[\Psi(\boldsymbol{x}, t)\Psi^+(\boldsymbol{x}_1, t_1)]T[\Psi(\boldsymbol{x}_1, t_1)\Psi^+(\boldsymbol{x}_2, t_1)]T[\Psi(\boldsymbol{x}_2, t_1)\Psi^+(\boldsymbol{x}', t')].
\end{aligned} \tag{9.109}$$

These six terms are the six possible combinations of the three creation operators and three annihilation operators, with the ordering and signs satisfying the above rules. This now leads to the first-order correction to the Green's function (we must remember to incorporate the needed factors of i in going from each time-ordered product to the Green's function):

$$\begin{aligned}
\delta G = &-\frac{\mathrm{i}}{\hbar}\int_{-\infty}^{\infty} \mathrm{d}t_1 \int \mathrm{d}\boldsymbol{x}_1 \int \mathrm{d}\boldsymbol{x}_2 \, \frac{e^2}{8\pi\varepsilon|\boldsymbol{x}_1 - \boldsymbol{x}_2|} \\
&\times \Big[ G^0(\boldsymbol{x}_2, t_1; \boldsymbol{x}_1, t_1) G^0(\boldsymbol{x}, t; \boldsymbol{x}_2, t_1) G^0(\boldsymbol{x}_1, t_1; \boldsymbol{x}', t') \\
&- G^0(\boldsymbol{x}_2, t_1; \boldsymbol{x}_1, t_1) G^0(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_1) G^0(\boldsymbol{x}, t; \boldsymbol{x}', t') \\
&+ G^0(\boldsymbol{x}_1, t_1; \boldsymbol{x}_1, t_1) G^0(\boldsymbol{x}_2, t_1; \boldsymbol{x}_2, t_1) G^0(\boldsymbol{x}, t; \boldsymbol{x}', t')
\end{aligned}$$

$$
\begin{aligned}
&- G^0(\boldsymbol{x}_1,t_1;\boldsymbol{x}_1,t_1)G^0(\boldsymbol{x},t;\boldsymbol{x}_2,t_1)G^0(\boldsymbol{x}_2,t_1;\boldsymbol{x}',t')\\
&- G^0(\boldsymbol{x},t;\boldsymbol{x}_1,t_1)G^0(\boldsymbol{x}_2,t_1;\boldsymbol{x}_2,t_1)G^0(\boldsymbol{x}_1,t_1;\boldsymbol{x}',t')\\
&+ G^0(\boldsymbol{x},t;\boldsymbol{x}_1,t_1)G^0(\boldsymbol{x}_1,t_1;\boldsymbol{x}_2,t_1)G^0(\boldsymbol{x}_2,t_1;\boldsymbol{x}',t')\Big]. \qquad (9.110)
\end{aligned}
$$

We can understand these various terms by making a series of *diagrams*, called Feynman diagrams. Each of the six terms in the square brackets leads to one diagram. Each Green's function is represented by a solid line with an arrow pointing in the direction of the time flow (time is taken to flow from the bottom of a diagram to the top), and each end of a line segment, termed a vertex, is at one of the position–time values of the two arguments. The interaction (the Coulomb potential in this case) connects two vertices by a wavy (or zigzag) line adjoining the two vertices for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Six products in (9.110), together with the interaction term, are shown by parts (*a*)–(*f*) of figure 9.1, respectively. Each vertex is identified with its position and time coordinates.

We note that the diagrams (*a*) and (*f*) are identical except for the interchange of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Since these are dummy variables which will be integrated out of the final result, they may be interchanged, so these two terms may be combined with a pre-factor of 2. Similarly, diagrams (*d*) and (*e*) are the same except for the interchange of these two variables, and these two terms can similarly be combined with the resultant factor of 2. This suggests that the result should contain only those diagrams that truly differ in their topology. Diagrams (*b*) and (*c*), however, are more complicated. These *disconnected* diagrams will drop out of the final result, by consideration of the denominator. Above, in (9.78), we set $\langle\mathcal{F}|\mathcal{F}\rangle = 1$, but this will no longer be the case with the introduction of the pertubation from the unperturbed ground state, which we have neglected up to now. We now reconsider this denominator term.

We have assumed that the Fermi ground state was normalized, but in computing the expectation value (9.106) we switched to considering the unperturbed Fermi ground state. To be consistent we must do the same for the denominator which has, until now, been ignored as a normalized factor of unity. We could continue in this, except that the perturbation series imposed upon the numerator may distort the normalization, and we must treat the denominator to make sure that normalization is maintained. Thus, the denominator is

$$\langle\mathcal{F}|\mathcal{F}\,\rangle = \langle\mathcal{F}_0|T[U(\infty,-\infty)]|\mathcal{F}_0\rangle. \qquad (9.111)$$

Here, we can now expand the interaction term to yield the zero- and first-order terms as

$$
\begin{aligned}
\langle\mathcal{F}|\mathcal{F}\rangle = \langle\mathcal{F}_0|\mathcal{F}_0\rangle - \frac{\mathrm{i}}{\hbar}\int_t^{t_0}\mathrm{d}t'\int\mathrm{d}\boldsymbol{x}_1\int\mathrm{d}\boldsymbol{x}_2\,\frac{e^2}{8\pi\varepsilon|\boldsymbol{x}_1-\boldsymbol{x}_2|}&\\
\times T[\Psi^+(\boldsymbol{x}_1,t)\Psi^+(\boldsymbol{x}_2,t)\Psi(\boldsymbol{x}_2,t)\Psi(\boldsymbol{x}_1,t)].& \qquad (9.112)
\end{aligned}
$$

This time-ordered product has only four terms and can be decomposed somewhat

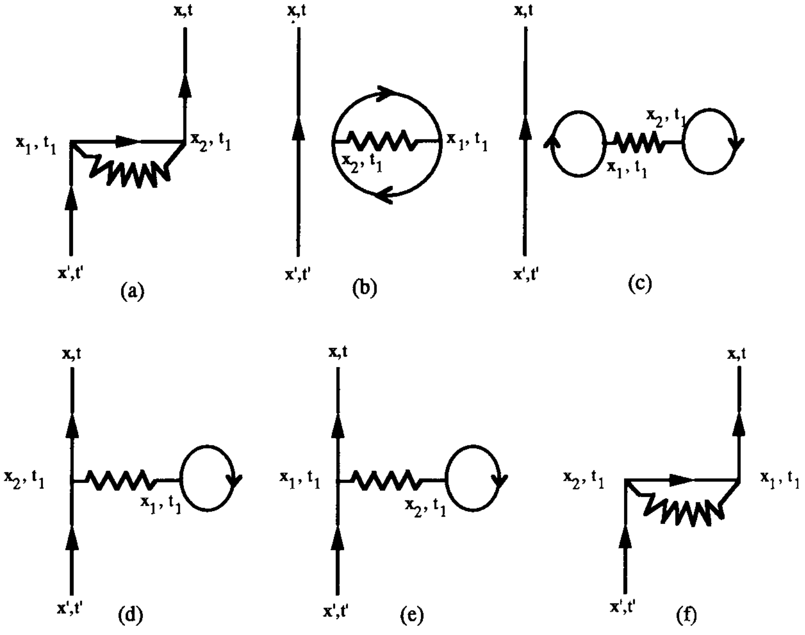


**Figure 9.1.** Diagrams for the triple products of unperturbed Green's functions appearing in (9.103).

more easily than the result above. This leads to the terms

$$\begin{aligned}\langle\mathcal{F}|\mathcal{F}\rangle = \langle\mathcal{F}_0|\mathcal{F}_0\rangle - \frac{\mathrm{i}}{\hbar}\int_t^{t_0}\mathrm{d}t'\int\mathrm{d}\boldsymbol{x}_1\int\mathrm{d}\boldsymbol{x}_2\,\frac{e^2}{8\pi\varepsilon|\boldsymbol{x}_1-\boldsymbol{x}_2|}\\ \times\Big[G^0(\boldsymbol{x}_2,t_1;\boldsymbol{x}_1,t_1)G^0(\boldsymbol{x}_1,t_1;\boldsymbol{x}_2,t_1)\\ - G^0(\boldsymbol{x}_1,t_1;\boldsymbol{x}_1,t_1)G^0(\boldsymbol{x}_2,t_1;\boldsymbol{x}_2,t_1)\Big].\end{aligned} \tag{9.113}$$

The three terms in this series (the leading unity factor plus the two products of Green's functions with their interaction term) are shown in figure 9.2 by diagrams (*a*) to (*c*) respectively.

The three terms of figure 9.2 are quite similar to diagrams (*b*) and (*c*) in figure 9.1. In fact, if we take the three diagrams of figure 9.2 and multiply them by the unperturbed Green's function, we arrive at the first term of the expansion for the numerator plus diagrams (*b*) and (*c*) of figure 9.1. This suggests that the denominator will cancel all such disconnected diagrams. While this is quite a leap of faith, it turns out to be the truth (Abrikosov *et al* 1963). This means that we need to treat only those factors of the Wick's theorem expansion that lead to connected diagrams. Thus, the denominator remains normalized and the

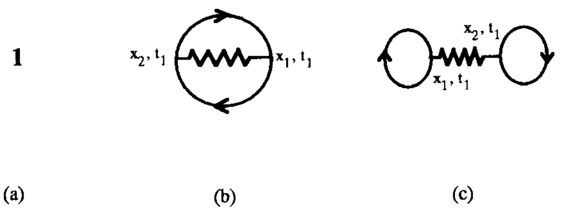


**Figure 9.2.** The first two terms (three factors) in the perturbation series for the normalization denominator.

numerator gives only two distinct terms in first-order perturbation theory. The expansion can be extended to higher order, and the number of distinct diagrams increases accordingly. For example, in second order, we expect the perturbing potential to appear twice, which leads to ten operators in the time-ordered product, four internal spatial integrations and two internal time integrations. These lead to diagrams with five unperturbed Green's functions and two interaction branches. In general, for the $r$th order of perturbation, there are $r$ interaction branches and $2r + 1$ unperturbed Green's function lines. The proof of this is beyond the treatment desired here, but one can immediately extend from the results so far obtained to see that this will be the case.

### 9.6.4 Dyson's equation

The final result of the expansion above can be combined into a simpler equation for the Green's function as follows:

$$\begin{aligned} G(\boldsymbol{x}, t; \boldsymbol{x}', t') &= G^0(\boldsymbol{x}, t; \boldsymbol{x}', t') \\ &\quad + \int \mathrm{d}^4 y_1 \int \mathrm{d}^4 y_2 \, G^0(\boldsymbol{x}, t; y_1) \Sigma(y_1, y_2) G^0(y_2; \boldsymbol{x}', t') \end{aligned} \tag{9.114}$$

where the short-hand notation $y_i = (\boldsymbol{x}_i, t_i)$ has been used, and we have introduced the *self-energy* $\Sigma(y_1, y_2)$. For the first-order expansion above, the self-energy may be written as

$$\begin{aligned} \Sigma(y_1, y_2) = \delta(t_1 - t_2) - \frac{\mathrm{i}}{\hbar} \biggl[ & \frac{e^2}{4\pi\varepsilon |\boldsymbol{x}_1 - \boldsymbol{x}_2|} G^0(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_1) \\ & - \int \mathrm{d}\boldsymbol{x}_3 \, \frac{e^2}{4\pi\varepsilon |\boldsymbol{x}_1 - \boldsymbol{x}_3|} G^0(\boldsymbol{x}_3, t_1; \boldsymbol{x}_3, t_1) \delta(\boldsymbol{x}_1 - \boldsymbol{x}_2) \biggr]. \end{aligned} \tag{9.115}$$

The second term is simply the Hartree energy discussed previously. The Green's function in this term is just the particle density, and the integration over the density produces the averaged interaction that we associated with the Hartree potential, although the formulation is different here. On the other hand, the first term is the leading term in the exchange energy since it involves the interaction between the electron at different points at the same time.

The integral in (9.114) is a multiple convolution integral. We want to examine the structure of this equation for a moment, and for this purpose will suppress the integrations and the internal variables, but recalling that all internal variables must be integrated in keeping with the concept of the convolution product. Thus, we can rewrite (9.114) in this reduced notation as

$$G = G^0 + G^0 \Sigma G^0. \tag{9.116}$$

Now, consider the product

$$G = G^0 + G^0 \Sigma G \tag{9.117}$$

in which the unperturbed Green's function at the end of the last term is replaced by the full interacting Green's function, calculated to all orders of perturbation theory. Let us follow the spirit of perturbation theory, discussed in chapter 7, where we first assume that the lowest-order approximation to $G$ is found by taking only the first term in (9.117), so that

$$G_0 = G^0. \tag{9.118}$$

The next approximation, the first-order approximation, is found by using this last result for $G$ on the right-hand side of (9.117):

$$G_1 = G^0 + G^0 \Sigma G^0. \tag{9.119}$$

This is just our result (9.116) obtained the hard way. If we continue, the next order in the approximate expansion is

$$G_2 = G^0 + G^0 \Sigma G^0 + G^0 \Sigma G^0 \Sigma G^0. \tag{9.120}$$

In fact, this argument is to convince ourselves that the form (9.117) is the properly re-summed perturbation expansion. The first intuitive step is to assume that all higher-order terms in the perturbation expansion will go into the change from $G^{\,0}$ to $G$, but this would be wrong. Certain of the terms in fact do contribute to this change. On the other hand, in particular, terms that expand the 'bubble' on the end of the interaction line (the combination of the interaction line and the bubble are often referred to as a 'tadpole diagram' for reasons that are too obvious to explain) in figure 9.1(*d*) or that expand the interaction line in figure 9.1(*a*) actually contribute to an expansion for the self-energy. In addition, when the problem is worked through self-consistently, the electrons can modify the strength of the interaction, which leads to a correction called a *vertex correction*. Most of this is just too complicated to worry about at this point, and we will limit ourselves to the corrections that arise in the lowest order of perturbation theory. However, we can immediately jump from (9.114) to the re-summed equation for the interacting Green's function

$$G(\boldsymbol{x}, t; \boldsymbol{x}', t') = G^0(\boldsymbol{x}, t; \boldsymbol{x}', t') + \int \mathrm{d}^4 y_1 \int \mathrm{d}^4 y_2 \, G^0(\boldsymbol{x}, t; y_1)\Sigma(y_1, y_2)G(y_2; \boldsymbol{x}', t'). \tag{9.121}$$

This is still a complicated integral equation for the Green's function. However, because this is a convolution product, it is possible to rid ourselves of the integrations by Fourier transforming in both space and time, using the results that the self-energy acts at a single time and that the Green's functions are functions only of the differences in the arguments. This result is just

$$G(\boldsymbol{k}, \omega) = G^0(\boldsymbol{k}, \omega) + G^0(\boldsymbol{k}, \omega)\Sigma(\boldsymbol{k})G(\boldsymbol{k}, \omega) \tag{9.122}$$

for which

$$G(\boldsymbol{k}, \omega) = \frac{1}{1/G^0(\boldsymbol{k}, \omega) - \Sigma(\boldsymbol{k})} = \frac{1}{\omega - \omega_{\boldsymbol{k}} - \Sigma(\boldsymbol{k}) \pm \mathrm{i}\gamma} \tag{9.123}$$

where we have used (9.85). This is, of course, a result quite similar to that obtained in the Hartree approximation, except that now the self-energy is a more complicated function. Moreover, the self-energy usually has an imaginary part and so the convergence factor can be ignored; for the retarded Green's function, we use the retarded self-energy with its negative imaginary part to produce the proper positive imaginary term in the denominator. The converse is taken for the advanced functions.

Equation (9.122) is termed Dyson's equation, and it should be compared with the equivalent formulation found in (7.85) and (7.86) (the latter differ by a factor of the reduced Planck constant). This equation is generally considered to be the basic starting point of any treatment of Green's functions. What we have done here is to develop just the starting point for these calculations. From this point, real results incorporate a number of further approximations, necessary because the interacting electron system, with its vast number of contributing particles in any real system, is a complicated beast. Only the beginnings have been explored but, with this opening to the unsolved, we are brought to a suitable point to stop. Before we do this, however, we need to finish our discussion of the self-energy in its lowest order.

### 9.6.5 The self-energy

Before completing this chapter, we want to examine in some more detail the self-energy that arises for the Coulomb interaction between electrons, at least to the lowest order encompassed in (9.114) and (9.115). For this purpose, we will want to work in the momentum space description of the Green's functions and the self-energy. We define the transforms as

$$G(\boldsymbol{x}-\boldsymbol{x}', t-t') = \int \frac{\mathrm{d}\boldsymbol{k}}{(2\pi)^3} \int \frac{\mathrm{d}\omega}{2\pi} G(\boldsymbol{k}, \omega) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')+\mathrm{i}\omega(t-t')} \tag{9.124}$$

and similarly for the self-energy and interaction terms (which form part of the self-energy). We will take the two terms that make up (9.114) and (9.115) separately. Let us first consider the leading term, which is composed of diagrams (*a*) and (*f*) of figure 9.1, as

$$\begin{aligned} G_1 = & \int \mathrm{d}^4 y_1 \int \mathrm{d}^4 y_2 \int \frac{\mathrm{d}\boldsymbol{k}}{(2\pi)^3} \int \frac{\mathrm{d}\omega}{2\pi} G^0(k, \omega) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y}_1)+\mathrm{i}\omega(t-t_1)} \\ & \times \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} \int \frac{\mathrm{d}\omega'}{2\pi} G(\boldsymbol{k}', \omega') \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot(\boldsymbol{y}_2-\boldsymbol{x}')+\mathrm{i}\omega'(t_1-t')} \delta(t_1-t_2) \\ & \times \frac{\mathrm{i}}{\hbar} \int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3} \frac{e^2}{\varepsilon q^2} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot(\boldsymbol{y}_2-\boldsymbol{y}_1)} \int \frac{\mathrm{d}\boldsymbol{k}''}{(2\pi)^3} G^0(\boldsymbol{k}'', 0) \mathrm{e}^{\mathrm{i}\boldsymbol{k}''\cdot(\boldsymbol{y}_2-\boldsymbol{y}_1)}. \end{aligned} \tag{9.125}$$

The first step is to notice that the integration over $t_1$ (part of $\boldsymbol{y}_1$) produces the factor $2\pi\,\delta(\omega-\omega')$, which can then be used in the integration over $\omega'$. Similarly, the integration over $t_2$ (part of $\boldsymbol{y}_2$) incorporates the delta function on the two times and yields unity. The integrations over $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ produce $(2\pi)^3\delta(\boldsymbol{q}+\boldsymbol{k}''-\boldsymbol{k})$ and $(2\pi)^3\delta(\boldsymbol{q}+\boldsymbol{k}''-\boldsymbol{k}')$, respectively, which also lead to $\delta(\boldsymbol{k}-\boldsymbol{k}')$. We can use these delta functions to integrate out $\boldsymbol{k}'$ and $\boldsymbol{k}''$. This leads to the result

$$\begin{aligned} G_1 = & \int \frac{\mathrm{d}\boldsymbol{k}}{(2\pi)^3} \int \frac{\mathrm{d}\omega}{2\pi} G^0(\boldsymbol{k}, \omega) G(\boldsymbol{k}, \omega) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')+\mathrm{i}\omega(t-t')} \\ & \times \frac{\mathrm{i}}{\hbar} \int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3} \frac{e^2}{\varepsilon q^2} G^0(\boldsymbol{k}-\boldsymbol{q}, 0) \end{aligned} \tag{9.126}$$

from which we can recognize the self-energy contribution by comparing with (9.114):

$$\Sigma_1 = \frac{\mathrm{i}}{\hbar} \int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3} \frac{e^2}{\varepsilon q^2} G^0(\boldsymbol{k}-\boldsymbol{q}, 0). \tag{9.127}$$

Now, the second term in (9.115) leads to

$$\begin{aligned} G_2 = & -\int \mathrm{d}^4 y_1 \int \mathrm{d}^4 y_2 \int \frac{\mathrm{d}\boldsymbol{k}}{(2\pi)^3} \int \frac{\mathrm{d}\omega}{2\pi} G^0(\boldsymbol{k}, \omega) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y}_1)+\mathrm{i}\omega(t-t_1)} \\ & \times \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} \int \frac{\mathrm{d}\omega'}{2\pi} G(\boldsymbol{k}', \omega') \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot(\boldsymbol{y}_2-\boldsymbol{x}')+\mathrm{i}\omega'(t_1-t')} \delta(t_1-t_2) \\ & \times \frac{\mathrm{i}}{\hbar} \int \mathrm{d}\boldsymbol{y}_3 \int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3} \frac{e^2}{\varepsilon q^2} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot(\boldsymbol{y}_3-\boldsymbol{y}_1)} \\ & \times \int \frac{\mathrm{d}\boldsymbol{k}''}{(2\pi)^3} G^0(\boldsymbol{k}'', 0) \mathrm{e}^{\mathrm{i}\boldsymbol{k}''\cdot(\boldsymbol{y}_3-\boldsymbol{y}_1)} \delta(\boldsymbol{x}_1-\boldsymbol{x}_2). \end{aligned} \tag{9.128}$$

Again, integrating over $t_1$ leads to a delta function for the two frequencies that can be absorbed in the integration over $\omega'$. Integration over $t_2$ involves the time-conserving delta function in the fourth line. Similarly, the integration over $\boldsymbol{y}_2$ involves the delta function and leads to a factor of unity. Integration over $\boldsymbol{y}_3$ leads to a factor of $(2\pi)^3\delta(\boldsymbol{q}+\boldsymbol{k}'')$, which is absorbed in the integration over $\boldsymbol{k}''$. Finally, the next integration over $\boldsymbol{y}_1$ leads to $(2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}')$, which is then absorbed in the integration over $\boldsymbol{k}'$. With these changes, the result is then

$$\begin{aligned} G_2 = &-\int \frac{\mathrm{d}\boldsymbol{k}}{(2\pi)^3}\int\frac{\mathrm{d}\omega}{2\pi}G^0(\boldsymbol{k},\omega)G(\boldsymbol{k},\omega)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')+\mathrm{i}\omega(t-t')} \\ &\times\frac{\mathrm{i}}{\hbar}\int\frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3}\frac{e^2}{\varepsilon q^2}G^0(0,0) \end{aligned} \tag{9.129}$$

from which we can see that

$$\Sigma_2 = -\frac{\mathrm{i}}{\hbar}\int\frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3}\frac{e^2}{\varepsilon q^2}G^0(0,0). \tag{9.130}$$

The transform of the unperturbed Green's function in the last term is the value at zero wave vector and zero frequency, which is normally written as the average density, and this term is recognized as the Hartree term. The problem is that the assignment of a non-zero wave vector to the interaction line normally means that the Green's functions on the input and output lines must have wave vectors that differ by this amount, but they have the same wave vector. This must be interpreted as meaning that we must seek the value for the interaction as $\boldsymbol{q}\to\boldsymbol{0}$. Since (9.130) is normally interpreted as a convolution integral, it is often seen in a form in which the dependence on $\boldsymbol{q}$ is transferred from the interaction to the Green's function in order to keep this consistency of notation. In any case, the expression (9.130) has no spatial variation and is an average potential—the Hartree potential. The final form of the self-energy is then

$$\Sigma(k) = \mathrm{i}\hbar\int\frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3}\frac{e^2}{\varepsilon q^2}[G^0(\boldsymbol{k}-\boldsymbol{q},0)-G^0(0,0)]. \tag{9.131}$$

This form can now be evaluated, but with care taken because of the problem with the second term and the Hartree energy.

## References

Abrikosov A A, Gorkov L P and Dzyaloshinskii I E 1963 *Methods of Quantum Field Theory in Statistical Physics* (Englewood Cliffs, NJ: Prentice-Hall) section 7

Haken H 1976 *Quantum Field Theory of Solids* (Amsterdam: North-Holland)

Pauli W 1925 *Z. Phys.* **31** 765

Pauli W 1927 *Z. Phys.* **43** 601

Slater J C 1929 *Phys. Rev.* **34** 1293

## Problems

1. Consider two electrons in a state in which the radius (around some orbit centre) is set normalized to unity, and the angular wave function is $\phi(\theta) = f[\cos(\theta)]$. If these two electrons have opposite spins and are located at $f[\cos(\theta)]$ and $f[\cos(\theta + \pi)]$, discuss the anti-symmetric properties of these electrons. Can you infer the nature of the function describing the angular variations?

2. Show that the number operator

$$N = \int \Psi^{+}(\xi)\Psi(\xi)\,\mathrm{d}\xi$$

commutes with the Hamiltonian (9.47).

3. Using the Hamiltonian (9.71), determine the dynamical equations of motion for the creation and annihilation operators for the electrons.

4. Expand the perturbation series to the second-order terms and draw the ten distinct connected diagrams. Explain how these can be separated into six diagrams for the second-order terms of the self-energy and the four second-order terms for $G$.

5. Consider a perturbation that scatters an electron from state $\boldsymbol{k}$ to state $\boldsymbol{k}'$ while creating or annihilating one unit of lattice vibration. In section 4.6, it was shown how the vibrating lattice could be represented in Fourier space as a summation of a set of harmonic oscillators. Devise a perturbing potential that accounts for the scattering of the above electron by the lattice (it will take three operators, two for the electrons and one for the boson). Carry out lowest-order perturbation and sketch the diagrams and Green's functions that will result.

# Solutions to selected problems

## Chapter 1

2. We use de Broglie's wave expression $\lambda = h/p = h/\sqrt{2mE}$. The frequency is always given by $E = hn$, so $n = 2.42 \times 10^{16}$ Hz. The wavelengths are

$$\lambda_{\mathrm{e}} = 1.22 \times 10^{-10}\ \mathrm{m} \qquad \lambda_{\mathrm{p}} = 2.85 \times 10^{-12}\ \mathrm{m}.$$

The group velocity is $h/\lambda m\ (= \hbar k/m)$ and the phase velocity is half of this value. This leads to

| | $v_\phi$ | $v_{\mathrm{gr}}$ |
|---|---|---|
| electron | $2.97 \times 10^6$ | $5.95 \times 10^6$ m s$^{-1}$ |
| proton | $6.93 \times 10^4$ | $1.38 \times 10^5$ m s$^{-1}$. |

3. We work in momentum space, with

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

The expectation value is then

$$\begin{aligned} \langle x \rangle &= \int_{-\infty}^{\infty} \Psi^*(x) x \Psi(x)\,\mathrm{d}x \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \varphi^*(k') \mathrm{e}^{-\mathrm{i}k'x}\,\mathrm{d}k' \right\} x \left\{ \int_{-\infty}^{\infty} \varphi(k) \mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k \right\} \mathrm{d}x \end{aligned}$$

which then is manipulated as

$$\begin{aligned} \langle x \rangle &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi^*(k')\,\mathrm{d}k' \int_{-\infty}^{\infty} \mathrm{d}x\,\mathrm{e}^{-\mathrm{i}k'x} \int_{-\infty}^{\infty} \mathrm{d}k\,\varphi(k) \left( -\mathrm{i}\frac{\partial}{\partial k}\,\mathrm{e}^{\mathrm{i}kx} \right) \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi^*(k')\,\mathrm{d}k' \int_{-\infty}^{\infty} \mathrm{d}x\,\mathrm{e}^{-\mathrm{i}k'x} \int_{-\infty}^{\infty} \mathrm{d}k\,\mathrm{e}^{\mathrm{i}kx} \left( \mathrm{i}\frac{\partial}{\partial k}\varphi(k) \right) \\ &= \int_{-\infty}^{\infty} \varphi^*(k')\,\mathrm{d}k' \int_{-\infty}^{\infty} \mathrm{d}k \left( \mathrm{i}\frac{\partial}{\partial k}\varphi(k) \right) \delta(k' - k) \end{aligned}$$

$$= \int_{-\infty}^{\infty} \mathrm{d}k \, \varphi^*(k) \left( \mathrm{i} \frac{\partial}{\partial k} \varphi(k) \right)$$
$$= \left\langle \mathrm{i} \frac{\partial}{\partial k} \right\rangle = \left\langle \mathrm{i}\hbar \frac{\partial}{\partial p} \right\rangle .$$

4. Using (1.52), with $\sigma = \Delta x$, gives

$$\Delta x = \sigma \sqrt{1 + \left( \frac{\hbar t}{2m\sigma^2} \right)^2} .$$

For $\Delta x$ to equal $2\sigma$, we require that the term in parentheses take the value $\sqrt{3}$. Thus,

$$t = \sqrt{3} \frac{2m\sigma^2}{\hbar} .$$

Now, the average energy $E = \langle p^2 \rangle / 2m$, but $\langle p^2 \rangle = \langle p \rangle^2 + \langle (\Delta p)^2 \rangle = 1.01 \langle p \rangle^2$. Thus, $\langle p \rangle = \sqrt{2mE}$ to within 0.01%, so

$$t = \sqrt{3} \frac{2m\sigma^2}{\hbar} = \sqrt{3} \frac{\hbar m}{2(\Delta p)^2} = 50\sqrt{3} \frac{\hbar m}{\langle p \rangle^2} = 25\sqrt{3} \frac{\hbar}{E} = 2.88 \times 10^{-16} \mathrm{s}.$$

## Chapter 2

1. We can summarize the picture, by defining the momentum wave function as

$$\phi(k) = 2a \left[ 1 - \left| \frac{ka}{\pi} \right| \right] .$$

This, however, is not normalized properly, and we must check the proper normalization:

$$\int_{-\pi/a}^{\pi/a} |\phi(k)|^2 \, \mathrm{d}k = 8a^2 \int_{-\pi/a}^{\pi/a} \left( 1 - \left| \frac{ka}{\pi} \right| \right)^2 \mathrm{d}k = \frac{8\pi a}{3} .$$

The normalized wave function is then

$$\phi(k) = \sqrt{\frac{3a}{8\pi}} 2a \left[ 1 - \left| \frac{ka}{\pi} \right| \right] = \sqrt{\frac{3a^3}{2\pi}} \left[ 1 - \left| \frac{ka}{\pi} \right| \right] .$$

This is now Fourier transformed into real space to give the wave function as

$$\psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}kx} \, \mathrm{d}k = \frac{\sqrt{3a^2}}{2\pi} \int_{-\pi/a}^{\pi/a} \left[ 1 - \left| \frac{ka}{\pi} \right| \right] \mathrm{e}^{\mathrm{i}kx} \, \mathrm{d}k$$
$$= \sqrt{\frac{3}{a}} \left( \frac{a}{\pi x} \right)^2 \left[ 1 - \cos\left( \frac{\pi x}{a} \right) \right] .$$

We note first that this wave function is *symmetrical* about $x = 0$, so $\langle x \rangle = 0$. Thus, we then find that $(\Delta x)^2 = \langle x^2 \rangle$, and

$$(\Delta x)^2 = \frac{3a^4}{\pi^4} \int_{-\infty}^{\infty} \frac{1}{x^2} \left[1 - \cos\left(\frac{\pi x}{a}\right)\right]^2 \mathrm{d}x = \frac{6a^2}{\pi^2}.$$

Using the momentum wave function, we similarly find

$$(\Delta k)^2 = \int_{-\pi/a}^{\pi/a} k^2 \frac{3a^3}{2\pi} \left[1 - \left|\frac{ka}{\pi}\right|\right]^2 \mathrm{d}k = \frac{\pi^2}{a^2 10}$$

so the uncertainty relation is found to be

$$\Delta x \Delta k = \sqrt{\frac{3}{10}}.$$

2. In general, one thinks that the kinetic energy is reduced once the wave passes over the barrier, and this leads to a smaller value of $k'$, and hence a longer wavelength $(2\pi/k)$. However, how does this appear in a wave packet? The important point is that *anything that narrows the momentum distribution will broaden the spatial distribution just due to the properties of the Fourier transform*. Hence, we consider a Gaussian wave packet, centred at $k_1$, as in (following (1.33))

$$\phi(k) = \left(\frac{2}{\pi}\right)^{1/4} \sqrt{\sigma} \exp[-\sigma^2 (k - k_0)^2].$$

For this wave packet, $\langle k \rangle = k_0$, and $\Delta p = \hbar/2\sigma$ from (1.34). *We note that $\Delta p$ is a function only of the width of the packet and not the centroid.* If we now pass this wave packet over the barrier $((k_0 > k_V = \sqrt{2mV_0/\hbar^2})$, then not much happens until the barrier begins to eat away part of the Gaussian distribution. This narrows the distribution, which means that the real space wave function (obtained from the Fourier transform) must become broader.

6. For this problem, we note that $\langle T \rangle = \langle V \rangle$ (which is true for any oscillating system, where the energy oscillates between being purely potential energy and purely kinetic energy), and this leads to (since we assume that $\langle x \rangle = \langle p \rangle = 0$)

$$(\Delta x)^2 = \frac{1}{m^2\omega^2} (\Delta p)^2 = \frac{1}{m^2\omega^2} \left(\frac{\hbar}{2\Delta x}\right)^2$$

and

$$(\Delta x) = \sqrt{\frac{\hbar}{2m\omega}} \qquad (\Delta p) = \sqrt{\frac{m\omega\hbar}{2}}$$

for which

$$E = \langle H \rangle = \frac{1}{2m} (\Delta p)^2 + \frac{m\omega^2}{2} (\Delta x)^2 = \frac{\hbar\omega}{2}.$$

7. First, one determines the value of

$$\beta = \sqrt{\frac{mV_0a^2}{2\hbar^2}} = \sqrt{7.277}$$

from which it may be determined that solutions can be found only for $ka/2 \leq 2.7$. This means that there is only one root from each of the two solution sets (e.g., one even-symmetry function and one odd-symmetry function). These two energy levels are (measured from the bottom of the well) 0.053 eV and 0.197 eV. When we shift to measuring the levels from the top of the well, the energy levels are at $-0.247$ eV and $-0.103$ eV. From this, the values of $k$ and $\gamma$ are easily determined and the wave functions plotted.

8. Here, it is easiest if an integral representation for the Airy function is used, e.g.

$$\mathrm{Ai}\left(\pm\frac{x}{\theta}\right) = \frac{1}{\pi}\int_{-\infty}^{\infty} \exp\left[\mathrm{i}\frac{\xi^3}{3} \pm \mathrm{i}\frac{x}{\theta}\xi\right] \mathrm{d}\xi$$

with

$$\theta = \left(\frac{\hbar^2}{2meE}\right)^{1/3}.$$

The wave function that we begin with is for the $i$th energy level:

$$\psi_i(x) = C_i\mathrm{Ai}\left(\frac{x - x_{0,i}}{\theta}\right) = \frac{1}{\pi}\int_{-\infty}^{\infty} \exp\left[\mathrm{i}\frac{\xi^3}{3} - \mathrm{i}\left(\frac{x - x_{0,i}}{\theta}\right)\xi\right] \mathrm{d}\xi$$

with

$$x_{0,i} = \left[\frac{3\pi}{8}(4i - 1)\right]^{2/3}.$$

The Airy functions with different values of $x_{0,i}$ are orthogonal to one another. Thus, we need only show that this property is maintained in the transforming to momentum space. Thus, we start with the latter, as

$$\begin{aligned}
\varphi_i(k) &= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} \psi_i(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x = \frac{C_i}{\sqrt{2\pi}}\int_0^{\infty} \mathrm{Ai}\left(\frac{x - x_{0,i}}{\theta}\right)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x \\
&= \frac{C_i}{\sqrt{2\pi}}\frac{1}{\pi}\int_0^{\infty} \mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x \int_{-\infty}^{\infty} \exp\left[\mathrm{i}\frac{\xi^3}{3} + \mathrm{i}\frac{(x - x_{0,i})}{\theta}\xi\right] \mathrm{d}\xi \\
&= \frac{C_i}{\sqrt{2\pi}}\frac{1}{\pi}\int_{-\infty}^{\infty} \exp\left[\mathrm{i}\frac{\xi^3}{3} - \mathrm{i}\frac{x_{0,i}}{\theta}\xi\right] \mathrm{d}\xi \int_0^{\infty} \exp\left[\mathrm{i}\left(\frac{\xi}{\theta} - k\right)x\right] \mathrm{d}x \\
&= \frac{C_i}{\sqrt{2\pi}}\frac{1}{\pi}\int_{-\infty}^{\infty} \exp\left[\mathrm{i}\frac{\xi^3}{3} - \mathrm{i}\frac{x_{0,i}}{\theta}\xi\right] \mathrm{d}\xi 2\pi\theta\delta(\xi - \theta k) \\
&= C_i\theta\sqrt{\frac{2}{\pi}} \exp\left[\mathrm{i}\left(\frac{\theta k}{3}\right)^3 - \mathrm{i}x_{0,i}k\right]
\end{aligned}$$

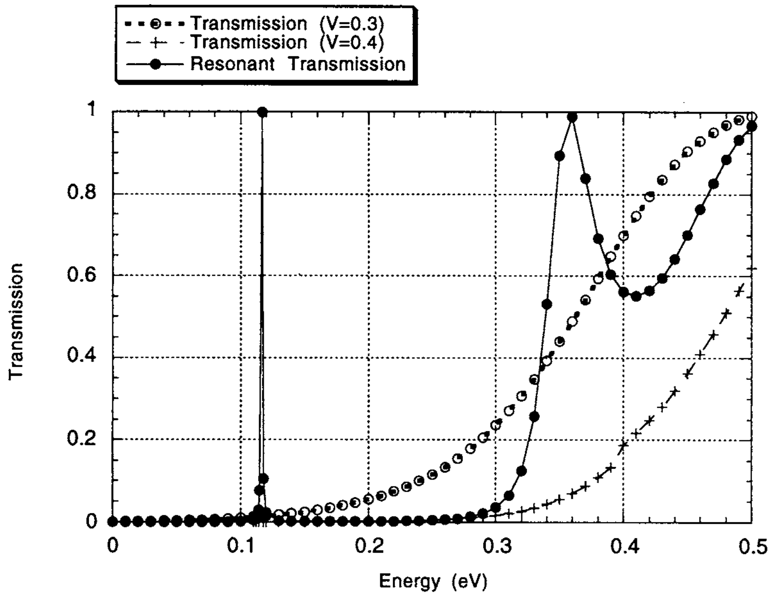
and we check the orthogonality through

$$\begin{aligned} I_{ij} &= \frac{C_i C_j \theta^2}{\pi} \int_{-\infty}^{\infty} \left\{ \exp\left[ \mathrm{i} \left( \frac{\theta k}{3} \right)^3 - \mathrm{i} x_{0,i} k \right] \right\}^* \\ &\quad \times \left\{ \exp\left[ \mathrm{i} \left( \frac{\theta k}{3} \right)^3 - \mathrm{i} x_{0,j} k \right] \right\} \mathrm{d}k \\ &\quad \times \frac{C_i C_j \theta^2}{\pi} \int_{-\infty}^{\infty} \exp\left[ \mathrm{i} k (x_{0,i} - x_{0,j}) \right] \mathrm{d}k = 2 C_i C_j \theta^2 \, \delta(x_{0,i} - x_{0,j}) \end{aligned}$$

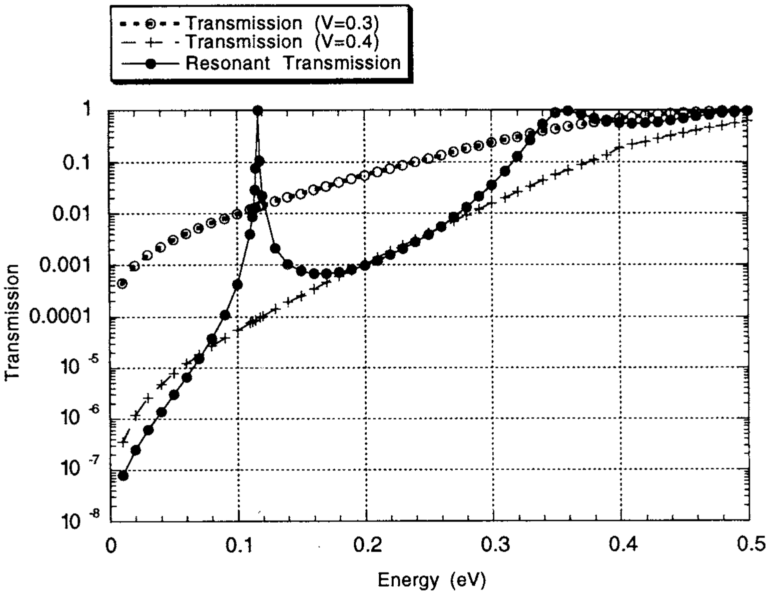
which clearly establishes the orthogonality, as the integral is zero unless the two zeros of the Airy functions coincide. Further, this may be used to show that $C_i = \theta/\sqrt{2}$.

## Chapter 3

1–3. For a linear axis, these may be plotted as

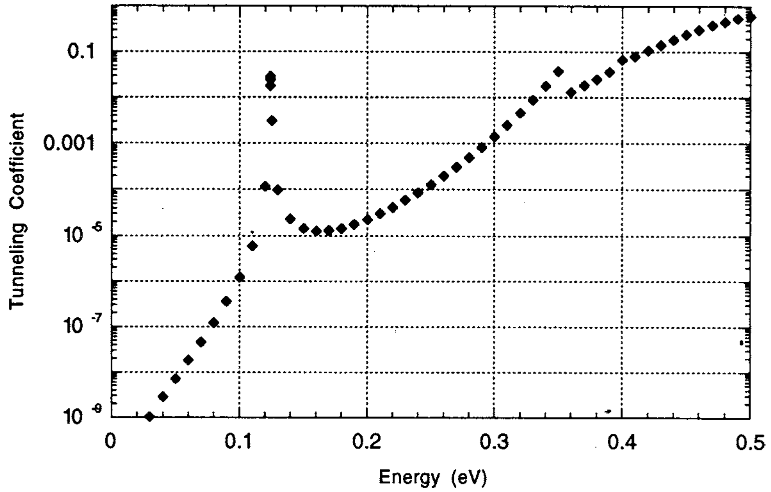


For a logarithmic transmission, the results are



For problem 1, the single bound-state resonance is at 0.1166 eV, although there is a resonance at about 0.36 eV.

4. The resonant peak is now at 0.124 eV, and the transmission at this point is 0.029. The values from problem 1 and problem 2 at this energy are 0.0156 and 0.000 116 85, respectively, so $4T_{\min}/T_{\max}$ is 0.0299, which is about as close as one can get on such a problem.



5. This problem is solved using the 'general' result:

$$\int_0^a k(x)\,\mathrm{d}x = \int_0^a \sqrt{\frac{2m}{\hbar^2}[\varepsilon + V_1 + eEx]}\,\mathrm{d}x = (2n+1)\frac{\pi}{2}$$

which leads to the result (with the energy $\varepsilon$ in eV)

$$(\varepsilon + 0.4)^{3/2} - (\varepsilon + 0.3)^{3/2} = \frac{3E\pi}{4e^{1/2}}\sqrt{\frac{\hbar^2}{2m}}(2n+1)$$

This leads to a single level in the well, given by $\varepsilon_0 = -0.2886$ eV.

6. For this problem, we cannot use the normal formula, because of the sharp potential at $x = 0$. At the energy of the bound states, each energy $\varepsilon_i$ can be related to a turning point $x_i$ via $\varepsilon_i = eEx_i$. For $x > x_i$, the decaying wave function is given by

$$\psi \sim \frac{1}{\sqrt{\gamma}} \exp\left(-\int_{x_i}^{x} \gamma(x)\,\mathrm{d}x\right)$$

and this must be connected to the cosine function

$$\psi \sim \frac{2}{\sqrt{k}} \cos\left(\int_{x}^{x_i} k(x)\,\mathrm{d}x - \frac{\pi}{4}\right).$$

Now, this latter wave function must vanish at $x = 0$, so the bound states are found from

$$\cos\left(\int_0^{x_i} k(x)\,\mathrm{d}x - \frac{\pi}{4}\right) = 0.$$

Using the above energy relation, we can write this as

$$\int_0^{x_i} k(x)\,\mathrm{d}x = \sqrt{\frac{2meE}{\hbar^2}} \int_0^{x_i} \sqrt{x_i - x}\,\mathrm{d}x = (2i+1)\frac{\pi}{2} + \frac{\pi}{4}$$

which leads to

$$\varepsilon_i = eEx_i = \left(\frac{\hbar^2 e^2 E^2}{2m}\right)^{1/3} \left[\frac{3\pi}{4}\left(2i + \frac{3}{2}\right)\right]^{2/3}.$$

It turns out that these are precisely the values found from solving the Airy equation (once we adjust the latter for $i = 0, 1, 2, \ldots$, and not $n = 1, 2, \ldots$).

8. For this problem, we can return to our 'general' formula. Again, noting that the energy eigenvalues will have corresponding turning points $x_n$ through $E_n = a(x_n)^4$, we have

$$\begin{aligned}(2n+1)\frac{\pi}{2} &= \int_{-x_n}^{x_n} k(x)\,\mathrm{d}x = \sqrt{\frac{2ma}{\hbar^2}} \int_{-x_n}^{x_n} \sqrt{x_n^4 - x^4}\,\mathrm{d}x \\ &= 2x_n^3 \sqrt{\frac{2ma}{\hbar^2}}\end{aligned}$$

and the energy eigenvalues are given by

$$E_n = a\left[\sqrt{\frac{\hbar^2}{2ma}}\left(n+\frac{1}{2}\right)\frac{\pi}{2}\right]^{4/3}.$$

**Chapter 4**

1. The potential is given by $m\omega^2x^2/2$. We seek energy levels given by $E_n$, with the turning points

$$E_n = \frac{1}{2}m\omega^2a^2 \to a = \sqrt{\frac{2E_n}{m\omega^2}}.$$

Via the basic approach we now seek solutions to the phase integral

$$\begin{aligned}\int_{-a}^{a} k(x)\,\mathrm{d}x &= \int_{-a}^{a}\sqrt{\frac{2m}{\hbar^2}\left(E-\frac{1}{2}m\omega^2x^2\right)}\,\mathrm{d}x\\ &= \frac{m\omega}{\hbar}\int_{-a}^{a}\sqrt{a^2-x^2}\,\mathrm{d}x = (2n+1)\frac{\pi}{2}\end{aligned}$$

which leads to

$$\frac{m\omega s^2}{\hbar}\frac{\pi}{2} = (2n+1)\frac{\pi}{2} \to a = \frac{(2n+1)\hbar}{m\omega} \to E_n = \left(n+\frac{1}{2}\right)\hbar\omega$$

which is the exact answer. We conclude that a quadratic potential is a 'soft' potential.

3. The energy $5\hbar\omega/2$ corresponds to the energy level $n = 2$. Using the creation operators, we find

$$\Psi_2(x) = \frac{1}{\sqrt{2}}\left(\frac{m\omega}{\pi\hbar}\right)^{1/2}\mathrm{e}^{-m\omega x^2/\hbar}\left(\frac{2m\omega}{\hbar}x^2-1\right).$$

Then,

$$P[|x|>x_0] = \left(\frac{m\omega}{\pi\hbar}\right)^{1/2}\int_{x_0}^{\infty}\left(\frac{2m\omega}{\hbar}x^2-1\right)\mathrm{e}^{-m\omega x^2/\hbar}\,\mathrm{d}x$$

which leads to

$$P[|x|>x_0] = \frac{4}{\sqrt{\pi}}\left(6\sqrt{5}\mathrm{e}^{-5}+\frac{\sqrt{\pi}}{8}\mathrm{erfc}(\sqrt{5})\right).$$

6. This may be carried out with the operator algebra. However, to understand the result, we must define wave functions. Since we do not know which state is occupied, we will take an arbitrary occupation of each state via the definition

$$\Psi(x) = \sum_n a_n\psi_n(x)$$

where the $\psi_n$ are the harmonic oscillator wave functions for state $n$. Then, the expectation values are found from ($p^2$ is used as the example)

$$\begin{aligned}\langle p^2\rangle = &\left\langle\left[-\sqrt{\frac{m\hbar\omega}{2}}(a-a^+)\right]^2\right\rangle = \left\langle -\frac{m\hbar\omega}{2}(a^2+a^{+2}-aa^+-a^+a)\right\rangle \\ = &-\frac{m\hbar\omega}{2}\{-\langle 2a^+a+1\rangle + \langle a^2\rangle + \langle a^{+2}\rangle\}\end{aligned}$$

for which the $nk$ matrix element is

$$\begin{aligned}\langle p^2\rangle_{nk} = \frac{m\hbar\omega}{2}\Big\{&(2n+1)\delta_{nk} - \sqrt{(n+1)(n+2)}\delta_{n+2,k} \\ &- \sqrt{n(n-1)}\delta_{n-2,k}\Big\}.\end{aligned}$$

Similarly,

$$\begin{aligned}\langle x^2\rangle_{nk} = \frac{\hbar}{2m\omega}\Big\{&(2n+1)\delta_{nk} + \sqrt{(n+1)(n+2)}\delta_{n+2,k} \\ &+ \sqrt{n(n-1)}\delta_{n-2,k}\Big\}.\end{aligned}$$

Thus, in both cases, only states either equal in index or separated by 2 in index can contribute to the expectation values.

7. Using the results of problem 6, we have

$$\begin{aligned}T_{nk} &= \frac{1}{2m}\langle p^2\rangle_{nk} \\ &= \frac{\hbar\omega}{4}\left\{(2n+1)\delta_{nk} - \sqrt{(n+1)(n+2)}\delta_{n+2,k} - \sqrt{n(n-1)}\delta_{n-2,k}\right\} \\ V_{nk} &= \frac{m\omega^2}{2}\langle x^2\rangle_{nk} \\ &= \frac{\hbar\omega}{4}\left\{(2n+1)\delta_{nk} + \sqrt{(n+1)(n+2)}\delta_{n+2,k} + \sqrt{n(n-1)}\delta_{n-2,k}\right\}.\end{aligned}$$

11. We take $\boldsymbol{B} = B_0\boldsymbol{a}_z$, which we can get from the vector potential $\boldsymbol{A} = (0, Bx, 0)$. This will give a harmonic oscillator in the $x$-direction, and for homogeneity in the $y$-direction, we take $\psi(x,y) = \phi(x)\exp(\mathrm{i}ky)$. Then, the Schrödinger equation becomes

$$\left[\frac{1}{2m}(\hbar k - eB_0x)^2 - \frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{m\omega_0^2}{2}x^2\right]\phi(x) = E\phi(x)$$

which with the change of variables

$$x_0 = \frac{\hbar k}{m}\frac{\omega_\mathrm{c}}{\omega_0^2+\omega_\mathrm{c}^2} \qquad \omega_\mathrm{c} = \frac{eB_0}{m} \qquad \Omega^2 = \omega_0^2+\omega_\mathrm{c}^2$$

gives

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}+\frac{1}{2}m\Omega^2(x-x_0)^2\phi=\left[E-\frac{\hbar^2k^2}{2m}+\frac{1}{2}m\Omega^2x_0^2\right]\phi$$

which is a shifted harmonic oscillator and leads to the energy levels

$$E_n=\left(n+\frac{1}{2}\right)\hbar\Omega+\frac{\hbar^2k^2}{2m}-\frac{1}{2}m\Omega^2x_0^2.$$

## Chapter 5

1. The Hamiltonian is given as $H=p^2+Ax+Bpx$. From (5.16)

$$\begin{aligned}\frac{\mathrm{d}\langle p\rangle}{\mathrm{d}t}&=\frac{\mathrm{i}}{\hbar}\langle[H,p]\rangle=\frac{\mathrm{i}}{\hbar}\langle[p^2,p]+A[x,p]+B[px,p]\rangle\\&=\frac{\mathrm{i}}{\hbar}\langle 0+\mathrm{i}\hbar A+B(pxp-p^2x)\rangle=\frac{\mathrm{i}}{\hbar}\langle 0+\mathrm{i}\hbar A+Bp(xp-px)\rangle\\&=-A-B\langle p\rangle\\\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t}&=\frac{\mathrm{i}}{\hbar}\langle[H,x]\rangle=\frac{\mathrm{i}}{\hbar}\langle[p^2,x]+A[x,x]+B[px,x]\rangle\\&=\frac{\mathrm{i}}{\hbar}\langle p^2x-xp^2+0+B(px^2-xpx)\rangle\\&=\frac{\mathrm{i}}{\hbar}\langle p(xp-\mathrm{i}\hbar)-(px-\mathrm{i}\hbar)p-\mathrm{i}\hbar Bx\rangle\\&=2\langle p\rangle+B\langle x\rangle.\end{aligned}$$

2. The eigenvalues are found from

$$\frac{1}{2}\begin{vmatrix}(\sqrt{2}-2E)&\sqrt{2}&0\\-\sqrt{2}&(\sqrt{2}-2E)&0\\0&0&(2-2E)\end{vmatrix}=0$$

(the factor of 2 multiplying the $E$ is due to the pre-factor) and this leads to $E_3=1$, with the other two roots being given by

$$\begin{aligned}(\sqrt{2}-2E)^2+2&=0\\E_{1,2}=\frac{1}{\sqrt{2}}\pm\sqrt{\frac{1}{2}-1}&=\frac{1\pm\mathrm{i}}{\sqrt{2}}\end{aligned}$$

3. (i) $H=2x^2p^2$

$$\begin{aligned}\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t}&=\frac{\mathrm{i}}{\hbar}\langle[2x^2p^2,x]\rangle=4\langle x^2p\rangle\\\frac{\mathrm{d}\langle p\rangle}{\mathrm{d}t}&=\frac{\mathrm{i}}{\hbar}\langle[2x^2p^2,p]\rangle=-4\langle xp^2\rangle\end{aligned}$$

(ii) $H = x^2p^2 + p^2x^2$

$$\begin{aligned}\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t} &= +2\langle x^2p\rangle + \frac{\mathrm{i}}{\hbar}\langle p^2x^3 - xp^2x^2\rangle = -4\langle xpx\rangle\\ \frac{\mathrm{d}\langle p\rangle}{\mathrm{d}t} &= -2\langle xp^2\rangle + \frac{\mathrm{i}}{\hbar}\langle p^2x^2p - p^3x^2\rangle = -4\langle pxp\rangle\end{aligned}$$

(iii) $H = 2(xp)^2$

$$\begin{aligned}\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t} &= \frac{\mathrm{i}}{\hbar}\langle[2xpxp, x]\rangle = 4\langle x^2p\rangle - 2\mathrm{i}\hbar\langle x\rangle\\ \frac{\mathrm{d}\langle p\rangle}{\mathrm{d}t} &= \frac{\mathrm{i}}{\hbar}\langle[2xpxp, p]\rangle = -4\langle p^2x\rangle + 2\mathrm{i}\hbar\langle p\rangle\end{aligned}$$

4. $[A, B] = C$, in which $C$ is a $c$-number. Thus,

$$\begin{aligned}A^3B^3 - (AB)^3 &= A^3B^3 - ABABAB\\ &= A^3B^3 - A(AB - C)(AB - C)B\\ &= A^3B^3 - A^2BAB^2 + 2CA^2B^2 - C^2AB\\ &= A^3B^3 - A^2(AB - C)B^2 + 2CA^2B^2 - C^2AB\\ &= 3CA^2B^2 - C^2AB = 2C^2AB + 3C(AB)^2.\end{aligned}$$

## Chapter 6

1. We have $V(x) = 0$ for $0 < x < a$, and $\to \infty$ elsewhere. The perturbing potential is given by $V_1 = (x - a)^2/2$. The unperturbed wave functions and energies are then given by

$$\begin{aligned}\varphi_n(x) &= \sqrt{\frac{2}{a}}\sin\left(\frac{n\pi x}{a}\right) \qquad 0 \le x \le a\\ E_n &= \frac{n^2\pi^2\hbar^2}{2ma^2}.\end{aligned}$$

For the effect of the perturbation, we find the matrix elements

$$\begin{aligned}V_{nm} &= \frac{1}{a}\int_0^a \sin\left(\frac{n\pi x}{a}\right)(x - a)^2 \sin\left(\frac{m\pi x}{a}\right)\,\mathrm{d}x\\ &= \begin{cases}\dfrac{a^2}{\pi^2}\left[\dfrac{1}{(n-m)^2} - \dfrac{1}{(n+m)^2}\right] & n \neq m\\ \dfrac{a^2}{6}\left[1 - \dfrac{3}{2n^2\pi^2}\right] & n = m.\end{cases}\end{aligned}$$

The first term can be simplified, but the perturbed wave function can be written as

$$\varphi_n(x) = \sqrt{\frac{2}{a}}\sin\left(\frac{n\pi x}{a}\right)$$

$$+\sum_{l\neq n}\frac{a^2}{\pi^2}\frac{4nl}{(n^2-l^2)^2}\left\{\frac{\hbar^2\pi^2}{2ma^2}(l^2-n^2)\right\}^{-1}\sqrt{\frac{2}{a}}\sin\left(\frac{l\pi x}{a}\right)$$
$$=\sqrt{\frac{2}{a}}\sin\left(\frac{n\pi x}{a}\right)-\frac{2ma^2}{\hbar^2\pi^4}\sqrt{\frac{2}{a}}\sum_{l\neq n}\frac{4nl}{(n^2-l^2)^3}\sin\left(\frac{l\pi x}{a}\right).$$

The energy levels are now given by

$$\begin{aligned}E_n &= E_n^{(0)}+E_n^{(1)}+E_n^{(2)}\\ &=\frac{n^2\pi^2\hbar^2}{2ma^2}+\frac{a^2}{6}\left[1-\frac{3}{2n^2\pi^2}\right]-\left(\frac{a^2}{\pi^2}\right)^2\frac{2ma^2}{\pi^2\hbar^2}\sum_{l\neq n}\frac{16n^2l^2}{(n^2-l^2)^5}.\end{aligned}$$

2. The third-order correction to the energy is given by

$$E_n^{(3)}=\sum_{l\neq n}\left\{\frac{V_{nl}(V_{ll}-V_{nn})V_{ln}}{(E_n^{(0)}-E_l^{(0)})^2}+\sum_{k\neq n,l}\frac{V_{nl}V_{lk}V_{kn}}{(E_n^{(0)}-E_l^{(0)})(E_n^{(0)}-E_k^{(0)})}\right\}.$$

4. The lowest energy level is for $n=m=1$,

$$E_{11}=\frac{\pi^2\hbar^2}{2ma^2}(1^2+1^2)=\frac{\pi^2\hbar^2}{ma^2}$$

since the wave functions are

$$\varphi_{x,n}(x)=\sqrt{\frac{2}{a}}\sin\left(\frac{n\pi x}{2}+\frac{n\pi}{2}\right)\qquad\varphi_{y,m}(y)=\sqrt{\frac{2}{a}}\sin\left(\frac{m\pi y}{2}+\frac{m\pi}{2}\right).$$

The lowest degenerate energy level occurs for the sets of $(n,m)=(2,1),(1,2)$, which both give the energy level of

$$E_{21}=E_{12}=\frac{\pi^2\hbar^2}{2ma^2}(2^2+1^2)=\frac{5\pi^2\hbar^2}{2ma^2}.$$

If we call the $(2,1)$ set $a$ and the $(1,2)$ set $b$, then the matrix element which couples these two wave functions is

$$\begin{aligned}V_{ab}&=V_0\int_{-a/4}^{a/4}\mathrm{d}x\int_{-a/4}^{a/4}\mathrm{d}y\,\varphi_{21}^*(x,y)\varphi_{12}(x,y)\\ &=\frac{4V_0}{a^2}\int_{-a/4}^{a/4}\mathrm{d}x\,\sin\left(\frac{2\pi x}{a}\right)\cos\left(\frac{\pi x}{a}\right)\int_{-a/4}^{a/4}\mathrm{d}y\,\cos\left(\frac{\pi y}{a}\right)\sin\left(\frac{2\pi y}{a}\right)\\ &=\frac{4V_0}{a^2}\left\{\frac{1}{2}\int_{-a/4}^{a/4}\mathrm{d}x\left[\sin\left(\frac{3\pi x}{a}\right)+\sin\left(\frac{\pi x}{a}\right)\right]\right\}=0.\end{aligned}$$

Hence this potential does not split the degeneracy. The reason for this is that the potential is still a *separable* potential (it can be split into $x$ and $y$ parts). Since the potential is an even function around the centre of the well, it can only couple two even or two odd wave functions, hence it does not split the lowest level which is composed of even and odd functions.

5. We use the fact that $[a, a^+] = aa^+ - a^+a = 1$ to get

$$\begin{aligned}
x &= \left(\frac{\hbar}{2m\omega}\right)^{1/2} (a + a^+) \\
x^2 &= \frac{\hbar}{2m\omega}(a + a^+)^2 = \frac{\hbar}{2m\omega}(a^2 + (a^+)^2 + 2n + 1) \\
x^4 &= \left(\frac{\hbar}{2m\omega}\right)^2 (a + a^+)^4 \\
&= \left(\frac{\hbar}{2m\omega}\right)^2 [a^4 + (a^+)^4 + 2(2n + 1)(a^2 + (a^+)^2) + 4n^2 + 4n + 1].
\end{aligned}$$

This leads to the matrix elements

$$\begin{aligned}
\langle k|a^4|n\rangle &= \delta_{k,n-4}\sqrt{n(n-1)(n-2)(n-3)} \\
\langle k|(a^+)^4|n\rangle &= \delta_{k,n+4}\sqrt{(n+1)(n+2)(n+3)(n+4)} \\
\langle k|a^2|n\rangle &= \delta_{k,n-2}\ \sqrt{n(n-1)} \\
\langle k|(a^+)^2|n\rangle &= \delta_{k,n+2}\sqrt{(n+1)(n+2)}.
\end{aligned}$$

The matrix has elements on the main diagonal $[4n^2 + 4n + 1]$, and elements given two and four units off the main diagonal, given by the above matrix elements (plus the pre-factor listed above).

6. By the same procedures as the previous problem, the perturbation can be written as

$$\begin{aligned}
V_1 &= \alpha x^3 = \alpha \left(\frac{\hbar}{2m\omega}\right)^{3/2} (a + a^+)^3 \\
&= \alpha \left(\frac{\hbar}{2m\omega}\right)^{3/2} [a^3 + (a^+)^3 + (3n + 1)a + (3n + 2)a^+].
\end{aligned}$$

We note that the $nn$-term is zero, so there is no first-order correction to the energy. Thus, we need only calculate the second-order correction using the matrix elements

$$\begin{aligned}
(V_1)_{kn} &= \alpha \left(\frac{\hbar}{2m\omega}\right)^{3/2} \Big[\delta_{k,n-3}\sqrt{n(n-1)(n-2)} + \delta_{k,n-1}(3n + 1)\sqrt{n} \\
&\qquad + \delta_{k,n+3}\sqrt{(n+1)(n+2)(n+3)} + \delta_{k,n+1}(3n + 2)\sqrt{n+1}\Big]
\end{aligned}$$

which leads to

$$E_n^{(2)} = \sum_{k\neq n} \frac{(V_1)_{nk}(V_1)_{kn}}{E_n^{(0)} - E_k^{(0)}}$$

$$= \alpha^2 \left(\frac{\hbar}{2m\omega}\right)^3 \left\{\frac{n(n-1)(n-2)}{3\hbar\omega} - \frac{(n+1)(n+2)(n+3)}{3\hbar\omega} + \frac{1}{\hbar\omega}[(3n+1)n^2 - (3n+2)^2(n+1)]\right\}$$

$$= \frac{\alpha^2\hbar^2}{8m^3\omega^4}\left\{-\left(\frac{9n^2+13n+6}{3}\right) - [15(n+1)n+4]\right\}.$$

**Chapter 7**

4. From (7.40*d*), we may write the general propagator as

$$U(t,0) = \exp\left[-\frac{\mathrm{i}Ex\tau}{\hbar}\right]$$

for $t > \tau$. Thus, the total wave function is given by

$$\Psi(t) = U(t,0)\varphi_3(x) = \sqrt{\frac{2}{a}}\mathrm{e}^{-\mathrm{i}Ex\tau/\hbar}\sin\left[\frac{3\pi}{a}\left(x+\frac{a}{2}\right)\right]$$

and the connection to an arbitrary state $|k\rangle$ becomes

$$(\varphi_k, \Psi(t)) = \frac{2}{a}\int_{-a/2}^{a/2}\sin\left[\frac{k\pi}{a}\left(x+\frac{a}{2}\right)\right]\left[\cos\left(\frac{E\tau x}{\hbar}\right) - \mathrm{i}\sin\left(\frac{E\tau x}{\hbar}\right)\right] \times \sin\left[\frac{3\pi}{a}\left(x+\frac{a}{2}\right)\right].$$

These integrals may be computed with some care. It should be noted, however, that the field term does not have the periodicity of the quantum well, but does have symmetry properties. Thus, for example, the initial state has even symmetry in the well. Thus, the cosine term in the field will only couple to states that have even symmetry. On the other hand, the sine term will only couple to states that have odd symmetry since it is odd. In general, all states are coupled to the $n = 3$ state by this perturbation.

**Chapter 8**

1. This zero-angular-momentum state requires $n_a = n_b = 1$. Thus, the state is excited as

$$|1,1\rangle = a^+ b^+ |0,0\rangle .$$

The various quantities are given by

$$a^+ = \frac{\mathrm{e}^{\mathrm{i}\varphi}}{2}\left[Br - \frac{1}{B}\frac{\partial}{\partial r} - \frac{\mathrm{i}}{Br}\frac{\partial}{\partial\varphi}\right]$$

$$b^+ = (a^+)^* \qquad |0,0\rangle = \frac{B}{\sqrt{\pi}}\exp\left(-\frac{B^2r^2}{2}\right) \qquad B = \sqrt{\frac{m\omega}{\hbar}}.$$

This leads to

$$|1,1\rangle = \frac{B}{\sqrt{\pi}}[B^2r^2 - 1]\exp\left(-\frac{B^2r^2}{2}\right).$$

2. The perturbation may be found by expanding the frequency in the form $V_1 = m(\omega_0\,\delta\omega + \delta\omega^2)x^2/2$. The operator $x$ may be expressed in terms of the various creation and annihilation operators as

$$x = \sqrt{\frac{\hbar}{2m\omega}}(a_x + a_x^+) = \frac{1}{2}\sqrt{\frac{\hbar}{m\omega}}(a + a^+ + b + b^+).$$

This leads to the crucial part, which is the $x^2$-operator:

$$x^2 = \frac{\hbar}{4m\omega}\big[a^2 + b^2 + (a^+)^2 + (b^+)^2 \\ + 2(ab + ab^+ + a^+b + a^+b^+) + 2(n_a + n_b + 1)\big].$$

We now define the matrix elements between the state $|k,l\rangle$ and the state $|n_a, n_b\rangle$ to be

$$(\bullet)^{kl}_{n_a n_b}$$

and the various terms in $x^2$ can be evaluated as

$$\begin{aligned}
(a^2)^{kl}_{n_a n_b} &= \frac{1}{\sqrt{2}}\delta_{k,n_a-2}\delta_{l,n_b} & (a^{+2})^{kl}_{n_a n_b} &= \frac{1}{\sqrt{2}}\delta_{k,n_a+2}\delta_{l,n_b} \\
(b^2)^{kl}_{n_a n_b} &= \frac{1}{\sqrt{2}}\delta_{k,n_a}\delta_{l,n_b-2} & (b^2)^{kl}_{n_a n_b} &= \frac{1}{\sqrt{2}}\delta_{k,n_a}\delta_{l,n_b+2} \\
(ab)^{kl}_{n_a n_b} &= \delta_{k,n_a-1}\delta_{l,n_b-1} & (a^+b^+)^{kl}_{n_a n_b} &= \delta_{k,n_a+1}\delta_{l,n_b+1} \\
(ab^+)^{kl}_{n_a n_b} &= \delta_{k,n_a-1}\delta_{l,n_b+1} & (a^+b)^{kl}_{n_a n_b} &= \delta_{k,n_a+1}\delta_{l,n_b-1}.
\end{aligned}$$

The term in the number operators is diagonal. Thus, there are a variety of shifting operations in the perturbation, which generally mixes a number of modes.

3. We write the Schrödinger equation with the potential $V = -eEx$ as

$$\left[-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + \frac{m\omega_0^2}{2}(x^2 + y^2) - eEx\right]\psi(x,y) = E\psi(x,y).$$

Introducing the parameter

$$x_0 = \frac{eE}{m\omega_0^2}$$

this may be written as a shifted harmonic oscillator in the form

$$\left[-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + \frac{m\omega_0^2}{2}[(x - x_0)^2 + y^2]\right]\psi(x,y) \\ = \left[E + m\omega_0^2\frac{x_0^2}{2}\right]\psi(x,y)$$

so the new energies are given by

$$E_{n_x n_y} = (n_x + n_y + 1)\hbar\omega_0 - \frac{e^2 E^2}{2m\omega_0^2}.$$

# Index