Preprint typeset in JHEP style - HYPER VERSION

Lent Term, 2017

Applications of Quantum Mechanics

University of Cambridge Part II Mathematical Tripos

David Tong

Department of Applied Mathematics and Theoretical Physics,
Centre for Mathematical Sciences,
Wilberforce Road,
Cambridge, CB3 OBA, UK

http://www.damtp.cam.ac.uk/user/tong/aqm.html
d.tong@damtp.cam.ac.uk

Recommended Books and Resources

There are many good books on quantum mechanics. Here's a selection that I like:

- Griffiths, *Introduction to Quantum Mechanics*

An excellent way to ease yourself into quantum mechanics, with uniformly clear explanations. For this course, it covers both approximation methods and scattering.

- Shankar, *Principles of Quantum Mechanics*
- James Binney and David Skinner, *The Physics of Quantum Mechanics*
- Weinberg, *Lectures on Quantum Mechanics*

These are all good books, giving plenty of detail and covering more advanced topics. Shankar is expansive, Binney and Skinner clear and concise. Weinberg likes his own notation more than you will like his notation, but it's worth persevering.

This course also contains topics that cannot be found in traditional quantum textbooks. This is especially true for the condensed matter aspects of the course, covered in Sections 3, 4 and 5. Some good books include

- Ashcroft and Mermin, *Solid State Physics*
- Kittel, *Introduction to Solid State Physics*
- Steve Simon, *Solid State Physics Basics*

Ashcroft & Mermin and Kittel are the two standard introductions to condensed matter physics, both of which go substantially beyond the material covered in this course. I have a slight preference for the verbosity of Ashcroft and Mermin. The book by Steve Simon covers only the basics, but does so very well. (An earlier draft can be downloaded from his homepage; see below for a link.)

A number of lecture notes are available on the web. Links can be found on the course webpage: http://www.damtp.cam.ac.uk/user/tong/aqm.html

Contents

Acknowledgements

This course is built on the foundation of previous courses, given in Cambridge by Ron Horgan and Nick Dorey. I'm supported by the Royal Society and Alex Considine Tong.

0. Introduction

> "The true meaning of quantum mechanics can be found in the answers it gives about the world we inhabit."
>
> *Me, in a previous set of lecture notes.*

Our previous courses on quantum mechanics were largely focussed on understanding the mathematical formalism of the subject. The purpose of this course is to put this understanding to use.

The applications of quantum mechanics are many and varied, and vast swathes of modern physics fall under this rubric. Here we tell only a few of the possible stories, laying the groundwork for future exploration. There are two major topics.

Much of these lectures is devoted to *condensed matter physics* or, more precisely, *solid state physics*. This is the study of "stuff", of how the wonderfully diverse properties of solids can emerge from the simple laws that govern electrons and atoms. We will develop the basics of the subject, learning how electrons glide through seemingly impenetrable solids, how their collective motion is described by a Fermi surface, and how the vibrations of the underlying atoms get tied into bundles of energy known as phonons. We will learn that electrons in magnetic fields can do strange things and start to explore some of the roles that geometry and topology play in quantum physics.

The second major topic is *scattering theory*. In the past century, physicists have developed a foolproof and powerful method to understand everything and anything: you take the object that you're interested in and you throw something at it. This technique was pioneered by Rutherford who used it to understand the structure of the atom. It was used by Franklin, Crick and Watson to understand the structure of DNA. And, more recently, it was used at the LHC to demonstrate the existence of the Higgs boson. In fact, throwing stuff at other stuff is the single most important experimental method known to science. It underlies much of what we know about condensed matter physics and all of what we know about high-energy physics.

In many ways, these lectures are where theoretical physics starts to fracture into separate sub-disciplines. Yet areas of physics which study systems separated by orders of magnitude — from the big bang, to stars, to materials, to information, to atoms and beyond — all rest on a common language and background. The purpose of these lectures is to build this shared base of knowledge.

1. Scattering Theory

The basic idea behind *scattering theory* is simple: there's an object that you want to understand. So you throw something at it. By analysing how that something bounces off, you can glean information about the object itself.

A very familiar example of scattering theory is called "looking at things". In this section we're going to explore what happens when you look at things by throwing a quantum particle at an object.

1.1 Scattering in One Dimension

We start by considering a quantum particle moving along a line. The maths here will be simple, but the physics is sufficiently interesting to exhibit many of the key ideas.

The object that we want to understand is some potential $V(x)$. Importantly, the potential is localised to some region of space which means that $V(x) \to 0$ as $x \to \pm\infty$. An example is shown to the right. We will need the potential to fall-off to be suitably fast in what follows although, for now, we won't be careful about what this means. A quantum particle moving along the line is governed by the Schrödinger equation,

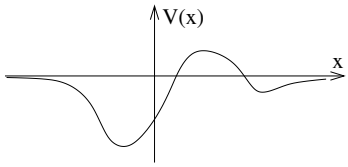


Figure 1:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + V(x)\psi = E\psi \tag{1.1}$$

Solutions to this equation are energy eigenstates. They evolve in time as $\psi(x,t) = e^{-iEt/\hbar}\psi(x)$. For any potential, there are essentially two different kinds of states that we're interested in.

- *Bound States* are states that are localised in some region of space. The wavefunctions are normalisable and have profiles that drop off exponentially far from the potential

  $$\psi(x) \sim e^{-\lambda|x|} \quad \text{as} \quad |x| \to \infty$$

  Because the potential vanishes in the asymptotic region, the Schrödinger equation (1.1) relates the asymptotic fall-off to the energy of the state,

  $$E = -\frac{\hbar^2\lambda^2}{2m} \tag{1.2}$$

  In particular, bound states have $E < 0$. Indeed, it is this property which ensures that the particle is trapped within the potential and cannot escape to infinity.

Bound states are rather special. In the absence of a potential, a solution which decays exponentially to the left will grow exponentially to the far right. But, for the state to be normalisable, the potential has to turn this behaviour around, so the wavefunction decreases at both $x \to -\infty$ and $x \to +\infty$. This will only happen for specific values of λ. Ultimately, this is why the spectrum of bound states is discrete, like in the hydrogen atom. It's where the name "quantum" comes from.

- *Scattering States* are not localised in space and, relatedly, the wavefunctions are not normalisable. Instead, asymptotically, far from the potential, scattering states take the form of plane waves. In one dimension, there are two possibilities

$$\begin{aligned} \text{Right moving:} \quad & \psi \sim e^{ikx} \\ \text{Left moving:} \quad & \psi \sim e^{-ikx} \end{aligned}$$

  where $k > 0$. To see why these are left or right moving, we need to put the time dependence back in. The wavefunctions then take the form $e^{\pm ikx - iEt/\hbar}$. The peaks and troughs of the wave move to the right with the plus sign, and to the left with the minus sign. Solving the Schrödinger equation in the asymptotic region with $V = 0$ gives the energy

$$E = \frac{\hbar^2 k^2}{2m}$$

  Scattering states have $E > 0$. Note that, in contrast, to bound states, nothing special has to happen to find scattering solutions. We expect to find solutions for any choice of k.

This simple classification of solutions already tells us something interesting. Suppose, for example, that the potential looks something like the one shown in the figure. You might think that we could find a localised solution that is trapped between the two peaks, with $E > 0$. But this can't happen because if the wavefunction is to be normalisable, it must have $E < 0$. The physical reason, of course, is quantum tunnelling which allows the would-be bound state to escape to infinity. We will learn more about this situation in Section 1.1.5.

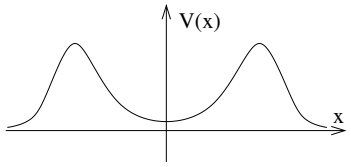


Figure 2:

1.1.1 Reflection and Transmission Amplitudes

Suppose that we stand a long way from the potential and throw particles in. What comes out? This is answered by solving the Schrödinger equation for the scattering

states. Because we have a second order differential equation, we expect that there are two independent solutions for each value of k. We can think of these solutions physically as what you get if you throw the particle in from the left or in from the right. Let's deal with each in turn.

Scattering from the Left

We throw the particle in from the left. When it hits the potential, one of two things can happen: it can bounce back, or it can pass straight through. Of course, this being quantum mechanics, it can quite happily do both at the same time. Mathematically, this means that we are looking for a solution which asymptotically takes the form

$$\psi_R(x) \sim \begin{cases} e^{ikx} + re^{-ikx} & x \to -\infty \\ te^{ikx} & x \to +\infty \end{cases} \tag{1.3}$$

We've labelled this state ψ_R because the ingoing wave is *right*-moving. This can be seen in the first term e^{ikx} which represents the particle we're throwing in from $x \to -\infty$. The second term re^{-ikx} represents the particle that is reflected back to $x \to -\infty$ after hitting the potential. The coefficient $r \in \mathbf{C}$ is called the *reflection amplitude*. Finally, the term te^{ikx} at $x \to +\infty$ represents the particle passing through the potential. The coefficient $t \in \mathbf{C}$ is called the *transmission coefficient*. (Note: in this formula t is a complex number that we have to determine; it is not time!) There is no term e^{-ikx} at $x \to +\infty$ because we're not throwing in any particles from that direction. Mathematically, we have chosen the solution in which this term vanishes.

Before we proceed, it's worth flagging up a conceptual point. Scattering is clearly a dynamical process: the particle goes in, and then comes out again. Yet there's no explicit time dependence in our ansatz (1.3); instead, we have a solution formed of plane waves, spread throughout all of space. It's best to think of these plane waves as describing a beam of particles, with the ansatz (1.3) giving us the steady-state solution in the presence of the potential.

The probability for reflection R and transmission T are given by the usual quantum mechanics rule:

$$R = |r|^2 \quad \text{and} \quad T = |t|^2$$

In general, both R and T will be functions of the wavenumber k. This is what we would like to calculate for a given potential and we will see an example shortly. But, before we do this, there are some observations that we can make using general statements about quantum mechanics.

Given a solution $\psi(x)$ to the Schrödinger equation, we can construct a conserved probability current

$$J(x) = -i\frac{\hbar}{2m}\left(\psi^\star\frac{d\psi}{dx} - \psi\frac{d\psi^\star}{dx}\right)$$

which obeys $dJ/dx = 0$. This means that $J(x)$ is constant. (Mathematically, this is the statement that the Wronskian is constant for the two solutions to the Schrödinger equation). For our scattering solution ψ_R, with asymptotic form (1.3), the probability current as $x \to -\infty$ is given by

$$\begin{aligned} J(x) &= \frac{\hbar k}{2m}\Big[\left(e^{-ikx} + r^\star e^{+ikx}\right)\left(e^{ikx} - re^{-ikx}\right) + \left(e^{ikx} + re^{-ikx}\right)\left(e^{-ikx} - r^\star e^{+ikx}\right)\Big] \\ &= \frac{\hbar k}{m}\left(1 - |r|^2\right) \qquad \text{as} \quad x \to -\infty \end{aligned}$$

Meanwhile, as $x \to +\infty$, we have

$$J(x) = \frac{\hbar k}{m}|t|^2 \qquad \text{as} \quad x \to +\infty$$

Equating the two gives

$$1 - |r|^2 = |t|^2 \quad \Rightarrow \quad R + T = 1 \tag{1.4}$$

This should make us happy as it means that probabilities do what probabilities are supposed to do. The particle can only get reflected or transmitted and the sum of the probabilities to do these things equals one.

Scattering from the Right

This time, we throw the particle in from the right. Once again, it can bounce back off the potential or pass straight through. Mathematically, we're now looking for solutions which take the asymptotic form

$$\psi_L(x) \sim \begin{cases} t'e^{-ikx} & x \to -\infty \\ e^{-ikx} + r'e^{+ikx} & x \to +\infty \end{cases} \tag{1.5}$$

where we've now labelled this state ψ_L because the ingoing wave, at $x \to +\infty$, is *left*-moving. We've called the reflection and transmission amplitudes r' and t'.

There is a simple relation between the two solutions ψ_R in (1.3) and ψ_L in (1.5). This follows because the potential $V(x)$ in (1.1) is a real function, so if ψ_R is a solution then so is $\psi_R^\star$. And, by linearity, so is $\psi_R^\star - r^\star\psi_R$ which is given by

$$\psi_R^\star(x) - r^\star\psi_R(x) \sim \begin{cases} (1-|r|^2)e^{-ikx} & x \to -\infty \\ t^\star e^{-ikx} - r^\star t e^{ikx} & x \to +\infty \end{cases}$$

This takes the same functional form as (1.5) except we need to divide through by $t^\star$ to make the normalisations agree. (Recall that scattering states aren't normalised anyway so we're quite at liberty to do this.) Using $1-|r|^2 = |t|^2$, this tells us that there is a solution of the form (1.5) with

$$t' = t \quad \text{and} \quad r' = -\frac{r^\star t}{t^\star} \tag{1.6}$$

Notice that the transition amplitudes are always the same, but the reflection amplitudes can differ by a phase. Nonetheless, this is enough to ensure that the reflection probabilities are the same whether we throw the particle from the left or right: $R = |r|^2 = |r'|^2$.

An Example: A Pothole in the Road

Let's compute r and t for a simple potential, given by

$$V(x) = \begin{cases} -V_0 & -a/2 < x < a/2 \\ 0 & \text{otherwise} \end{cases}$$

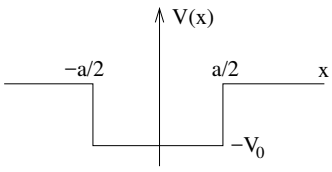


Figure 3:

with $V_0 > 0$. This looks like a pothole in the middle of an, otherwise, flat potential.

Outside the potential, we have the usual plane waves $\psi \sim e^{\pm ikx}$. In the middle of the potential, the solutions to the Schrödinger equation (1.1) take the form

$$\psi(x) = Ae^{iqx} + Be^{-iqx} \qquad x \in [-a/2, a/2] \tag{1.7}$$

where

$$q^2 = \frac{2mV_0}{\hbar^2} + k^2$$

To compute the reflection and transmission amplitudes, r, r' and t, we need to patch the solution (1.7) with either (1.3) or (1.5) at the edges of the potential.

Let's start by scattering from the left, with the solution (1.3) outside the potential. Continuity of the wavefunction at $x = \pm a/2$ tells us that

$$e^{-ika/2} + re^{ika/2} = Ae^{-iqa/2} + Be^{iqa/2} \quad \text{and} \quad te^{ika/2} = Ae^{iqa/2} + Be^{-iqa/2}$$

Meanwhile, matching the derivatives of ψ at $x = \pm a/2$ gives

$$\frac{k}{q}\left(e^{-ika/2} - re^{ika/2}\right) = Ae^{-iqa/2} - Be^{iqa/2} \quad \text{and} \quad \frac{kt}{q}e^{ika/2} = Ae^{iqa/2} - Be^{-iqa/2}$$

These are four equations with four unknowns: A, B, r and t. One way to proceed is to add and subtract the two equations on the right, and then do the same for the two equations on the left. This allows us to eliminate A and B

$$A = t\left(1 + \frac{k}{q}\right)e^{i(k-q)a/2} = \left(1 + \frac{k}{q}\right)e^{-i(k-q)a/2} + r\left(1 - \frac{k}{q}\right)e^{i(k+q)a/2}$$

$$B = t\left(1 - \frac{k}{q}\right)e^{i(k+q)a/2} = \left(1 - \frac{k}{q}\right)e^{-i(k+q)a/2} + r\left(1 + \frac{k}{q}\right)e^{i(k-q)a/2}$$

We've still got some algebraic work ahead of us. It's grungy but straightforward. Solving these two remaining equations gives us the reflection and transmission coefficients that we want. They are

$$\begin{aligned} r &= \frac{(k^2 - q^2)\sin(qa)e^{-ika}}{(q^2 + k^2)\sin(qa) + 2iqk\cos(qa)} \\ t &= \frac{2iqke^{-ika}}{(q^2 + k^2)\sin(qa) + 2iqk\cos(qa)} \end{aligned} \tag{1.8}$$

Even for this simple potential, the amplitudes are far from trivial. Indeed, they contain a lot of information. Perhaps the simplest lesson we can extract comes from looking at the limit $k \to 0$, where $r \to -1$ and $t \to 0$. This means that if you throw the particle very softly ($k \to 0$), then it won't make it through the potential; it's guaranteed to bounce back.

Conversely, in the limit $k \to \infty$, we have $r = 0$. (Recall that $q^2 = k^2 + 2mV_0/\hbar^2$ so we also have $q \to \infty$ in this limit.) By conservation of probability, we must then have $|t| = 1$ and the particle is guaranteed to pass through. This is what you might expect; if you throw the particle hard enough, it barely notices that the potential is there.

There are also very specific values of the incoming momenta for which $r = 0$ and the particle is assured of passage through the potential. This occurs when $qa = n\pi$ with $n \in \mathbf{Z}$ for which $r = 0$. Notice that you have to fine tune the incoming momenta so that it depends on the details of the potential which, in this example, means V_0 and a.

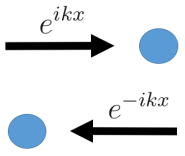
We can repeat the calculation above for scattering from the right. In fact, for our pothole potential, the result is exactly the same and we have $r = r'$. This arises because $V(x) = V(-x)$ so it's no surprise that scattering from the left and right are the same. We'll revisit this in Section 1.1.3.

1.1.2 Introducing the S-Matrix

The *S-matrix* is a convenient way of packaging the information about reflection and transmission coefficients. It is useful both because it highlights new features of the problem, and because it generalises to scattering in higher dimensions.

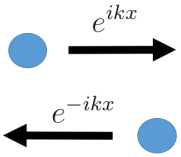
We will start by writing the above solutions in slightly different notation. We have two ingoing asymptotic wavefunctions, one from the left and one from the right

$$\text{Ingoing} \begin{cases} \text{right-moving:} & \mathcal{I}_R(x) = e^{+ikx} \quad x \to -\infty \\ \text{left-moving:} & \mathcal{I}_L(x) = e^{-ikx} \quad x \to +\infty \end{cases}$$



Similarly, there are two outgoing asymptotic wavefunctions,

$$\text{Outgoing} \begin{cases} \text{right-moving:} & \mathcal{O}_R(x) = e^{+ikx} \quad x \to +\infty \\ \text{left-moving:} & \mathcal{O}_L(x) = e^{-ikx} \quad x \to -\infty \end{cases}$$



The two asymptotic solutions (1.3) and (1.5) can then be written as

$$\begin{pmatrix} \psi_R \\ \psi_L \end{pmatrix} = \begin{pmatrix} \mathcal{I}_R \\ \mathcal{I}_L \end{pmatrix} + \mathcal{S} \begin{pmatrix} \mathcal{O}_R \\ \mathcal{O}_L \end{pmatrix} \tag{1.9}$$

where

$$\mathcal{S} = \begin{pmatrix} t & r \\ r' & t' \end{pmatrix} \tag{1.10}$$

This is the *S-matrix*. As we've seen, for any given problem the entries of the matrix are rather complicated functions of k.

The S-matrix has many nice properties, some of which we will describe in these lectures. One of the simplest and most important is that $\mathcal{S}$ is unitary. To see this note that

$$\mathcal{S}\mathcal{S}^\dagger = \begin{pmatrix} |t|^2 + |r|^2 & tr'^\star + rt'^\star \\ t^\star r' + t' r^\star & |t'|^2 + |r'|^2 \end{pmatrix}$$

Unitarity then follows from the conservation of probability. The off-diagonal elements vanish by virtue of the relations $t' = t$ and $r' = -r^\star t/t^\star$ that we found in (1.6). Meanwhile, the diagonal elements are equal to one by (1.4) and so $\mathcal{S}\mathcal{S}^\dagger = \mathbf{1}$. The equivalence between conservation of probability and unitarity of the S-matrix is important, and will generalise to higher dimensions. Indeed, in quantum mechanics the word "unitarity" is often used synonymously with "conservation of probability".

One further property follows from the fact that the wavefunctions $\psi_R(x)$ and $\psi_L(x)$ do not change under complex conjugation if we simultaneously flip $k \to -k$. In other words $\psi(x;k) = \psi^\star(x;-k)$. This means that the S-matrix obeys

$$\mathcal{S}^\star(k) = \mathcal{S}(-k)$$

There are a number of other, more hidden properties of the S-matrix that we will uncover below.

1.1.3 A Parity Basis for Scattering

As we've seen above, for symmetric potentials, with $V(x) = V(-x)$, scattering from the left and right is the same. Let's first make this statement more formal.

We introduce the *parity* operator P which acts on functions $f(x)$ as

$$P : f(x) \to f(-x)$$

For symmetric potentials, we have $[P, H] = 0$ which means that eigenstates of the Hamiltonian can be chosen so that they are also eigenstates of P. The parity operator is Hermitian, $P^\dagger = P$, so its eigenvalues λ are real. But we also have $P^2 f(x) = f(x)$, which means that the eigenvalues must obey $\lambda^2 = 1$. Clearly there are only two possibilities: $\lambda = +1$ and $\lambda = -1$, This means that eigenstates of the Hamiltonian can be chosen to be either even functions ($\lambda = +1$) or odd functions ($\lambda = -1$).

Above we worked with scattering eigenstates ψ_R and ψ_L. These are neither odd nor even. Instead, for a symmetric potential, they are related by $\psi_L(x) = \psi_R(-x)$. This is the reason that symmetric potentials have $r = r'$. If we want to work with the parity eigenstates, we take

$$\begin{aligned}\psi_+(x) &= \psi_R(x) + \psi_L(x) = \psi_R(x) + \psi_R(-x) \\ \psi_-(x) &= -\psi_R(x) + \psi_L(x) = -\psi_R(x) + \psi_R(-x)\end{aligned}$$

which obey $P\psi_\pm(x) = \pm\psi_\pm(x)$.

Often, working with parity eigenstates makes the algebra a little easier. This is particularly true if our problem has a parity-invariant potential, $V(x) = V(-x)$.

The Pothole Example Revisited

Let's see how the use of parity eigenstates can make our calculations simpler. We'll redo the scattering calculation in the pothole, but now we'll take the asymptotic states to be ψ_+ and ψ_-. Physically, you can think of this experiment as throwing in particles from both the left and right at the same time, with appropriate choices of signs.

We start with the even parity wavefunction ψ_+. We want to patch this onto a solution in the middle, but this too must have even parity. This mean that the solution in the pothole takes the form

$$\psi_+(x) = A(e^{iqx} + e^{-iqx}) \qquad x \in [-a/2, a/2]$$

which now has only one unknown coefficient, A. As previously, $q^2 = k^2 + 2mV_0/\hbar^2$. We still need to make sure that both the wavefunction and its derivative are continuous at $x = \pm a/2$. But, because we're working with even functions, we only need to look at one of these points. At $x = a/2$ we get

$$\begin{aligned}e^{-ika/2} + (r+t)e^{ika/2} &= A(e^{iqa/2} + e^{-iqa/2}) \\ \left(-e^{-ika/2} + (r+t)e^{ika/2}\right) &= \frac{q}{k}A(e^{iqa/2} - e^{-iqa/2})\end{aligned}$$

Notice that only the combination $(r + t)$ appears. We have two equations with two unknowns. If we divide the two equations and rearrange, we get

$$r + t = -e^{-ika}\,\frac{q\tan(qa/2) - ik}{q\tan(qa/2) + ik} \tag{1.11}$$

which is all a lot easier than the messy manipulations we had to do when working with ψ_L and ψ_R. Of course, we've only got an expression for $(r + t)$. But we can play the

same game for the odd parity eigenstates to get a corresponding expression for $(r-t)$. Now, the solution in the pothole takes the form

$$\psi_-(x) = B(e^{iqx} - e^{-iqx}) \qquad x \in [-a/2, a/2]$$

Requiring continuity of the wavefunction and its derivative at $x = a/2$ we get

$$\begin{aligned} e^{-ika/2} + (r-t)e^{ika/2} &= B(e^{iqa/2} - e^{-iqa/2}) \\ \left(-e^{-ika/2} + (r-t)e^{ika/2}\right) &= \frac{q}{k}B(e^{iqa/2} + e^{-iqa/2}) \end{aligned}$$

Once again, dividing we find

$$r - t = e^{-ika}\frac{q + ik\tan(qa/2)}{q - ik\tan(qa/2)} \tag{1.12}$$

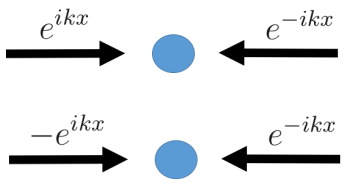
It's not immediately obvious that the expressions (1.11) and (1.12) are the same as those for r and t that we derived previously. But a little bit of algebra should convince you that they agree.

[A helping hand: this little bit of algebra is extremely fiddly if you don't go about it in the right way! Here's a reasonably a streamlined approach. First define the denominator of (1.8) as $D(k) = (q^2+k^2)\sin(qa) + 2iqk\cos(qa)$. Using the double-angle formula from trigonometry, we can write this as $D(k) = 2\cos^2(qa/2)(q\tan(qa/2) + ik)(q - ik\tan(qa/2))$. We can then add the two expressions in (1.8), and use the double-angle formula again, to get $r + t = 2e^{-ika}\cos^2(qa/2)(q\tan(qa/2) - ik)(ik\tan(qa/2) - q)/D(k)$ This coincides with our formula (1.11). Similar games give us the formula (1.12).]

The S-Matrix in the Parity Basis

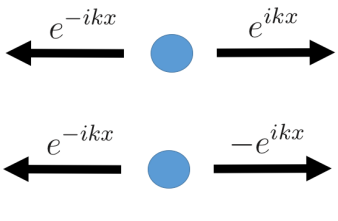
We can also think about the S-matrix using our new basis of states. The asymptotic ingoing modes are even and odd functions, given at $|x| \to \infty$ by

Ingoing
- parity-even: $\mathcal{I}_+(x) = e^{-ik|x|}$
- parity-odd: $\mathcal{I}_-(x) = \text{sign}(x)\, e^{-ik|x|}$



The two asymptotic outgoing modes are

$$\text{Outgoing} \begin{cases} \text{parity-even:} & \mathcal{O}_+(x) = e^{+ik|x|} \\ \text{parity-odd:} & \mathcal{O}_-(x) = -\text{sign}(x)\, e^{+ik|x|} \end{cases}$$



These are related to our earlier modes by a simple change of basis,

$$\begin{pmatrix} \mathcal{I}_+ \\ \mathcal{I}_- \end{pmatrix} = \mathcal{M} \begin{pmatrix} \mathcal{I}_R \\ \mathcal{I}_L \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mathcal{O}_+ \\ \mathcal{O}_- \end{pmatrix} = \mathcal{M} \begin{pmatrix} \mathcal{O}_R \\ \mathcal{O}_L \end{pmatrix} \quad \text{with} \quad \mathcal{M} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

We can define an S-matrix with respect to this parity basis. In analogy with (1.9), we write asymptotic solutions as

$$\begin{pmatrix} \psi_+ \\ \psi_- \end{pmatrix} = \begin{pmatrix} \mathcal{I}_+ \\ \mathcal{I}_- \end{pmatrix} + \mathcal{S}^P \begin{pmatrix} \mathcal{O}_+ \\ \mathcal{O}_- \end{pmatrix} \tag{1.13}$$

where we use the notation $\mathcal{S}^P$ to denote the S-matrix with respect to the parity basis. We write

$$\mathcal{S}^P = \begin{pmatrix} S_{++} & S_{+-} \\ S_{-+} & S_{--} \end{pmatrix}$$

This is related to our earlier S-matrix by a change of basis. We have

$$\mathcal{S}^P = \mathcal{M}S\mathcal{M}^{-1} = \begin{pmatrix} t + (r+r')/2 & (r-r')/2 \\ (r'-r)/2 & t - (r+r')/2 \end{pmatrix}$$

As you may expect, this basis is particularly useful if the underlying potential is symmetric, so $V(x) = V(-x)$. In this case we have $r = r'$ and the S-matrix becomes diagonal. The diagonal components are simply

$$S_{++} = t + r \quad \text{and} \quad S_{--} = t - r$$

In fact, because $\mathcal{S}^p$ is unitary, each of these components must be a phase. This follows because r and t are not independent. First, they obey $|r|^2 + |t|^2 = 1$. Moreover, when $r' = r$, the relation (1.6) becomes

$$rt^\star + r^\star t = 0 \qquad \Rightarrow \qquad \text{Re}(rt\star) = 0$$

This is enough to ensure that both S_{++} and S_{--} are indeed phases. We write them as

$$S_{++} = e^{2i\delta_+(k)} \quad \text{and} \quad S_{--} = e^{2i\delta_-(k)}$$

We learn that for scattering off a symmetric potential, all the information is encoded in two momentum-dependent *phase shifts*, $\delta_\pm(k)$ which tell us how the phases of the outgoing waves $\mathcal{O}_\pm$ are changed with respect to the ingoing waves $\mathcal{I}_\pm$.

1.1.4 Bound States

So far we've focussed only on the scattering states of the problem. We now look at the bound states, which have energy $E < 0$ and are localised near inside the potential. Here, something rather magical happens. It turns out that the information about these bound states can be extracted from the S-matrix, which we constructed purely from knowledge of the scattering states.

To find the bound states, we need to do something clever. We take our scattering solutions, which depend on momentum $k \in \mathbf{R}$, and extend them to the complex momentum plane. This means that we analytically continue out solutions so that they depend on $k \in \mathbf{C}$.

First note that the solutions with $k \in \mathbf{C}$ still obey our original Schrödinger equation (1.1) since, at no point in any of our derivation did we assume that $k \in \mathbf{R}$. The only difficulty comes when we look at how the wavefunctions behave asymptotically. In particular, any putative solution will, in general, diverge exponentially as $x \to +\infty$ or $x \to -\infty$, rendering the wavefunction non-normalisable. However, as we will now show, there are certain solutions that survive.

For simplicity, let's assume that we have a symmetric potential $V(x) = V(-x)$. As we've seen above, this means that there's no mixing between the parity-even and parity-odd wavefunctions. We start by looking at the parity-even states. The general solution takes the form

$$\psi_+(x) = \mathcal{I}_+(x) + S_{++}\mathcal{O}_+(x) = \begin{cases} e^{+ikx} \;+\; S_{++}e^{-ikx} & x \to -\infty \\ e^{-ikx} \;+\; S_{++}e^{+ikx} & x \to +\infty \end{cases}$$

Suppose that we make k pure imaginary and write

$$k = i\lambda$$

with $\lambda > 0$. Then we get

$$\psi_+(x) = \begin{cases} e^{-\lambda x} \;+\; S_{++}e^{+\lambda x} & x \to -\infty \\ e^{+\lambda x} \;+\; S_{++}e^{-\lambda x} & x \to +\infty \end{cases} \tag{1.14}$$

Both terms proportional to S_{++} decay asymptotically, but the other terms diverge. This is bad. However, there's a get-out. For any fixed k (whether real or complex), S_{++} is simply a number. That means that we're quite at liberty to divide by it. Indeed, the wavefunction above isn't normalised anyway, so dividing by a constant isn't going to change anything. We get

$$\psi_+(x) = \begin{cases} S_{++}^{-1}\, e^{-\lambda x} \;+\; e^{+\lambda x} & x \to -\infty \\ S_{++}^{-1}\, e^{+\lambda x} \;+\; e^{-\lambda x} & x \to +\infty \end{cases} \tag{1.15}$$

Now we can see the loop-hole. The wavefunction above is normalisable whenever we can find a $\lambda > 0$ such that

$$S_{++}(k) \to \infty \quad \text{as } k \to i\lambda$$

This, then, is the magic of the S-matrix. Poles in the complex momentum plane that lie on the positive imaginary axis (i.e. $k = i\lambda$ with $\lambda > 0$) correspond to bound states. This information also tells us the energy of the bound state since, as we saw in (1.2), it is given by

$$E = -\frac{\hbar^2\lambda^2}{2m}$$

We could also have set $k = -i\lambda$, with $\lambda > 0$. In this case, it is the terms proportional to S_{++} in (1.14) which diverge and the wavefunction is normalisable only if $S_{++}(k = -i\lambda) = 0$. However, since S_{++} is a phase, this is guaranteed to be true whenever $S_{++}(k = i\lambda)$ has a pole, and simply gives us back the solution above.

Finally, note that exactly the same arguments hold for parity-odd wavefunctions. There is a bound state whenever $S_{--}(k)$ has a pole at $k = i\lambda$ with $\lambda > 0$.

An Example: Stuck in the Pothole

We can illustrate this with our favourite example of the square well, of depth $-V_0$ and width a. We already computed the S-matrix in (1.11) and (1.12). We have,

$$S_{++}(k) = r + t = -e^{-ika}\, \frac{q\tan(qa/2) - ik}{q\tan(qa/2) + ik}$$

where $q^2 = 2mV_0/\hbar^2 + k^2$. Setting $k = i\lambda$, we see that this has a pole when

$$\lambda = q\tan\left(\frac{qa}{2}\right) \quad \text{with} \quad \lambda^2 + q^2 = \frac{2mV_0}{\hbar^2}$$

These are the usual equations that you have to solve when finding parity-even bound states in a square well. The form of the solutions is simplest to see if we plot these equations, as shown in the left-hand of Figure 4. There is always at least one bound state, with more appearing as the well gets deeper.

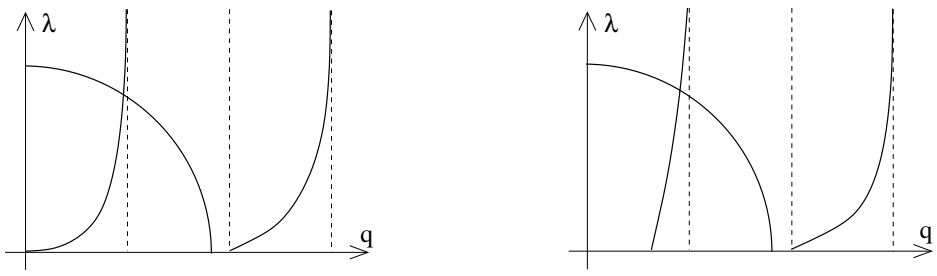


Figure 4: Bound state of even parity always exist, since the two equations shown on the left always have a solution with $\lambda, q > 0$. Bound states of odd parity, shown on the right, exist if the potential is deep enough.

Similarly, if we look at the parity-odd wavefunctions, we have

$$S_{--}(k) = t - r = e^{-ika}\frac{q + ik\tan(qa/2)}{q - ik\tan(qa/2)}$$

which has a pole at $k = i\lambda$ when

$$q = -\lambda\tan\left(\frac{qa}{2}\right) \quad \text{with} \quad \lambda^2 + q^2 = \frac{2mV_0}{\hbar^2} \tag{1.16}$$

This too reproduces the equations that we found in earlier courses in quantum mechanics when searching for bound states in a square well. Now there is no guarantee that a bound state exists; this only happens if the potential is deep enough.

1.1.5 Resonances

We might wonder if there's any other information hidden in the analytic structure of the S-matrix. In this section, we will see that there is, although its interpretation is a little more subtle.

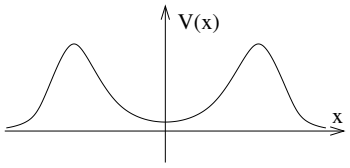


Figure 5:

First, the physics. Let's think back again to the example shown on the right. One the one hand, we know that there can be no bound states in such a trap because they will have $E > 0$. Any particle that we place in the trap will ultimately tunnel out. On the other hand, if the walls of the trap are very large then we might expect that the particle stays there for a long time before it eventually escapes. In this situation, we talk of a *resonance*. These are also referred to as *unstable* or *metastable* states. Our goal is to show how such resonances are encoded in the S-matrix.

Now, the maths. We'll restrict attention to parity-even functions. Suppose that the S-matrix S_{++} has a pole that lies on the complex momentum plane at position

$$k = k_0 - i\gamma$$

We'd like to interpret this pole. First note that the energy is also imaginary

$$E = \frac{\hbar^2 k^2}{2m} \equiv E_0 - i\frac{\Gamma}{2} \tag{1.17}$$

with $E_0 = \hbar^2(k_0^2 - \gamma^2)/2m$ and $\Gamma = 2\hbar^2\gamma k_0/m$. An imaginary energy may sound strange, but it is has a very natural interpretation. Recall that the time dependence of the wavefunction is given by

$$e^{-iEt/\hbar} = e^{-iE_0 t/\hbar}\, e^{-\Gamma t/2\hbar} \tag{1.18}$$

This is the first clue that we need. We see that, for $\gamma > 0$, the overall form of the wavefunction decays exponentially with time. This is the characteristic behaviour of unstable states. A wavefunction that is initially supported inside the trap will be very small there at time much larger than $\tau = 1/\Gamma$. Here τ is called the *half-life* of the state, while Γ is usually referred to as the *width* of the state. (We'll see why in Section 1.2).

Where does the particle go? Including the time dependence (1.18), the same argument that led us to (1.15) now tells us that when $S_{++} \to \infty$, the solution takes the asymptotic form

$$\psi_+(x,t) = \begin{cases} e^{-iE_0 t/\hbar}\, e^{-ik_0 x}\, e^{-\gamma x - \Gamma t/2\hbar} & x \to -\infty \\ e^{-iE_0 t/\hbar}\, e^{+ik_0 x}\, e^{+\gamma x - \Gamma t/2\hbar} & x \to +\infty \end{cases} \tag{1.19}$$

The first two exponential factors oscillate. But the final factor varies as

$$e^{\pm\gamma(x \mp vt)} \quad \text{where} \quad v = \frac{\Gamma}{2\hbar\gamma} = \frac{\hbar k_0}{m}$$

This has the interpretation of a particle moving with momentum $\hbar k_0$. This, of course, is the particle which has escaped the trap.

Note that for fixed time t, these wavefunctions are not normalisable: they diverge at both $x \to \pm\infty$. This shouldn't concern us, because, although our wavefunctions are eigenstates of the Hamiltonian, they are not interpreted as stationary states. Indeed, it had to be the case. An unstable state has complex energy, but standard theorems in linear algebra tell us that a Hermitian operator like the Hamiltonian must have real eigenvalues. We have managed to evade this theorem only because these wavefunctions are non-normalisable and so do not, strictly speaking, live in the Hilbert space.

There's a lesson buried in all of this. If we were to take the standard axioms of quantum mechanics, we would simply throw away wavefunctions of the form (1.19) on the grounds that they do not lie in the Hilbert space and so are unphysical. But this would be a mistake: the wavefunctions do contain interesting physics, albeit of a slightly different variety than we are used to. Sometimes it's worth pushing our physical theories beyond our comfort zone to see what is lurking there.

The upshot of this discussion is that poles of the S-matrix in the lower-half complex plane correspond to resonances. It is often useful to write S_{++} as a function of energy rather than momentum. (They are related by (1.17)). Since S_{++} is a phase, close to a resonance it necessarily takes the form

$$S_{++} = \frac{E - E_0 - i\Gamma/2}{E - E_0 + i\Gamma/2}$$

The fact that the S-matrix is a phase means that any pole in the complex energy plane necessarily comes with a zero at the conjugate point.

An Example: A Pair of Delta-Functions

A pair of delta functions provide a simple and tractable example to illustrate the idea of resonances. The potential is given by

$$V(x) = V_0\Big[\delta(x-1) + \delta(x+1)\Big]$$

Recall that the effect of the delta-functions is simply to change the boundary conditions at $x = \pm 1$ when solving the Schrödinger equation. All wavefunctions should be continuous at $x = \pm 1$, but their derivatives are discontinuous. For example, at $x = +1$, solutions obey

$$\lim_{\epsilon\to 0}\ \Big[\psi'(1+\epsilon) - \psi'(1-\epsilon)\Big] = U_0\psi(1) \quad \text{with} \quad U_0 = \frac{2mV_0}{\hbar^2}$$

Working in the parity basis makes life simpler, not least because you only need to consider the matching at one of the delta-functions, with the other then guaranteed. The computation of the S-matrix is a problem on the exercise sheet. You will find

$$S_{++} = e^{-2ik}\left[\frac{(2k - iU_0)e^{ik} - iU_0e^{-ik}}{(2k + iU_0)e^{-ik} + iU_0e^{ik}}\right]$$

Note that the denominator is the complex conjugate of the numerator, ensuring that S_{++} is a phase, as expected. The poles of this S-matrix are given by solutions to the equation

$$e^{2ik} = -\left(1 - \frac{2ik}{U_0}\right) \tag{1.20}$$

To understand the physics behind this, let's first look at the situation where $U_0 \to \infty$, so that the weight of the delta-functions gets infinitely large. Then the poles sit at

$$e^{2ik} = -1 \quad \Rightarrow \quad k = k_n = \left(n + \frac{1}{2}\right)\pi$$

These correspond to bound states trapped between the two wavefunctions. For example, the $n = 0$ state is shown in the figure. Note that they're rather unusual because the poles sit on the real k-axis, rather than the imaginary k-axis. Correspondingly, these bound states have $E > 0$. This strange behaviour is only allowed because we have an infinitely large potential which forbids particles on one side of the barrier to cross to the other.

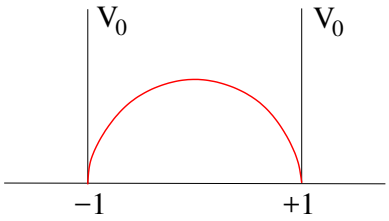


Figure 6:

As a side remark, we note that this same impenetrable behaviour is seen in scattering. When $U_0 \to \infty$, the S-matrix becomes $S_{++} \to -e^{2ik}$. This tells us that a particle coming from outside is completely reflected off the infinitely large barrier. The minus sign is the standard phase change after reflection. The factor of e^{2ik} is because the waves are forbidden from travelling through the region between the delta functions, which has width $x = 2$. As a result, the phase is shifted by e^{ikx} from what it would be if the barriers were removed.

Let's now look at what happens when U_0 is large, but finite? We'll focus on the lowest energy bound state with $n = 0$. We can expand (1.20) in $1/U_0$. (This too is left as a problem on the exercise sheet.) We find

$$k = \frac{\pi}{2} + \alpha - i\gamma$$

with

$$\alpha \approx -\frac{\pi}{2U_0} + \frac{\pi}{2U_0^2} + \mathcal{O}\left(\frac{1}{U_0^3}\right) \quad \text{and} \quad \gamma \approx \frac{\pi^2}{4U_0^2} + \mathcal{O}\left(\frac{1}{U_0^3}\right)$$

Note, in particular, that $\gamma > 0$, so the pole moves off the real axis and into the lower half-plane. This pole now has all the properties that we described at the beginning of this section. It describes a state, trapped between the two delta-functions, which decays with half-life

$$\tau = \frac{\hbar}{\Gamma} = \frac{4mU_0^2}{\hbar\pi^3}\left(1 + \mathcal{O}\left(\frac{1}{U_0}\right)\right)$$

This is the resonance.

1.2 Scattering in Three Dimensions

Our real interest in scattering is for particles moving in three spatial dimensions, with Hamiltonian

$$H = \frac{\mathbf{p}^2}{2m} + V(\mathbf{r})$$

Recall that there are two distinct interpretations for such a Hamiltonian

- We could think of this as the motion of a single particle, moving in a fixed background potential $V(\mathbf{r})$. This would be appropriate, for example, in Rutherford's famous experiment where we fire an alpha particle at a gold nucleus.

- Alternatively, We could think of this as the relative motion of two particles, separated by distance $\mathbf{r}$, interacting through the force $\mathbf{F} = -\nabla V(\mathbf{r})$. We could take $V(r)$ to be the Coulomb force, to describe the scattering of electrons, or the Yukawa force to describe the scattering of neutrons.

In this section, we will use language appropriate to the first interpretation, but everything we say holds equally well in the second. Throughout this section, we will work with rotationally invariant (i.e. central) potentials, so that $V(\mathbf{r}) = V(|\mathbf{r}|)$.

1.2.1 The Cross-Section

Our first goal is to decide what we want to calculate. The simple reflection and transmission coefficients of the one-dimensional problem are no longer appropriate. We need to replace them by something a little more complicated. We start by thinking of the classical situation.

Classical Scattering

Suppose that we throw in a single particle with kinetic energy E. Its initial trajectory is characterised by the *impact parameter* b, defined as the closest the particle would get to the scattering centre at $r = 0$ if there were no potential. The particle emerges with *scattering angle* θ, which is the angle between the asymptotic incoming and outgoing trajectories, as shown in the figure. By solving the classical equations of motion, we can compute $\theta(b; E)$ or, equivalently, $b(\theta; E)$.

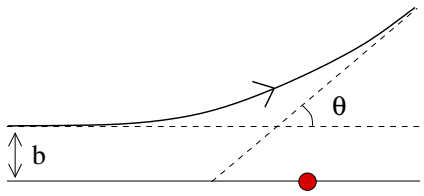


Figure 7:

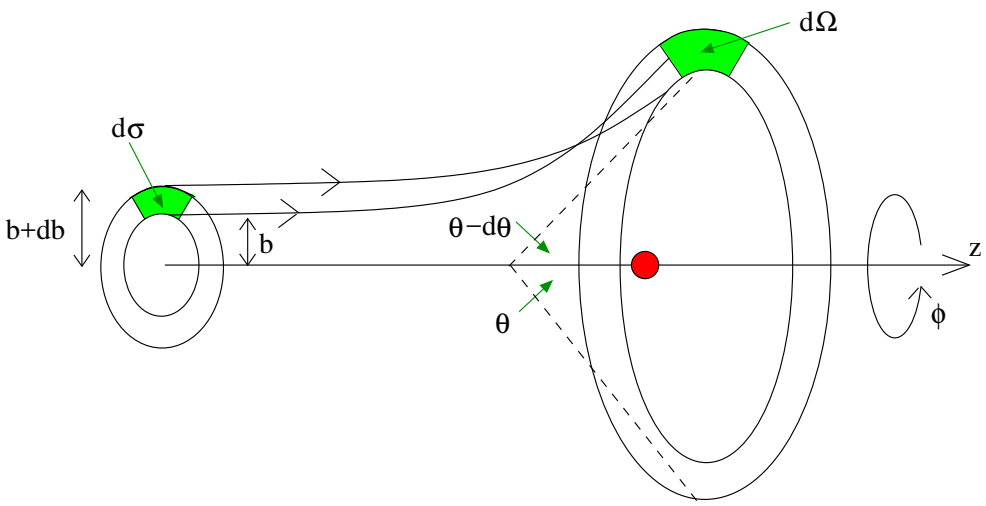


Figure 8: What becomes of an infinitesimal cross-sectional area after scattering.

Now consider a uniform beam of particles, each with kinetic energy E. We want to understand what becomes of this beam. Consider the cross-sectional area, denoted $d\sigma$ in Figure 8. We write this as

$$d\sigma = b\, d\phi\, db$$

The particles within $d\sigma$ will evolve to the lie in a cone of solid angle $d\Omega$, given by

$$d\Omega = \sin\theta\, d\phi\, d\theta$$

where, for central potentials, the infinitesimal angles $d\phi$ are the same in both these formulae. The *differential cross-section* is defined to be

$$\frac{d\sigma}{d\Omega} = \frac{b}{\sin\theta}\left|\frac{db}{d\theta}\right|$$

The left-hand side should really be $|d\sigma/d\Omega|$, but we'll usually drop the modulus. The differential cross-section is a function of incoming momentum k, together with the outgoing angle θ.

More colloquially, the differential cross-section can be thought of as

$$\frac{d\sigma}{d\Omega}\, d\Omega = \frac{\text{Number of particles scattered into } d\Omega \text{ per unit time}}{\text{Number of incident particles per area } d\sigma \text{ per unit time}}$$

We write this in terms of *flux*, defined to be the number of particles per unit area per unit time. In this language, the differential cross-section is

$$\frac{d\sigma}{d\Omega} = \frac{\text{Scattered flux}}{\text{Incident flux}}$$

We can also define the total cross-section

$$\sigma_T = \int d\Omega \, \frac{d\sigma}{d\Omega}$$

Both the differential cross-section and the total cross-section have units of area. The usual unit used in particle physics, nuclear physics and atomic physics is the *barn*, with 1 barn $= 10^{-28}\ m^2$. The total cross-section is a crude characterisation of the scattering power of the potential. Roughly speaking, it can be thought of as the total area of the incoming beam that is scattered. The differential cross-section contains more detailed information.

An Example: The Hard Sphere

Suppose that our particle bounces off a hard sphere, described by the potential $V(r) = \infty$ for $r \leq R$. By staring at the geometry shown in the figure, you can convince yourself that $b = R\sin\alpha$ and $\theta = \pi - 2\alpha$. So in this case

$$b = R\sin\left(\frac{\pi}{2} - \frac{\theta}{2}\right) = R\cos\frac{\theta}{2}$$

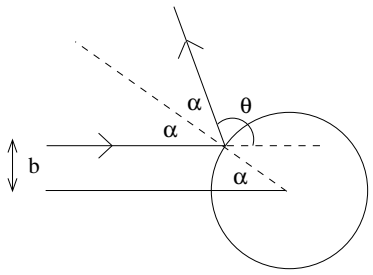


Figure 9:

If $b > R$, clearly there is no scattering. The differential cross-section is

$$\frac{d\sigma}{d\Omega} = \frac{R^2\cos(\theta/2)\sin(\theta/2)}{2\sin\theta} = \frac{R^2}{4}$$

Rather unusually, in this case $d\sigma/d\Omega$ is independent of both θ and E. The total cross-section is

$$\sigma_T = \int_0^{2\pi} d\phi \int_{-1}^{+1} d(\cos\theta)\, \frac{d\sigma}{d\Omega} = \pi R^2 \tag{1.21}$$

which, happily, coincides with the geometrical cross-section of the sphere.

This result reinforces the interpretation of the total cross-section that we mentioned above; it is the area of the beam that is scattered. In general, the area of the beam that is scattered will depend on the energy E of the incoming particles.

Another Example: Rutherford Scattering

Rutherford scattering is the name given to scattering off a repulsive Coulomb potential of the form

$$V(r) = \frac{A}{r} \qquad \text{with} \qquad A > 0$$

where, for two particles of charge q_1 and q_2, we have $A = q_1 q_2/4\pi\epsilon_0$. We studied Rutherford scattering in the lectures on Dynamics and Relativity. We found[1]

$$2bE = A \cot \frac{\theta}{2}$$

This gives the differential cross-section,

$$\frac{d\sigma}{d\Omega} = \frac{b}{\sin\theta}\left|\frac{db}{d\theta}\right| = \left(\frac{A}{4E}\right)^2 \frac{1}{\sin^4(\theta/2)} \tag{1.22}$$

This scattering amplitude played an important role in the history of physics. Rutherford, together with Geiger and Marsden, fired alpha particles (a helium nucleus) at gold foil. They discovered that the alpha particles could be deflected by a large angle, with the cross-section given by (1.22). Rutherford realised that this meant the positive charge of the atom was concentrated in a tiny, nucleus.

There is, however, a puzzle here. Rutherford did his experiment long before the discovery of quantum mechanics. While his data agreed with the classical result (1.22), there is no reason to believe that this classical result carries over to a full quantum treatment. We'll see how this pans out later in this section.

There's a surprise when we try to calculate the total cross-section σ_T. We find that it's infinite! This is because the Coulomb force is long range. The potential decays to $V(r) \to 0$ as $r \to \infty$, but it drops off very slowly. This will mean that we will have to be careful when applying our formalism to the Coulomb force.

1.2.2 The Scattering Amplitude

The language of cross-sections is also very natural when we look at scattering in quantum mechanics. As in Section 1.1, we set up the scattering problem as a solution to the time-independent Schrödinger equation, which now reads

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(r)\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{1.23}$$

We will send in a plane wave with energy E which we choose to propagate along the z-direction. This is just

$$\psi_{\text{incident}}(\mathbf{r}) = e^{ikz}$$

[1]See equation (4.20) of the Dynamics and Relativity lecture notes, where we denoted the scattering angle by ϕ instead of θ.

where $E = \hbar^2 k^2/2m$. However, after scattering off the potential, the wave doesn't only bounce back in the z direction. Instead, it spreads out spherically, albeit with a phase and amplitude which can vary around the sphere. It's hard to take photographs of quantum wavefunctions, but the water waves shown on the right give a good analogy for what's going on. Asymptotically, as $r \to \infty$, this scattered wave takes the form

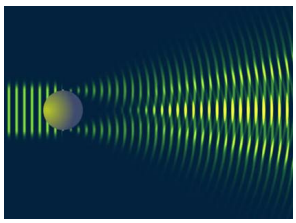


Figure 10:

$$\psi_{\text{scattered}}(\mathbf{r}) = f(\theta, \phi)\, \frac{e^{ikr}}{r} \tag{1.24}$$

The $1/r$ fall-off follows from solving the free Schrödinger equation; we'll see this explicitly below. However, there is a simple intuition for this behaviour which follows from thinking of $|\psi|^2$ as a probability, spreading over a sphere which grows as r^2 as $r \to \infty$. The $1/r$ fall-off ensures that this probability is conserved. Our final ansatz for the asymptotic wavefunction is then

$$\psi(\mathbf{r}) = \psi_{\text{incident}}(\mathbf{r}) + \psi_{\text{scattered}}(\mathbf{r}) \tag{1.25}$$

The function $f(\theta, \phi)$ is called the *scattering amplitude*. For the central potentials considered here it is independent of ϕ, so $f = f(\theta)$. It is the 3d generalisation of the reflection and transmission coefficients that we met in the previous section. Our goal is to calculate it.

The scattering amplitude is very closely related to the differential cross-section. To see this, we can look at the probability current

$$\mathbf{J} = -i\frac{\hbar}{2m}\Big(\psi^\star \nabla\psi - (\nabla\psi^\star)\psi\Big)$$

which obeys $\nabla \cdot \mathbf{J} = 0$. For the incident wave, we have

$$\mathbf{J}_{\text{incident}} = \frac{\hbar k}{m}\hat{\mathbf{z}}$$

This is interpreted as a beam of particles with velocity $v = \hbar k/m$ travelling in the z-direction. Meanwhile, the for the scattered wave we use the fact that

$$\nabla\psi_{\text{scattered}} = \frac{ikf(\theta)e^{ikr}}{r}\hat{\mathbf{r}} + \mathcal{O}\left(\frac{1}{r^2}\right)$$

to find

$$\mathbf{J}_{\text{scattered}} = \frac{\hbar k}{m}\frac{1}{r^2}|f(\theta)|^2\,\hat{\mathbf{r}} + \mathcal{O}\left(\frac{1}{r^3}\right)$$

This means that, as $r \to \infty$, the flux of outgoing particles crossing an area dA subtended by the solid angle $d\Omega$

$$\mathbf{J}_{\text{scattered}} \cdot \hat{\mathbf{r}}\, dA = \frac{\hbar k}{m} |f(\theta)|^2\, d\Omega$$

The differential cross-section is defined to be the ratio of the scattered flux through $d\Omega$, divided by the incident flux. In other words, it is

$$\frac{d\sigma}{d\Omega} = \frac{\hbar k |f(\theta)|^2/m}{\hbar k/m} = |f(\theta)|^2$$

This is rather nice. It means that if we can compute the scattering amplitude $f(\theta)$, it immediately tells us the differential cross-section. The total cross-section is defined, as before, as

$$\sigma_T = \int d\Omega\ |f(\theta)|^2$$

1.2.3 Partial Waves

To make progress, we need to start to look in a more detail at the solutions to the Schrödinger equation (1.23). Because we've decided to work with rotationally invariant potentials, it makes sense to label our wavefunctions by their angular momentum, l. Let's quickly review what this looks like.

A general wavefunction $\psi(r, \theta, \phi)$ can be expanded in terms of spherical harmonics. In this section, however, we only need to deal with wavefunctions of the for form $\psi(r, \theta)$, which are independent of ϕ. Such functions have an expansion in terms of *partial waves*

$$\psi(r, \theta) = \sum_{l=0} R_l(r)\, P_l(\cos\theta)$$

Here the $P_l(\cos\theta)$ are *Legendre polynomials*. They appear by virtue of being eigenstates of the angular momentum operator $\mathbf{L}^2$,

$$\mathbf{L}^2\, P_l(\cos\theta) = \hbar^2 l(l+1) P_l(\cos\theta)$$

In more concrete terms, this is the statement that the Legendre polynomials $P_l(w)$ obey the differential equation

$$\frac{d}{dw}(1-w^2)\frac{dP_l}{dw} + l(l+1)P_l(w) = 0$$

Meanwhile, the original Schrödinger equation (1.23) becomes an ordinary differential equation for the radial functions R_l,

$$\left(\frac{d^2}{dr^2} + \frac{2}{r}\frac{d}{dr} - \frac{l(l+1)}{r^2} - U(r) + k^2\right) R_l(r) = 0 \tag{1.26}$$

where we've used the expression for the energy, $E = \hbar^2 k^2/2m$, and rescaled the potential

$$U(r) = \frac{2m}{\hbar^2} V(r)$$

Spherical Waves when $U(r) = 0$

We will assume that our potential drops off sufficiently quickly so that asymptotically our waves obey (1.26) with $U(r) = 0$. (We will be more precise about how fast $U(r)$ must fall off later.) We can write the equation obeyed by R_l as

$$\left(\frac{d^2}{dr^2} - \frac{l(l+1)}{r^2} + k^2\right) (rR_l(r)) = 0 \tag{1.27}$$

There are two s-wave solutions with $l = 0$, given by

$$R_0(r) = \frac{e^{\pm ikr}}{r} \tag{1.28}$$

These are ingoing (minus sign) and outgoing (plus sign) spherical waves.

The solutions for $l \neq 0$ are more known as *spherical Bessel functions* and are described below.

Plane Waves when $U(r) = 0$

Of course, when $U = 0$, the plane wave

$$\psi_{\text{incident}}(\mathbf{r}) = e^{ikz} = e^{ikr\cos\theta}$$

is also a solution to the Schrödinger equation. Although it feels rather unnatural, it must be possible to expand these solutions in terms of the spherical waves. To do this, it is convenient to briefly introduce the coordinate $\rho = kr$. We write the plane wave solution as

$$\psi_{\text{incident}}(\rho, \theta) = e^{i\rho\cos\theta} = \sum_l (2l+1) u_l(\rho) P_l(\cos\theta) \tag{1.29}$$

where the factor of $(2l+1)$ is for convenience and the function $u_l(\rho)$ are what we want to determine. The Legendre polynomials have a nice orthogonality property,

$$\int_{-1}^{+1} dw\ P_l(w) P_m(w) = \frac{2}{2l+1}\delta_{lm} \tag{1.30}$$

We can use this to write

$$u_l(\rho) = \frac{1}{2}\int_{-1}^{+1} dw\ e^{i\rho w} P_l(w) \tag{1.31}$$

Our interest is only in the behaviour of the plane wave as $\rho \to \infty$. To extract this, we start by integrating by parts

$$u_l(\rho) = \frac{1}{2}\left[\frac{e^{i\rho w}P_l(w)}{i\rho}\right]_{-1}^{+1} - \frac{1}{2i\rho}\int_{-1}^{+1} dw\ e^{i\rho w}\,\frac{dP_l}{dw}$$

The Legendre polynomials obey $P_l(1) = 1$ and $P_l(-1) = (-1)^l$. We then find

$$u_l(\rho) = \frac{1}{2i\rho}\left[e^{i\rho} - (-1)^l e^{-i\rho}\right] + \mathcal{O}\left(\frac{1}{\rho^2}\right) \tag{1.32}$$

where a further integration by parts will convince you that the remaining terms do indeed drop off as $1/\rho^2$. This is the result we need. As $r \to \infty$, the incident plane wave can be written as

$$\psi_{\text{incident}} = \sum_{l=0}^{\infty} \frac{2l+1}{2ik}\left[\frac{e^{ikr}}{r} - (-1)^l\frac{e^{-ikr}}{r}\right] P_l(\cos\theta) \tag{1.33}$$

We learn that the ingoing plane wave decomposes into an outgoing spherical wave (the first term) together with an ingoing spherical wave (the second term).

Phase Shifts

It's been quite a long build up, but we now know what we want to calculate, and how to do it! To recapitulate, we'd like to calculate the scattering amplitude $f(\theta)$ by finding solutions of the asymptotic form

$$\psi(\mathbf{r}) = e^{ikz} + f(\theta)\frac{e^{ikr}}{r} \qquad \text{as } r \to \infty$$

We still have a couple more definitions to make. First, we expand the scattering amplitude in partial waves as

$$f(\theta) = \sum_{l=0}^{\infty} \frac{2l+1}{k}\, f_l\, P_l(\cos\theta) \tag{1.34}$$

The normalisation coefficients of $1/k$ and $(2l+1)$ mean that the coefficients f_l sit nicely with the expansion (1.33) of the plane wave in terms of spherical waves. We can then write the asymptotic form of the wavefunction as a sum of ingoing and outgoing waves

$$\psi(\mathbf{r}) \sim \sum_{l=0}^{\infty} \frac{2l+1}{2ik}\left[(-1)^{l+1}\frac{e^{-ikr}}{r} + (1+2if_l)\frac{e^{ikr}}{r}\right] P_l(\cos\theta) \tag{1.35}$$

where the first term is ingoing, and the second term is outgoing. For a given potential $V(r)$, we would like to compute the coefficients f_l which, in general, are functions of k.

Note that the problem has decomposed into decoupled angular momentum sectors, labelled by $l = 0, 1, \ldots$. This is because we're working with a rotationally symmetric potential, which scatters an incoming wave, but does not change its angular momentum. Moreover, for each l, our ansatz consists of an ingoing wave, together with an outgoing wave. This is entirely analogous to our 1d solutions (1.9) when we first introduced the S-matrix. We identify the coefficients of the outgoing terms as the elements of the S-matrix. For rotationally invariant potentials, the 3d S-matrix $\mathcal{S}$ is diagonal in the angular momentum basis, with elements given by

$$S_l = 1 + 2if_l \quad \text{with} \quad l = 0, 1, 2, \ldots$$

Now unitarity of the S-matrix — which is equivalent to conservation of particle number — requires that these diagonal elements are a pure phase. We write

$$S_l = e^{2i\delta_l} \quad \Rightarrow \quad f_l = \frac{1}{2i}(e^{2i\delta_l} - 1) = e^{i\delta_l} \sin\delta_l$$

where δ_l are the phase shifts. Comparing back to (1.34), we see that the phase shifts and scattering amplitude are related by

$$f(\theta) = \frac{1}{2ik}\sum_{l=0}^{\infty}(2l+1)\left(e^{2i\delta_l} - 1\right)P_l(\cos\theta)$$

The picture that we have is entirely analogous to the 1d situation. A wave comes in, and a wave goes out. Conservation of probability ensures that the amplitudes of these waves are the same. All information about scattering is encoded in the phase shifts $\delta_l(k)$ between the ingoing and outgoing waves.

1.2.4 The Optical Theorem

The differential cross-section is $d\sigma/d\Omega = |f(\theta)|^2$. Using the partial wave decomposition (1.34), we have

$$\frac{d\sigma}{d\Omega} = \frac{1}{k^2}\sum_{l,l'}(2l+1)(2l'+1)f_l f_{l'}^\star P_l(\cos\theta)P_{l'}(\cos\theta)$$

In computing the total cross-section σ_T, we can use the orthogonality of Legendre polynomials (1.30) to write

$$\sigma_T = 2\pi\int_{-1}^{+1} d(\cos\theta)\ \frac{d\sigma}{d\Omega} = \frac{4\pi}{k^2}\sum_l(2l+1)|f_l|^2 = \frac{4\pi}{k^2}\sum_l(2l+1)\sin^2\delta_l \qquad (1.36)$$

We can compare this to our expansion (1.34). Using the fact that $P(1) = 1$, we have

$$f(0) = \sum_l \frac{2l+1}{k} e^{i\delta_l} \sin\delta_l$$

This tells us that the total cross-section is given by

$$\sigma_T = \frac{4\pi}{k} \mathrm{Im} f(0)$$

This is known as the *optical theorem*.

Here's some words that will hopefully build some intuition for the optical theorem. The potential causes scattering from the forward direction ($\theta = 0$) to other directions. Because total probability is conserved, clearly the amount of particles going in the forward direction must decrease. However, this decrease in the forward direction must be equal to the total increase in other directions – and this is what the total cross-section σ_T measures. Finally, the amount of decrease in forward scattering is due to interference between the incoming wave and outgoing waves, and so is proportional to $f(0)$.

Unitarity Bounds

If we think of the total cross-section as built from the cross-sections for each partial wave then, from (1.36), we have

$$\sigma_T = \sum_{l=0}^{\infty} \sigma_l \quad \text{with} \quad \sigma_l = \frac{4\pi}{k^2}(2l+1)\sin^2\delta_l \tag{1.37}$$

Clearly each contribution is bounded as $\sigma_l \leq 4\pi(2l+1)/k^2$, with the maximum arising when the phase shift is given by $\delta_l = \pm\pi/2$. This is called the *unitarity bound*.

There's a straightforward, semi-classical way to understand these unitarity bounds. If we send in a particle with momentum $\hbar k$ and impact parameter b, then it has angular momentum $L = \hbar k b$. This angular momentum is quantised. Roughly speaking, we might expect that the particle has angular momentum $\hbar l$, with $l \in \mathbf{Z}$, when the impact parameter lies in the window

$$\frac{l}{k} \leq b \leq \frac{l+1}{k} \tag{1.38}$$

If the particle gets scattered with 100% probability when it lies in this ring, then the cross-section is equal to the area of the ring. This is

$$\frac{(l+1)^2\pi}{k^2} - \frac{l^2\pi}{k^2} = \frac{(2l+1)\pi}{k^2}$$

This is *almost* the unitarity bound (1.37). It differs by a factor 4. As we will now see, that same factor of 4 difference often arises between simple classical arguments and a full quantum treatment of scattering processes.

1.2.5 An Example: A Hard Sphere and Spherical Bessel Functions

After all this formalism, let's finally do an example. Our scattering region will be a hard sphere of radius a, with potential

$$V(r) = \begin{cases} \infty & r < a \\ 0 & r > a \end{cases}$$

Since the wavefunction vanishes inside the sphere and is continuous, this potential is equivalent to imposing the boundary condition $\psi(a) = 0$.

For $r > a$, the wavefunction can be decomposed in partial waves

$$\psi(r, \theta) = \sum_{l=0} R_l(r)\, P_l(\cos\theta)$$

where the radial wavefunction obeys the free Schrödinger equation

$$\left(\frac{d^2}{d\rho^2} - \frac{l(l+1)}{\rho^2} + 1 \right) (\rho R_l(\rho)) = 0 \tag{1.39}$$

where we're again using the coordinate $\rho = kr$. Solutions $R_l(\rho)$ to this equation are known as *spherical Bessel functions* and are denoted $j_l(\rho)$ and $n_l(\rho)$. They are important enough that we take some time to describe their properties.

An Aside: Spherical Bessel Functions

The solutions to (1.39) are given by *spherical Bessel functions*, $R_l(\rho) = j_l(\rho)$ and $R_l(\rho) = n_l(\rho)$, and can be written as[2]

$$j_l(\rho) = (-\rho)^l \left(\frac{1}{\rho} \frac{d}{d\rho} \right)^l \frac{\sin\rho}{\rho} \quad \text{and} \quad n_l(\rho) = -(-\rho)^l \left(\frac{1}{\rho} \frac{d}{d\rho} \right)^l \frac{\cos\rho}{\rho}$$

Note that $j_0(\rho) = \sin\rho/\rho$ and $n_0(\rho) = -\cos\rho/\rho$, so the solutions (1.28) for free spherical waves can be written as $R_0(\rho) = n_0(\rho) \pm i n_0(\rho)$.

[2]Proofs of this statement, together with the asymptotic expansions given below, can be found in the handout http://www.damtp.cam.ac.uk/user/tong/aqm/bessel.pdf.

In what follows, it will be useful to have the asymptotic form of j_l and n_l. They are given by

$$j_l(\rho) \to \frac{\sin(\rho - \frac{1}{2}l\pi)}{\rho} \quad \text{and} \quad n_l(\rho) \to -\frac{\cos(\rho - \frac{1}{2}l\pi)}{\rho} \quad \text{as } \rho \to \infty \tag{1.40}$$

We see that at large r, the spherical Bessel functions look more or less the same for all l, differing only by a phase. In particular, the combinations $j_l \pm n_l$ look essentially the same as the $l = 0$ spherical waves that we met in (1.28). However, the spherical Bessel functions differ as we come in towards the origin. In particular, close to $\rho = 0$ we have

$$j_l(\rho) \to \frac{\rho^l}{(2l+1)!!} \quad \text{and} \quad n_l(\rho) \to -(2l-1)!!\,\rho^{-(l+1)} \quad \text{as } \rho \to 0 \tag{1.41}$$

where $(2l+1)!! = 1 \cdot 3 \cdot 5 \cdot \cdots (2l+1)$ is the product of all odd numbers up to $2l+1$. Note that $j_l(\rho)$ is regular near the origin, while n_l diverges.

Before we proceed, it's worth seeing how we write the plane wave e^{ikz} in terms of spherical Bessel functions. We wrote the partial wave expansion (1.29) in terms of functions $u_l(\rho)$, whose asymptotic expansion was given in (1.32). This can be rewritten as

$$u_l(\rho) \;\to\; i^l \frac{\sin(\rho - \frac{1}{2}l\pi)}{\rho} \quad \text{as} \quad \rho \to \infty$$

which tells us that we can identify the functions $u_l(\rho)$ as

$$u_l(\rho) = i^l j_l(\rho)$$

Back to the Hard Sphere

Returning to our hard sphere, the general solution for $r \geq a$ can be written in the form,

$$R_l(r) = A_l\Big[\cos\alpha_l\, j_l(\rho) - \sin\alpha_l\, n_l(\rho)\Big] \tag{1.42}$$

where, as before, $\rho = kr$. Here A_l and α_l are two integration constants which we will fix by the boundary condition. Because the Schrödinger equation is linear, nothing fixes the overall coefficient A_l. In contrast, the integration constant α_l will be fixed by the boundary conditions at $r = a$. Moreover, this integration constant turns out to be precisely the phase shift δ_l that we want to compute. To see this, we use the asymptotic form of the spherical Bessel functions (1.40) to find

$$R_l(r) \sim \frac{1}{\rho}\left[\cos\alpha_l\, \sin(\rho - \frac{1}{2}l\pi) + \sin\alpha_l\, \cos(\rho - \frac{1}{2}l\pi)\right] = \frac{1}{\rho}\sin(\rho - \frac{1}{2}l\pi + \alpha_l)$$

We can compare this to the expected asymptotic form (1.35) of the wavefunction

$$R_l(r) \sim \left[(-1)^{l+1}\frac{e^{-i\rho}}{\rho} + e^{2i\delta_l}\frac{e^{i\rho}}{\rho}\right] = \frac{e^{i\delta_l}e^{i\pi l/2}}{\rho}\left[- e^{-i(\rho+\delta_l-\pi l/2)} + e^{i(\rho+\delta_l-\pi l/2)}\right]$$

to see that, as a function of $\rho = kr$, the two expressions agree provided

$$\alpha_l = \delta_l$$

In other words, if we can figure out the integration constant α_l then we've found our sought-after phase shift.

The boundary condition imposed by the hard sphere is simply $R_l(a) = 0$. This tells us that

$$\cos\delta_l\, j_l(ka) = \sin\delta_l\, n_l(ka) \quad \Rightarrow \quad \tan\delta_l = \frac{j_l(ka)}{n_l(ka)}$$

This is the final result for this system. Now let's try to extract some physics from it.

First note that for the $l = 0$ s-wave, the phase shift is given by exactly by

$$\delta_0 = -ka$$

For small momenta, $ka \ll 1$, we can extract the behaviour of the higher l phase shifts from $\rho \to 0$ behaviour of the spherical Bessel functions (1.41). We have

$$\delta_l \approx -\frac{(ka)^{2l+1}}{(2l+1)!!\,(2l-1)!!}$$

We see that for low momentum the phase shifts decrease as l increases. This is to be expected: the higher l modes have to penetrate the repulsive angular momentum $\sim \hbar l(l+1)/r^2$. Classically, this would prohibit the low-momentum modes from reaching the sphere. Quantum mechanically, only the exponential tails of these modes reach $r = a$ which is why their scattering is suppressed.

For low momentum $ka \ll 1$, we now have all the information we need to compute the total cross-section. The sum (1.36) is dominated by the $l = 0$ s-wave, and given by

$$\sigma_T = 4\pi a^2\Big(1 + \mathcal{O}\left((ka)^4\right)\Big)$$

This is a factor of 4 bigger than the classical, geometric result (1.21)

It's also possible to extract analytic results for the phase shifts at high momentum $ka \gg 1$. For this we need further properties of the spherical Bessel functions. Here we simply state the results. The phase shifts δ_l vary between 0 and 2π for $l \lesssim ka$. However, when $l < ka$, the phase shifts quickly drop to zero. The intuition behind this follows from the semi-classical analysis (1.38) which tells us that for $l \gg ka$, the impact parameter is $b \gg a$. This makes it unsurprising that no scattering takes place in this regime. It turns out that as $ka \to \infty$, the total cross-section becomes $\sigma_T \to 2\pi a^2$.

The Scattering Length

The low-momentum behaviour $\delta_l \sim (ka)^{2l+1}$ that we saw is common to all scattering potentials. It means that low-energy scattering is always dominated by the s-wave whose phase shift scales as

$$\delta_0 \sim -ka_s + \mathcal{O}(k^3) \tag{1.43}$$

The coefficients a_s is called the *scattering length*. As we have seen, for the hard sphere $a_s = a$, the radius of the sphere. At low energies, the total cross-section is always given by

$$\sigma_T \approx \sigma_0 \sim 4\pi a_s^2$$

The scattering length is a useful way to characterise the low-energy behaviour of a potential. As we will see in examples below, a_s can be positive or negative and can, at times, diverge.

1.2.6 Bound States

In this section we describe the effects of bound states on scattering. Such states only occur for attractive potentials, so we again take a sphere of radius a, but this time with potential

$$V(r) = \begin{cases} -V_0 & r < a \\ 0 & r > a \end{cases} \tag{1.44}$$

It will be useful to define the following notation

$$U(r) = \frac{2mV(r)}{\hbar^2} \quad \text{and} \quad \gamma^2 = \frac{2mV_0}{\hbar^2} \tag{1.45}$$

We'll start by focussing on the $l = 0$ s-wave. Outside the sphere, the wavefunction satisfies the usual free Schrödinger equation (1.27)

$$\left(\frac{d^2}{dr^2} + k^2\right)(r\psi) = 0 \qquad r > a$$

with general solution

$$\psi(r) = \frac{A\sin(kr+\delta_0)}{r} \qquad r > a \tag{1.46}$$

The same argument that we made when discussing the hard sphere shows that the integration constant δ_0 is the phase shift that we want to calculate. We do so by matching the solution to the wavefunction inside the sphere, which satisfies

$$\left(\frac{d^2}{dr^2} + k^2 + \gamma^2\right)(r\psi) = 0 \qquad r < a$$

The requirement that the wavefunction is regular at the origin $r = 0$ picks the solution inside the sphere to be

$$\psi(r) = \frac{B\sin(\sqrt{k^2+\gamma^2}r)}{r} \qquad r < a \tag{1.47}$$

The solutions (1.46) and (1.47) must be patched at $r = a$ by requiring that both $\psi(a)$ and $\psi'(a)$ are continuous. We get the answer quickest if we combine these two and insist that ψ'/ψ is continuous at $r = a$, since this condition does not depend on the uninteresting integration constants A and B. A quick calculation shows that it is satisfied when

$$\frac{\tan(ka+\delta_0)}{ka} = \frac{\tan(\sqrt{k^2+\gamma^2}a)}{\sqrt{k^2+\gamma^2}a} \tag{1.48}$$

For very high momentum scattering, $k^2 \gg \gamma^2$, we have $\delta_0 \to 0$. This is to be expected: the energy of the particle is so large that it doesn't much care for the small, puny potential and there is no scattering.

Bound States and the Scattering Length

Things are more interesting at low energies, $k^2 \ll \gamma^2$ and $ka \ll 1$. We have

$$\frac{\tan(ka+\delta_0)}{ka} \approx \frac{\tan(\gamma a)}{\gamma a} \quad \Rightarrow \quad \frac{\tan(ka)+\tan(\delta_0)}{1-\tan(ka)\tan(\delta_0)} \approx \frac{k}{\gamma}\tan(\gamma a)$$

Rearranging, we get

$$\tan\delta_0 = ka\left(\frac{\tan(\gamma a)}{\gamma a} - 1\right) + \mathcal{O}(k^3) \tag{1.49}$$

If the phase shift δ_0 is small, then we can write $\tan\delta_0 \approx \delta_0$ and, from (1.43), read off the scattering length

$$a_s = a - \frac{\tan(\gamma a)}{\gamma} \tag{1.50}$$

Note that, for this approximation to hold, we need $ka_s \ll 1$, but the scattering length a_s exhibits somewhat surprising behaviour. For small γ, the scattering length is negative. This can be thought of as due to the attractive nature of the potential, which pulls the particle into the scattering region rather than repelling it. However, as γ is increased, the scattering length diverges to $-\infty$, before reappearing at $+\infty$. It continues this pattern, oscillating between $+\infty$ and $-\infty$. Our task is to understand why this striking behaviour is happening.

Before we proceed, note that all the calculations above also hold for repulsive potentials with $V_0 < 0$. In this case γ, defined in (1.45) is pure imaginary and the scattering length (1.50) becomes

$$a_s = a - \frac{\tanh(|\gamma|a)}{|\gamma|} \qquad (V_0 < 0)$$

Now the scattering length is always positive. It increases monotonically from $a_s = 0$ when $\gamma = 0$, corresponding to no scattering, through to $a_s = a$ when $|\gamma| \to \infty$, which is our previous result for the hard-sphere. We see that whatever is causing the strange oscillations in (1.50) does not occur for the repulsive potential.

The key to the divergent behaviour of the scattering length lies in the bound states of the theory. It's a simple matter to construct $l = 0$ bound states. We solve the Schrödinger equation with the form

$$r\psi(r) = \begin{cases} A\sin(\sqrt{\gamma^2 - \lambda^2}r) & r < a \\ Be^{-\lambda r} & r > a \end{cases}$$

The two solutions have the same energy $E = -\hbar^2\lambda^2/2m$. Matching the logarithmic derivatives across $r = a$ gives

$$\tan(\sqrt{\gamma^2 - \lambda^2}a) = -\frac{\sqrt{\gamma^2 - \lambda^2}}{\lambda} \tag{1.51}$$

This structure of the solutions is similar to what we saw in Section 1.1.4. Indeed, if we write $q^2 = \gamma^2 - \lambda^2$, then these equations take the same form as (1.16) that describe odd-parity states in one-dimension. In particular, this means that if the potential is too shallow then no bound states exist. As γ gets larger, and the potential gets deeper, bound states start to appear. They first arise when $\lambda = 0$ and $\tan(\gamma a) = \infty$, so that

$$\gamma = \gamma_\star = \left(n + \frac{1}{2}\right)\frac{\pi}{a} \qquad \text{with} \qquad n = 0, 1, \ldots$$

This coincides with the values for which the scattering length (1.50) diverges. For γ slightly less than $\gamma_\star$, the bound state has not yet appeared and the scattering length is very large and negative. For γ slightly greater than $\gamma_\star$, the new state exists and is weakly bound, and the scattering length is large and positive. Meanwhile, when $\gamma = \gamma_\star$, then there is a bound state which has energy $E = 0$. Such bound states are said to be "*at threshold*".

The incoming wave has energy slightly above $E = 0$ and mixes strongly with the state with bound state – or almost bound state – with energy a little below $E = 0$. This is what gives rise to the divergence in the cross-section. Specifically, when there is a bound state exactly at threshold, $\tan\delta_0 \to \infty$ and so the phase shift is $\delta_0 = (n + \frac{1}{2})\pi$. (Note that at this point, we can no longer write $\delta_0 \approx -ka_s$ because a_s this is valid only for $ka_s \ll 1$, but a_s is diverging.) The s-wave cross-section saturates the unitarity bound (1.37)

$$\sigma_0 = \frac{4\pi}{k^2}$$

To understand why the formation of bound states gives rise to a divergent scattering length, we can look at the analytic structure of the S-matrix at finite k. We know from (1.48) that the phase shift is given by

$$\tan(ka + \delta_0) = \frac{k}{\sqrt{k^2+\gamma^2}}\tan(\sqrt{k^2+\gamma^2}a) \equiv f(k)$$

Rearranging, we get the s-wave component of the S-matrix

$$S_0(k) = e^{2i\delta_0} = e^{-2ika}\frac{1+if(k)}{1-if(k)}$$

The S-matrix has a pole at $f(k) = -i$, or for values of k such that

$$\tan(\sqrt{k^2+\gamma^2}a) = \frac{\sqrt{k^2+\gamma^2}}{ik} \tag{1.52}$$

This has no solutions for real k. However, it does have solutions along the positive imaginary k axis. If we set $k = i\lambda$, the equation (1.52) coincides with the condition for bound states (1.51).

Close to the pole, the S-matrix takes the form

$$S_0(k) = e^{2i\delta_0} = \frac{i\lambda + k}{i\lambda - k}$$

When the bound state approaches threshold, λ is small and this form is valid in the region $k = 0$. For $k \ll \lambda$, we can expand in k/λ to find $\delta_0 \approx -k/\lambda$, which tells us that we should indeed expect to see a divergent scattering length $a_s = 1/\lambda$.

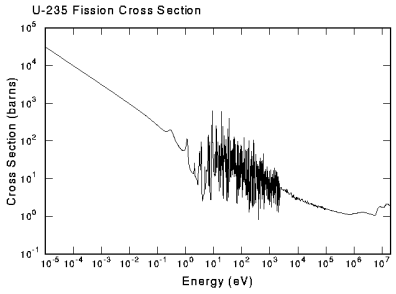


Figure 11: The cross-section for neutron scattering off U-235.

When neutrons scatter off large nuclei at low-energies they are very close to forming a threshold bound state. The total cross-section for neutron scattering off uranium 235 is shown in the figure[3]. You can see the large enhancement of the cross-section. This is partly due to the bound state, although it is complicated by the presence of a large number of resonances whose effects we'll discuss in the next section.

1.2.7 Resonances

We already met the idea of resonances in Section 1.1.5. These are unstable bound states, which appear as poles of the S-matrix in the lower-half complex plane. Here we see how these resonances affect scattering in 3d.

It's not hard to construct examples which exhibit resonances. Indeed, the attractive, spherical potential (1.44) which has bound states also exhibits resonances. These don't occur for s-waves, but only for higher l, where the effective potential includes an effective, repulsive angular momentum barrier. The algebra is not conceptually any more difficult than what we did above, but in practice rapidly becomes a blur of spherical Bessel functions.

Alternatively, we could look at the somewhat simpler example of a delta-function cage of the form $V(r) = V_0\delta(r-a)$, which is the obvious 3d generalisation of the example we looked at in Section 1.1.5 and has s-wave resonances.

Rather than getting bogged down in any of these details, here we focus on the features that are common to all these examples. In each case, the S-matrix has a pole. Thinking in terms of energy $E = \hbar^2k^2/2m$, these poles occur at

$$E = E_0 - \frac{i\Gamma}{2}$$

[3]The data is taken from the Los Alamos on-line nuclear information tour.

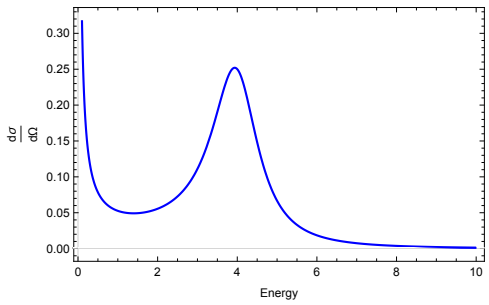


Figure 12: Distribution with $\Gamma^2 = 2$...

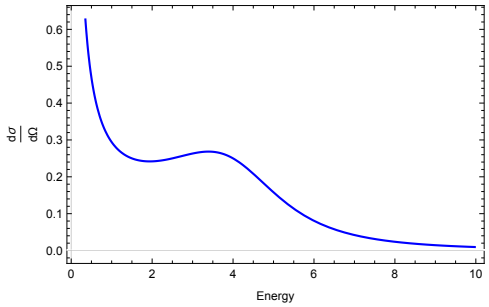


Figure 13: ...and with $\Gamma^2 = 15$

This is the same result (1.17) that we saw in our 1d example. Close to the pole, the S-matrix — which, by unitarity, is simply a phase — must take the form

$$S(E) = e^{2i\delta(E)} = e^{2i\theta(E)} \frac{E - E_0 - i\Gamma/2}{E - E_0 + i\Gamma/2} \tag{1.53}$$

Here $e^{2i\theta(E)}$ is the so-called continuum contribution; it is due to the usual, run-of-the-mill phase shift that arises from scattering off the potential. Here our interest is in the contributions that come specifically from the resonance, so we'll set $\theta = 0$. From (1.53), we have

$$\cos\, 2\delta = \frac{(E - E_0)^2 - \Gamma^2/4}{(E - E_0)^2 + \Gamma^2/4} \quad \Rightarrow \quad \sin^2 \delta = \frac{\Gamma^2}{4(E - E_0)^2 + \Gamma^2}$$

From this we can read off the contribution to the total cross-section using (1.36). If the pole occurs for a partial wave with angular momentum l, we have

$$\sigma_T \approx \frac{4\pi}{k^2}(2l + 1)\frac{\Gamma^2}{4(E - E_0)^2 + \Gamma^2}$$

This distribution is plotted in the figure, with $E_0 = 4$ and $\Gamma^2 = 2$ and 15. (Remember that there is an extra factor of E sitting in the k^2 in the formula above). It is called the *Breit-Wigner distribution*, or sometimes the *Lorentzian distribution* (although, strictly speaking, neither of these has the extra factor of $1/k^2$). It exhibits a clear peak at $E = E_0$, whose width is given by $\Gamma/2$. Comparing to our discussion in Section 1.1.5, we see that the lifetime of the resonance can be read off from the width of the peak: the narrower the peak, the longer lived the resonance.

The Breit-Wigner distribution is something of an iconic image in particle physics because this is the way that we discover new particles. To explain this fully would require us to move to the framework of quantum field theory, but we can get a sense

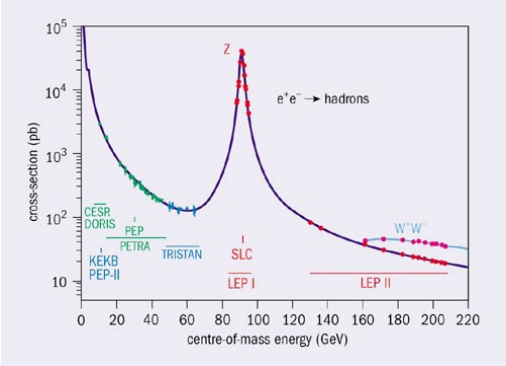


Figure 14: The cross-section for the Z-boson.

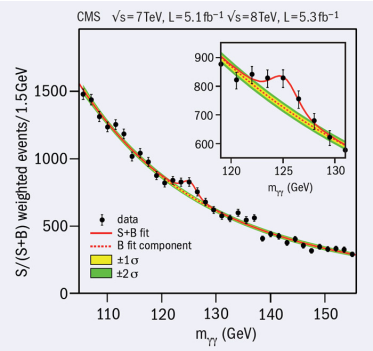


Figure 15: And for the Higgs boson.

for what's going on from what we've seen above. The key fact is that most particles in Nature are not stable. The exceptions are the electron, the proton, neutrinos and photons. All other decay with some lifetime τ. When we collide known particles — typically electrons or protons — we can create new particles which, since they are unstable, show up as resonances. The energy E_0 corresponds to the mass of the new particle through $E_0 = mc^2$, while the lifetime is seen in the width, $\tau = 1/\Gamma$.

Two examples are shown in the figures. The left-hand figure shows the cross-section, now measured in pico-barns $= 10^{-40}\ m^2$, for high-energy electron-positron scattering. We see a large resonance peak which sits at a centre of mass energy $E_0 \approx 91\ GeV$ with width $\Gamma \approx 2.5\ GeV$. Since we're measuring the width in unit of energy, we need a factor of $\hbar$ to convert to the lifetime

$$\tau = \frac{\hbar}{\Gamma}$$

Using $\hbar \approx 6.6 \times 10^{-16}\ eV$, we find the lifetime of the Z-boson to be $\tau \approx 3 \times 10^{-25}\ s$.

The right-hand figure shows the 2012 data from the discovery of the Higgs boson, with mass $E_0 \approx 125\ GeV$. I should confess that the experiment doesn't have the resolution to show the Breit-Wigner shape in this case. The best that can be extracted from this plot is a bound on the width of $\Gamma < 17\ MeV$ or so, while the true width is predicted by theory to be $\Gamma \sim 4\ MeV$.

1.3 The Lippmann-Schwinger Equation

So far, we've developed the machinery necessary to compute cross-sections, but our examples have been rather artificial. The interactions between particles do not look

like spherical potential wells or shells of delta-functions. Instead, they are smooth potentials $V(r)$, such as the Coulomb or Yukawa potentials. We would like to understand scattering in these more realistic settings .

In principle, this is straightforward: you simply need to solve the relevant Schrödinger equation, impose regularity at the origin, and then read off the appropriate phase shifts asymptotically. In practice, the solution to the Schrödinger equation is rarely known analytically. (A counterexample to this is the Coulomb potential which will be discussed in Section 1.4.) In this section, we present a different approach to scattering that makes use of Green's functions. This provides a platform to develop a perturbative approach to understanding scattering for potentials that we actually care about. Moreover, these Green's functions methods also have applications in other areas of physics.

Our starting point is the Schrödinger equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(r)\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{1.54}$$

We'll briefly use a more formal description of this equation, in order to write the Lippmann-Schwinger equation in its most general form. We'll then revert back to the form (1.54) which, for the purposes of these lectures, is all we really care about. With this in mind, we write the Schrödinger equation as

$$(H_0 + V)|\psi\rangle = E|\psi\rangle$$

The idea here is that we've split the Hamiltonian up into a piece that is simple to solve – in this case $H_0 = -\hbar^2\nabla^2/2m$ – and a more complicated piece, V. Trivially re-arranging this equation gives

$$(E - H_0)|\psi\rangle = V|\psi\rangle \tag{1.55}$$

We can then formally re-arrange this equation once more to become

$$|\psi\rangle = |\phi\rangle + \frac{1}{E - H_0}V|\psi\rangle \tag{1.56}$$

Here $|\phi\rangle$ is a zero mode which obeys $H_0|\phi\rangle = E|\phi\rangle$. If (1.56) is multiplied by $E - H_0$ then the state $|\phi\rangle$ is annihilated and we get back to (1.55). However, the inverse quantum operator $(E - H_0)^{-1}$ is somewhat subtle and, as we will see below, there is very often an ambiguity in its definition. This ambiguity is resolved by writing this inverse operator as $(E - H_0 + i\epsilon)^{-1}$, and subsequently taking the limit $\epsilon \to 0^+$. We then write

$$|\psi\rangle = |\phi\rangle + \frac{1}{E - H_0 + i\epsilon}V|\psi\rangle \tag{1.57}$$

This is the *Lippmann-Schwinger equation*. It is not really a solution to the Schrödinger equation (1.54) since $|\psi\rangle$ appears on both sides. It is more a rewriting of the Schrödinger equation, but one which gives us a new way to move forward.

The Green's Function

Let's now write down the Lippmann-Schwinger equation for our Schrödinger equation (1.54). We want the inverse operator $(E - H_0)^{-1}$. But this is precisely what we call the Green's function G_0. It obeys

$$\left(E + \frac{\hbar^2}{2m}\nabla^2\right) G_0(E; \mathbf{r}, \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}')$$

The formulae will be somewhat simpler if we scale out the factor $\hbar^2/2m$. We write

$$E = \frac{\hbar^2 k^2}{2m}$$

so that

$$\left(\nabla^2 + k^2\right) G_0(k; \mathbf{r}, \mathbf{r}') = \frac{2m}{\hbar^2}\delta(\mathbf{r} - \mathbf{r}') \tag{1.58}$$

We can solve for this Green's function using the Fourier transform. First, we note that translational invariance ensures that $G_0(k; \mathbf{r}, \mathbf{r}') = G_0(k; \mathbf{r} - \mathbf{r}')$. Then we define the Fourier transform

$$\tilde{G}_0(k; \mathbf{q}) = \int d^3x \; e^{-i\mathbf{q}\cdot\mathbf{x}} G_0(k; \mathbf{x}) \quad \Rightarrow \quad G_0(k; \mathbf{x}) = \int \frac{d^3q}{(2\pi)^3} \; e^{i\mathbf{q}\cdot\mathbf{x}} \tilde{G}_0(k; \mathbf{q})$$

Plugging this into our formula (1.58), we have

$$(-q^2 + k^2)\tilde{G}(k; \mathbf{q}) = \frac{2m}{\hbar^2} \quad \Rightarrow \quad \tilde{G}_0(k; \mathbf{q}) = -\frac{2m}{\hbar^2}\frac{1}{q^2 - k^2}$$

So it's simple to get the Green's function in momentum space. Now we must invert it. We have

$$G_0(k; \mathbf{x}) = -\frac{2m}{\hbar^2}\int \frac{d^3q}{(2\pi)^3}\frac{e^{i\mathbf{q}\cdot\mathbf{x}}}{q^2 - k^2}$$

Here we run into the ambiguity that we promised above. When we do the integral over $\mathbf{q}$, we run into a singularity whenever $q^2 = k^2$. To define the integral, when we integrate over $q = |\mathbf{q}|$, we should define a contour in the complex q plane which skips around the pole. We do this through the so-called "$i\epsilon$ prescription" which, as the name suggests, replaces the integral with

$$G_0^+(k; \mathbf{x}) = -\frac{2m}{\hbar^2}\int \frac{d^3q}{(2\pi)^3}\frac{e^{i\mathbf{q}\cdot\mathbf{x}}}{q^2 - k^2 - i\epsilon}$$

Where we subsequently take $\epsilon \to 0^+$. This shifts the pole slightly off the real q axis.

The simplest way to do this integral is to go to polar coordinates for the q variable. We have

$$\begin{aligned} G_0^+(k;\mathbf{x}) &= -\frac{2m}{\hbar^2}\frac{1}{(2\pi)^3}\int_0^{2\pi} d\phi \int_{-1}^{+1} d(\cos\theta)\int_0^\infty dq\ \frac{q^2\, e^{iqx\cos\theta}}{q^2-k^2-i\epsilon} \\ &= -\frac{2m}{\hbar^2}\frac{1}{(2\pi)^2}\int_0^\infty dq\ \frac{q}{ix}\frac{e^{iqx}-e^{-iqx}}{q^2-k^2-i\epsilon} \\ &= -\frac{2m}{\hbar^2}\frac{1}{(2\pi)^2}\frac{1}{ix}\int_{-\infty}^\infty dq\ \frac{qe^{iqx}}{(q-k-i\epsilon)(q+k+i\epsilon)} \end{aligned}$$

where we're allowed to factorise the denominator in this way, with $k > 0$, only because we're ultimately taking $\epsilon \to 0^+$. We can now complete the derivation by contour integral. Since $x > 0$, we can complete the contour in the upper half-plane, picking up the residue from the pole at $q = k + i\epsilon$. This gives our final answer,

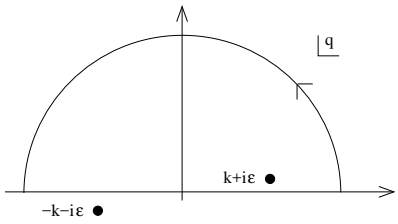


Figure 16:

$$G_0^+(k;\mathbf{r}-\mathbf{r}') = -\frac{2m}{\hbar^2}\frac{1}{4\pi}\frac{e^{+ik|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} \qquad (1.59)$$

Note that had we chosen to add $+i\epsilon$ rather than $-i\epsilon$ to the denominator, we would find the alternative Green's function $G_0^-(k;\mathbf{x}) \sim e^{-ikx}/4\pi x$. We will justify the choice of G_0^+ below.

Our Lippmann-Schwinger Equation

To finally write down the Lippmann-Schwinger equation, we need to determine the state $|\phi\rangle$ which is annihilated by $E - H_0$. But, for us, this is simply the plane wave solution

$$\phi(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}$$

We can now write the formal Lippmann-Schwinger equation (1.57) in more concrete form. It becomes

$$\psi(k;\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} - \frac{2m}{\hbar^2}\int d^3r'\ \frac{e^{+ik|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}\,V(\mathbf{r}')\psi(k;\mathbf{r}') \qquad (1.60)$$

It is simple to check that acting on this equation with the operator $(\nabla^2 + k^2)$ indeed brings us back to the original Schrödinger equation (1.54). The Lippmann-Schwinger equation is an integral equation, a reformulation of the more familiar Schrödinger differential equation. It is not solution to the Schrödinger equation because we still have to figure out what ψ is. We'll offer a strategy for doing this in Section 1.3.1.

The equation (1.60) has a very natural interpretation. The first term is simply the ingoing wave with momentum $\hbar\mathbf{k}$. The second term is the scattered wave. Note that the factor $e^{ik|\mathbf{r}-\mathbf{r}'|}$ tells us that this wave is moving outwards from the point $\mathbf{r}'$. Had we instead chosen the Green's function G_0^-, we would have found a wave moving inwards from infinity of the form $e^{-ik|\mathbf{r}-\mathbf{r}'|}$. This is unphysical. This is the reason that we pick the $-i\epsilon$ prescription rather than $+i\epsilon$.

To make contact with our earlier discussion of scattering, we look at the asymptotic form of this outgoing wave at $r \to \infty$. For this to work, we'll assume that $V(r')$ has support only in some finite region. We can then take the limit $r \gg r'$ and expand

$$|\mathbf{r}-\mathbf{r}'| = \sqrt{r^2 - 2\mathbf{r}\cdot\mathbf{r}' + r'^2} \approx r - \frac{\mathbf{r}\cdot\mathbf{r}'}{r}$$

With $V(\mathbf{r}')$ localised within some region, it makes sense to perform this expansion inside the integral. In this approximation the Green's function (1.59) can be written as

$$G_0^+(k;\mathbf{r}-\mathbf{r}') \approx -\frac{2m}{\hbar^2}\frac{1}{4\pi}\frac{e^{+ikr}}{r}e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}'}$$

and the Lippmann-Schwinger equation then becomes

$$\psi(k;\mathbf{r}) \sim e^{i\mathbf{k}\cdot\mathbf{r}} - \frac{2m}{\hbar^2}\frac{1}{4\pi}\left[\int d^3r'\; e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}'}V(\mathbf{r}')\psi(k;\mathbf{r}')\right]\frac{e^{ikr}}{r}$$

Although we derived this by assuming that $V(\mathbf{r})$ has compact support, we can actually be a little more relaxed about this. The same result holds if we require that $V(r') \to 0$ suitably quickly as $r' \to \infty$. Any potential which falls off exponentially, or as a power-law $V(r) \sim 1/r^n$ with $n \geq 2$, can be treated in this way. Note, however, that this excludes the Coulomb potential. We will deal with this separately in Section 1.4.

If we set the ingoing wave to be along the z-axis, $\mathbf{k} = k\hat{\mathbf{z}}$, then this takes the asymptotic form (1.25) that we discussed previously

$$\psi(\mathbf{r}) \sim e^{ikz} + f(\theta,\phi)\frac{e^{ikr}}{r} \tag{1.61}$$

The upshot of this analysis is that we identify the scattering amplitude as

$$f(\theta,\phi) = -\frac{2m}{\hbar^2}\frac{1}{4\pi}\int d^3r'\; e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}'}V(\mathbf{r}')\psi(k;\mathbf{r}')$$

where θ and ϕ are the usual polar angles such that $\hat{\mathbf{r}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. This gives a simple way to compute the scattering amplitude, but only if we already know the form of the wavefunction $\psi(\mathbf{r}')$ in the scattering region where $V(\mathbf{r}') \neq 0$. Our next task is to figure out how to compute $\psi(\mathbf{r}')$.

An Equation for Bound States

Above we've focussed on scattering states with energy $E = \hbar^2 k^2/2m > 0$. However, it is not difficult to repeat everything for bound states with energy $E = -\hbar^2\lambda^2/2m$. Indeed, in this case there is no ambiguity in the definition of the Green's function. We find that bound states must obey the integral equation

$$\psi(\mathbf{r}) = \frac{2m}{\hbar^2}\int d^3r' \frac{e^{-\lambda|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}\, V(\mathbf{r}')\psi(\mathbf{r}')$$

We won't attempt to solve this equation; instead our interest will focus on the Lippmann-Schwinger equation for scattering states (1.60).

1.3.1 The Born Approximation

In this section we describe a perturbative solution to the Lippmann-Schwinger equation,

$$\psi(k;\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} + \int d^3r'\; G_0^+(k;\mathbf{r}-\mathbf{r}')\, V(\mathbf{r}')\psi(k;\mathbf{r}') \tag{1.62}$$

This solution is known as the *Born series*.

We write ψ as a series expansion

$$\psi(\mathbf{r}) = \sum_{n=0}^{\infty}\phi_n(\mathbf{r}) \tag{1.63}$$

where we take the leading term to be the plane wave

$$\phi_0(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}$$

This series solves (1.62) if the ϕ_n obey the recursion relation

$$\phi_{n+1}(\mathbf{r}) = \int d^3r'\; G_0^+(k;\mathbf{r}-\mathbf{r}')\, V(\mathbf{r}')\phi_n(\mathbf{r}')$$

We will not be very precise here about the convergent properties of this series. Roughly speaking, things will work nicely if the potential V is small, so each successive term is smaller than those preceding it.

The *Born approximation* consists of taking just the leading order term ϕ_1 in this expansion. (Strictly speaking this is the first Born approximation; the n^{th} Born approximation consists of truncating the series at the n^{th} term.) This is

$$\psi(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} - \frac{2m}{\hbar^2}\frac{1}{4\pi}\left[\int d^3r'\; e^{i\mathbf{q}\cdot\mathbf{r}'}\, V(\mathbf{r}')\right]\frac{e^{ikr}}{r} \tag{1.64}$$

where

$$\mathbf{q} = \mathbf{k} - k\hat{\mathbf{r}}$$

can be thought of as the momentum transferred from the incoming wave to the outgoing wave. With this in mind, it's traditional to define the momentum of the outgoing wave to be

$$\mathbf{k}' = k\hat{\mathbf{r}}$$

so that $\mathbf{q} = \mathbf{k} - \mathbf{k}'$. Comparing the Born approximation (1.64) to the asymptotic form (1.61), we see that the scattering amplitude is simply the Fourier transform of the potential,

$$f(\theta,\phi) \approx f_0(\theta,\phi) = -\frac{2m}{\hbar^2}\frac{1}{4\pi}\left[\int d^3r'\; e^{i\mathbf{q}\cdot\mathbf{r}'}\, V(\mathbf{r}')\right] \equiv -\frac{m}{2\pi\hbar^2}\tilde{V}(\mathbf{q})$$

Note that the scattering amplitude is a function of θ and ϕ, but these variables are somewhat hidden on the notation of the right-hand side. They're sitting in the definition of $\mathbf{q}$, with $\mathbf{k}\cdot\mathbf{k}' = k^2\cos\theta$, and the variable ϕ determining the relative orientation as shown in the figure. As we've seen before, for a central potential $V(\mathbf{r}) = V(r)$, the resulting scattering amplitude will be independent of ϕ. Because the angular variables are somewhat disguised, the scattering amplitude is sometimes written as $f(\mathbf{k},\mathbf{k}')$ instead of $f(\theta,\phi)$. Indeed, we'll adopt this notation in Section 3.4.

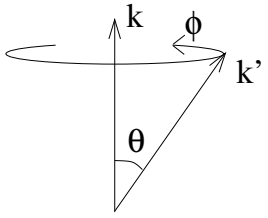


Figure 17:

Finally, the cross-section in the Born approximation is simply

$$\frac{d\sigma}{d\Omega} \approx |f_0|^2 = \left(\frac{m}{2\pi\hbar^2}\right)^2 |\tilde{V}(\mathbf{q})|^2 \tag{1.65}$$

There's some physics in this simple formula. Suppose that your potential has some short-distance structure on scales $\sim L$. Then the Fourier transform $\tilde{V}(\mathbf{q})$ is only sensitive to this when the momentum transfer is of order $q \sim 1/L$. This is a manifestation of the uncertainty principle: if you want to probe short distance physics, you need high momentum transfer.

1.3.2 The Yukawa Potential and the Coulomb Potential

At long distances, the strong nuclear force between, say, a proton and a neutron is well modelled by the Yukawa potential

$$V(r) = \frac{Ae^{-\mu r}}{r}$$

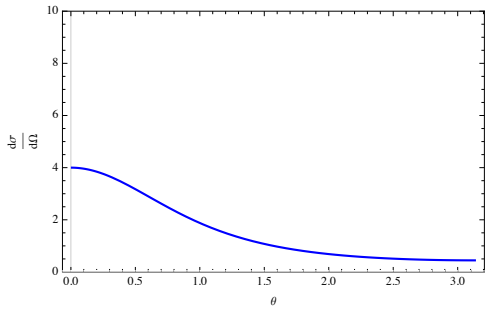


Figure 18: The cross-section for the Yukawa potential...

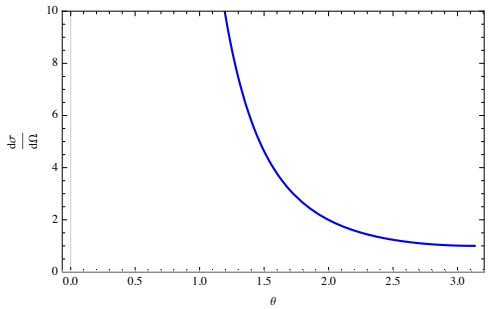


Figure 19: ...and for the Coulomb potential.

where $1/\mu$ is said to be the *range of the force*. We can compute the Fourier transform using the same kind of contour methods that we used in the previous section. We have

$$\tilde{V}(\mathbf{q}) = \frac{4\pi A}{q^2 + \mu^2}$$

Writing this in terms of the scattering angle θ, we recall that $\mathbf{q} = \mathbf{k} - \mathbf{k}'$ with $\mathbf{k}' = k\hat{\mathbf{r}}$, so that

$$q^2 = 2k^2 - 2\mathbf{k}\cdot\mathbf{k}' = 2k^2(1 - \cos\theta) = 4k^2\sin^2(\theta/2)$$

If we translate from momentum k to energy $E = \hbar^2k^2/2m$, then from (1.65), we have the leading order contribution to the cross-section for the Yukawa potential given by

$$\frac{d\sigma}{d\Omega} = \left(\frac{2Am}{\hbar^2\mu^2 + 8mE\sin^2(\theta/2)}\right)^2 \tag{1.66}$$

This is shown in the left-hand figure (for values $A = m = \hbar\mu = 1$ and $E = 1/4$).

An Attempt at Rutherford Scattering

It's tempting to look at what happens when $\mu \to 0$, so that the Yukawa force becomes the Coulomb force. For example, for electron-electron or proton-proton scattering, the strength of the Coulomb force is $A = e^2/4\pi\epsilon_0$. In this case, the cross-section (1.66) becomes,

$$\frac{d\sigma}{d\Omega} = \left(\frac{A}{4E}\right)^2 \frac{1}{\sin^4(\theta/2)} \tag{1.67}$$

This is shown in the right-hand figure (with the same values). Note that there is an enhancement of the cross-section at all scattering angles, but a divergence at forward scattering.

Rather remarkably, the quantum result (1.67) agrees with the classical cross-section that we found in (1.22)! This is a surprise and is special to the Coulomb potential. Rutherford was certainly a great scientist but, like many other great scientists before him, he had his fair share of luck.

In fact, Rutherford's luck ran deeper than you might think. It turns out that the Born approximation is valid for the Yukawa potential in certain regimes, but is never valid for the Coulomb potential! The difficulty stems from the long range nature of the Coulomb force which means that the plane wave solutions $\phi_0 \sim e^{i\mathbf{k}\cdot\mathbf{r}}$ are never really good approximations to the asymptotic states. We will describe the correct treatment of the Coulomb potential in Section 1.4 where we will see that, although our approximation wasn't valid, the result (1.67) is correct after all.

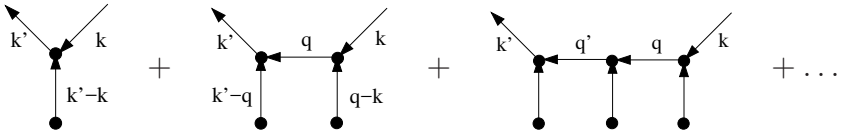
1.3.3 The Born Expansion

One can continue the Born expansion to higher orders. In compressed notation, the solution (1.63) takes the form

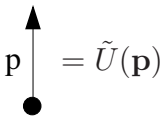
$$\psi = \phi_0 + \int G_0^+ V\phi_0 + \int\int G_0^+ V G_0^+ V\phi_0 + \int\int\int G_0^+ V G_0^+ V G_0^+ V\phi_0 + \ldots$$

This has a natural interpretation. The first term describes the incident plane wave which doesn't scatter at all. The second term describes the wave scattering once of the potential, before propagating by G_0^+ to the asymptotic regime. The third term describes the wave scattering off the potential, propagating some distance by G_0^+ and then scattering for a second time before leaving the region with the potential. In general, the term with n copies of V should be thought of as the wave scattering n times from the potential region.

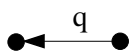
There's a useful diagrammatic way to write the resulting scattering amplitude. It is given by

$f(\mathbf{k},\mathbf{k}') =$ 

Each diagram is shorthand for an integral. Every black dot describes an insertion



while each line describes an insertion of



$$= \frac{-1}{q^2 - k^2 - i\epsilon}$$

Meanwhile, for each internal line we include the integral

$$-\frac{1}{4\pi}\int \frac{d^3q}{(2\pi)^3}$$

Although we're dealing with wave scattering, it's tempting to think of the lines as describing the trajectory of a particle. Indeed, this diagrammatic picture is a precursor to Feynman diagrams that occur in quantum field theory, where there's a much closer connection to the underlying particles.

1.4 Rutherford Scattering

> "How can a fellow sit down at a table and calculate something that would take me – *me* – six months to measure in a laboratory?"
>
> *Ernest Rutherford*

Historically, some of the most important scattering problems in particle physics involved the Coulomb potential. This is the problem of Rutherford scattering. Yet, as we mentioned above, none of the techniques that we've mentioned so far are valid for the Coulomb potential. This is mitigated somewhat by the fact that we get the right answer whether we work classically (1.22) or using the Born approximation (1.67). Nonetheless, this is a little unsatisfactory. After all, how do we know that this is the right answer!

Here we show how to do Rutherford scattering properly. We want to solve the Schrödinger equation

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + \frac{A}{r}\right)\psi(\mathbf{r}) = E\psi(\mathbf{r})$$

where $A > 0$ for repulsive interactions and $A < 0$ for attractive interactions. It will prove useful to rewrite this as

$$\left(\nabla^2 + k^2 - \frac{2\gamma k}{r}\right)\psi(\mathbf{r}) = 0 \tag{1.68}$$

where, as usual, $E = \hbar^2k^2/2m$ while $\gamma = mA/\hbar^2k$ is a dimensional parameter which characterises the strength of the Coulomb force.

The Asymptotic Form of the Wavefunction

Let's start by understanding what the wavefunctions look like asymptotically. Repeating the analysis of Section 1.2.3, the radial wavefunction $R_l(r)$ satisfies

$$\left(\frac{d^2}{dr^2} + \frac{2}{r}\frac{d}{dr} + k^2 - \frac{l(l+1)}{r^2} - \frac{2\gamma k}{r}\right) R_l(r) = 0$$

Already here we can see what the issue is. At large distances, $r \to \infty$, the Coulomb force is more important than the angular momentum barrier. We saw in previous sections that when $\gamma = 0$, the asymptotic form of the wavefunction is given by $R_l(r) = e^{\pm ikr}/r$ regardless of the value of l. However, when $\gamma \neq 0$ we have to revisit this conclusion.

With the previous solution in mind, we will look for solutions which asymptotically take the form

$$R_l(r) \sim \frac{e^{\pm ikr + g(r)}}{r}$$

for some function $g(r)$. Inserting this ansatz, we find that $g(r)$ must satisfy

$$\frac{d^2g}{dr^2} + \left(\frac{dg}{dr}\right)^2 \pm 2ik\frac{dg}{dr} = \frac{2\gamma k}{r}$$

But, for now, we care only about the asymptotic expression where the left-hand side is dominated by the last term. We then have

$$\pm i\frac{dg}{dr} = \frac{\gamma}{r} \quad \text{as} \quad r \to \infty$$

which is solved, up to some constant, by $g = \mp i\gamma \log(kr)$. Clearly this diverges as $r \to \infty$ and so should be included in the asymptotic form. We learn that asymptotically the radial wavefunctions take the form

$$R_l(r) \sim \frac{e^{\pm i(kr - \gamma \log(kr))}}{r}$$

This extra logarithm in the phase of the wavefunction means that the whole framework we described previously needs adjusting.

Note that this same analysis tells us that our previous formalism for scattering works fine for any potential $V(r) \sim 1/r^n$ with $n \geq 2$. It is just the long-range Coulomb potential that needs special treatment.

1.4.1 The Scattering Amplitude

To compute the amplitude for Rutherford scattering, we don't need any new conceptual ideas. But we do need to invoke some technical results about special functions. This is because the solution to the Schrödinger equation (1.68) can be written as

$$\psi(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} e^{-\pi\gamma/2}\Gamma(1+i\gamma)\,{}_1F_1\big(-i\gamma;\,1;\,i(kr-\mathbf{k}\cdot\mathbf{r})\big)$$

where ${}_1F_1\big(a;b;w\big)$ is the confluent hypergeometric function, defined by the series expansion

$${}_1F_1\big(a;b;w\big) = 1 + \frac{a}{b}w + \frac{a(a+1)}{b(b+1)}\frac{w^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{w^3}{3!} + \ldots$$

We won't prove that this is a solution to the Schrödinger equation. Moreover, the only fact we'll need about the hypergeometric function is its expansion for large $|w|$. For our solution, this is an expansion in $1/(kr-\mathbf{k}\cdot\mathbf{r})$ and so is valid at large distance, but not along the direction of the incident beam $\mathbf{k}$. If we take $\mathbf{k} = k\hat{\mathbf{z}}$, we have

$$\psi(\mathbf{r}) \sim e^{ikz+i\gamma\log(k(r-z))} - \frac{\gamma}{k(r-z)}\frac{\Gamma(1+i\gamma)}{\Gamma(1-i\gamma)}e^{ikr-i\gamma\log(k(r-z))} + \ldots$$

where the $+\ldots$ are corrections to both terms which are suppressed by $1/k(r-z)$. This is now very similar to our usual asymptotic form (1.61), but with the corrected phases. The first term describes the ingoing wave, the second term the scattered outgoing wave. We can therefore write

$$\psi(\mathbf{r}) \sim e^{ikz+i\gamma\log(k(r-z))} + f(\theta)\,\frac{e^{ikz-i\gamma\log(k(r-z))}}{r}$$

where the scattering amplitude is given by

$$f(\theta) = -\frac{\gamma}{k}\frac{\Gamma(1+i\gamma)}{\Gamma(1-i\gamma)}\frac{r}{r-z} = -\frac{\gamma}{2k}\frac{\Gamma(1+i\gamma)}{\Gamma(1-i\gamma)}\frac{1}{\sin^2(\theta/2)} \tag{1.69}$$

We learn that the cross-section is

$$\frac{d\sigma}{d\Omega} = |f(\theta)|^2 = \left(\frac{mA}{2\hbar^2k^2}\right)^2 \frac{1}{\sin^4(\theta/2)}$$

This is the same result as we saw using the invalid Born approximation (1.67) and the same result that we saw from a classical analysis (1.22). This shouldn't give you the wrong idea. In most situations if you use the wrong method you will get the wrong answer! The Coulomb potential is an exception.

Recovering the Hydrogen Atom

There's a rather nice exercise we can do with the scattering amplitude (1.69). When $\gamma < 0$, the Coulomb potential is attractive and has bound states. Moreover, these bound states are simply those of the hydrogen atom that we met in our first course on quantum mechanics. From our earlier analysis, we should be able to recover this from the poles in the scattering amplitude.

These arise from the gamma function $\Gamma(z)$ which has no zeros, but has poles at $z = 0, -1, -2, \ldots$. The scattering amplitude therefore has poles when

$$1 + i\gamma = -(n-1) \quad \Rightarrow \quad k = -i\,\frac{mA}{\hbar^2}\frac{1}{n} \qquad \text{with} \;\; n = 1, 2, 3, \ldots$$

For an attractive potential with $A < 0$, these poles lie along the positive imaginary k-axis, as they should. We see that they correspond to bound states with energy

$$E_n = \frac{\hbar^2 k^2}{2m} = -\frac{mA^2}{2\hbar^2}\frac{1}{n^2}$$

This, of course, is the familiar spectrum of the hydrogen atom.

2. Approximation Methods

Physicists have a dirty secret: we're not very good at solving equations. More precisely, humans aren't very good at solving equations. We know this because we have computers and they're much better at solving things than we are.

We usually do a good job of hiding this secret when teaching physics. In quantum physics we start with examples like the harmonic oscillator or the hydrogen atom and then proudly demonstrate how clever we all are by solving the Schrödinger equation exactly. But there are very very few examples where we can write down the solution in closed form. For the vast majority of problems, the answer is something complicated that isn't captured by some simple mathematical formula. For these problems we need to develop different tools.

You already met one of these tools in an earlier course: it's called *perturbation theory* and it's useful whenever the problem we want to solve is, in some sense, close to one that we've already solved. This works for a surprisingly large number of problems. Indeed, one of the arts of theoretical physics is making everything look like a coupled harmonic oscillator so that you can use perturbation theory. But there are also many problems for which perturbation theory fails dismally and we need to find another approach. In general, there's no panacea, no universal solution to all problems in quantum mechanics. Instead, the best we can hope for is to build a collection of tools. Then, whenever we're faced with a new problem we can root around in our toolbox, hoping to find a method that works. The purpose of this chapter is to stock up your toolbox.

2.1 The Variational Method

The *variational method* provides a simple way to place an upper bound on the ground state energy of any quantum system and is particularly useful when trying to demonstrate that bound states exist. In some cases, it can also be used to estimate higher energy levels too.

2.1.1 An Upper Bound on the Ground State

We start with a quantum system with Hamiltonian H. We will assume that H has a discrete spectrum

$$H|n\rangle = E_n|n\rangle \qquad n = 0, 1, \ldots$$

with the energy eigenvalues ordered such that $E_n \leq E_{n+1}$. The simplest application of the variational method places an upper bound on the value of the ground state energy E_0.

Theorem: Consider an arbitrary state $|\psi\rangle$. The expected value of the energy obeys the inequality

$$\langle E\rangle = \langle\psi|H|\psi\rangle \geq E_0$$

Proof: The proposed claim is, hopefully, intuitive and the proof is straightforward. We expand $|\psi\rangle = \sum_n a_n|n\rangle$ with $\sum_n |a_n|^2 = 1$ to ensure that $\langle\psi|\psi\rangle = 1$. Then

$$\begin{aligned}\langle E\rangle &= \sum_{n,m=0}^{\infty} a_m^\star a_n \langle m|H|n\rangle = \sum_{n,m=0}^{\infty} a_m^\star a_n E_n \delta_{mn} \\ &= \sum_{n=0}^{\infty} |a_n|^2 E_n = E_0 \sum_{n=0}^{\infty} |a_n|^2 + \sum_{n=0}^{\infty} |a_n|^2 (E_n - E_0) \geq E_0\end{aligned}$$

In the case of a non-degenerate ground state, we have equality only if $a_0 = 1$ which implies $a_n = 0$ for all $n \neq 0$. □

Now consider a family of states, $|\psi(\alpha)\rangle$, depending on some number of parameters α_i. If we like, we can relax our assumption that the states are normalised and define

$$E(\alpha) = \frac{\langle\psi(\alpha)|H|\psi(\alpha)\rangle}{\langle\psi(\alpha)|\psi(\alpha)\rangle}$$

This is sometimes called the *Rayleigh-Ritz quotient*. We still have

$$E(\alpha) \geq E_0 \quad \text{for all } \alpha$$

The most stringent bound on the ground state energy comes from the minimum value of $E(\alpha)$ over the range of α. This, of course, obeys

$$\left.\frac{\partial E}{\partial \alpha_i}\right|_{\alpha=\alpha^\star} = 0$$

giving us the upper bound $E_0 \leq E(\alpha_\star)$. This is the essence of the variational method.

The variational method does not tell us how far above the ground state $E(\alpha_\star)$ lies. It would be much better if we could also get a lower bound for E_0 so that we can say for sure that ground state energy sits within a particular range. However, for particles moving in a general potential $V(\mathbf{x})$, the only lower bound that is known is $E_0 > \min V(\mathbf{x})$. Since we're often interested in potentials like $V(\mathbf{x}) \sim -1/r$, which have no lower bound this is not particularly useful.

Despite these limitations, when used cleverly by choosing a set of states $|\psi(\alpha)\rangle$ which are likely to be fairly close to the ground state, the variational method can give remarkably accurate results.

An Example: A Quartic Potential

Consider a particle moving in one-dimension in a quartic potential. The Hamiltonian, written in units where everything is set to one, is

$$H = -\frac{d^2}{dx^2} + x^4$$

Unlike the harmonic oscillator, this problem does not a have simple solution. Nonetheless, it is easy to solve numerically where one finds

$$E_0 \approx 1.06$$

Let's see how close we get with the variational method. We need to cook up a trial wavefunction which we think might look something like the true ground state. The potential is shown on the right and, on general grounds, the ground state wavefunction should have support where the potential is smallest; an example is shown in orange. All we need to do is write down a function which has vaguely this shape. We will take

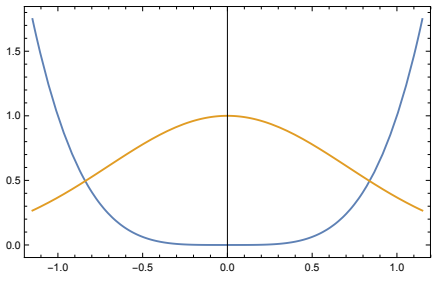


Figure 20:

$$\psi(x;\alpha) = \left(\frac{\alpha}{\pi}\right)^{1/4} e^{-\alpha x^2/2}$$

where the factor in front ensures that this wavefunction is normalised. You can check that this isn't an eigenstate of the Hamiltonian. But it does have the expected crude features of the ground state: e.g. it goes up in the middle and has no nodes. (Indeed, it's actually the ground state of the harmonic oscillator). The expected energy is

$$E(\alpha) = \sqrt{\frac{\alpha}{\pi}} \int dx\ (\alpha - \alpha^2 x^2 + x^4) e^{-\alpha x^2} = \frac{\alpha}{2} + \frac{3}{4\alpha^2}$$

The minimum value occurs at $\alpha_\star^3 = 3$, giving

$$E(\alpha_\star) \approx 1.08$$

We see that our guess does pretty well, getting within 2% of the true value. You can try other trial wavefunctions which have the same basic shape and see how they do.

How Accurate is the Variational Method?

Formally, we can see why a clever application of the variational method will give a good estimate of the ground state energy. Suppose that the trial wavefunction which minimizes the energy differs from the true ground state by

$$|\psi(\alpha_\star)\rangle = \frac{1}{\sqrt{1+\epsilon^2}}\left(|0\rangle + \epsilon|\phi\rangle\right)$$

where $|\phi\rangle$ is a normalised state, orthogonal to the ground state, $\langle 0|\phi\rangle = 0$, and ϵ is assumed to be small. Then our guess at the energy is

$$E(\alpha_\star) = \frac{1}{1+\epsilon^2}\left[\langle 0|H|0\rangle + \epsilon(\langle 0|H|\phi\rangle + \langle\phi|H|0\rangle) + \epsilon^2\langle\phi|H|\phi\rangle\right]$$

Importantly the terms linear in ϵ vanish. This is because $\langle\phi|H|0\rangle = E_0\langle\phi|0\rangle = 0$. We can then expand the remaining terms as

$$E(\alpha_\star) = E_0 + \epsilon^2\left(\langle\phi|H|\phi\rangle - E_0\right) + \mathcal{O}(\epsilon^2)$$

This means that if the difference from the true ground state is $\mathcal{O}(\epsilon)$, then the difference from the ground state energy is $\mathcal{O}(\epsilon^2)$. This is the reason that the variational method often does quite well.

Nonetheless, one flaw with the variational method is that unless someone tells us the true answer, we have no way of telling how good our approximation is. Or, in the language above, we have no way of estimating the size of ϵ. Despite this, we will see below that there are some useful things we can do with it.

2.1.2 An Example: The Helium Atom

One important application of quantum mechanics is to explain the structure of atoms. Here we will look at two simple approaches to understand an atom with two electrons. This atom is helium.

The Hamiltonian for two electrons, each of charge $-e$, orbiting a nucleus of charge Ze is

$$H = \frac{\mathbf{p}_1^2}{2m} - \frac{Ze^2}{4\pi\epsilon_0}\frac{1}{r_1} + \frac{\mathbf{p}_2^2}{2m} - \frac{Ze^2}{4\pi\epsilon_0}\frac{1}{r_2} + \frac{e^2}{4\pi\epsilon_0}\frac{1}{|\mathbf{x}_1 - \mathbf{x}_2|} \tag{2.1}$$

For helium, $Z = 2$ but, for reasons that will become clear, we will leave it arbitrary and only set it to $Z = 2$ at the end of the calculation.

If we ignore the final term, then this Hamiltonian is easy to solve: it simply consists of two independent copies of the hydrogen atom. The eigenstates would be

$$\Psi(\mathbf{x}_1, \mathbf{x}_2) = \psi_{n_1,l_1,m_1}(\mathbf{x}_1)\psi_{n_2,l_2,m_2}(\mathbf{x}_2)$$

where $\psi_{n,l,m}(r)$ are the usual energy eigenstates of the hydrogen atom. We should remember that the electrons are fermions so we can't put them in the same state. However, electrons also have a spin degree of freedom which we have neglected above. This means that two electrons can have the same spatial wavefunction as long as one is spin up and the other spin down.

Ignoring the interaction term between electrons gives the energy

$$E = -Z^2 \left(\frac{1}{n_1^2} + \frac{1}{n_2^2} \right) Ry \tag{2.2}$$

where Ry is the *Rydberg constant*, given by

$$Ry = \frac{me^4}{32\pi^2\epsilon_0^2\hbar^2} \approx 13.6\ eV$$

Setting $Z = 2$ and $n_1 = n_2 = 1$, this very naive approach suggests that the ground state of helium has energy $E_0 = -8\,Ry \approx -109\,eV$. The true ground state of helium turns out to have energy

$$E_0 \approx -79.0\ eV \tag{2.3}$$

Our task is to find a method to take into account the final, interaction term between electrons in (2.1) and so get closer to the true result (2.3) Here we try two alternatives.

Perturbation Theory

Our first approach is to treat the Coulomb energy between two electrons as a perturbation on the original problem. Before proceeding, there is a question that we should always ask in perturbation theory: what is the small, dimensionless parameter that ensures that the additional term is smaller than the original terms?

For us, we need a reason to justify why the last term in the Hamiltonian (2.1) is likely to be smaller than the other two potential terms. All are due to the Coulomb force, so come with a factor of $e^2/4\pi\epsilon_0$. But the interactions with the nucleus also come with a factor of Z. This is absent in the electron-electron interaction. This, then, is what we hang our hopes on: the perturbative expansion will be an expansion in $1/Z$. Of course, ultimately we will set $1/Z = 1/2$ which is not a terribly small number. This might give us concern that perturbation theory will not be very accurate for this problem.

We now place each electron in the usual hydrogen ground state $\psi_{1,0,0}(\mathbf{x})$, adapted to general Z

$$\psi_{1,0,0}(\mathbf{x}) = \sqrt{\frac{Z^3}{\pi a_0^3}}\, e^{-Zr/a_0} \tag{2.4}$$

where a_0 is the Bohr radius, defined as

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{me^2} \approx 5 \times 10^{-11}\ m$$

To leading order, the shift of the ground state energy is given by the standard result of first order perturbation theory,

$$\Delta E = \frac{e^2}{4\pi\epsilon_0}\int d^3x_1 d^3x_2\ \frac{|\psi_{1,0,0}(\mathbf{x}_1)|^2|\psi_{1,0,0}(\mathbf{x}_2)|^2}{|\mathbf{x}_1 - \mathbf{x}_2|}$$

We need to compute this integral.

The trick is to pick the right coordinate system. We will work in spherical polar coordinates for both particles. However, we will choose the z axis for the second particle to lie along the direction $\mathbf{x}_1$ set by the first particle. The advantage of this choice is that the angle θ between the two particles coincides with the polar angle θ_2 for the second particle. In particular, the separation between the two particles particles can be written as

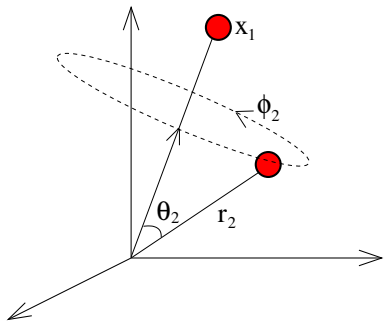


Figure 21:

$$|\mathbf{x}_1 - \mathbf{x}_2| = \sqrt{(\mathbf{x_1} - \mathbf{x}_2)^2} = \sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos\theta_2}$$

In these coordinates, it is simple to do the integration over the angular variables for the first particle, and over ϕ_2 for the second. The shift in the energy then becomes

$$\begin{aligned}\Delta E &= \frac{8\pi^2e^2}{4\pi\epsilon_0}\left(\frac{Z^3}{\pi a_0^3}\right)^2 \int dr_1\ r_1^2 e^{-2Zr_1/a_0}\int dr_2\ r_2^2 e^{-2Zr_2/a_0} \\ &\qquad\qquad \times \int_{-1}^{+1} d(\cos\theta_2)\ \frac{1}{\sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos\theta_2}} \\ &= -\frac{2\pi e^2}{\epsilon_0}\left(\frac{Z^3}{\pi a_0^3}\right)^2 \int dr_1\ r_1^2 e^{-2Zr_1/a_0}\int dr_2\ r_2^2 e^{-2Zr_2/a_0}\ \frac{\sqrt{(r_1-r_2)^2} - \sqrt{(r_1+r_2)^2}}{r_1r_2} \\ &= -\frac{2\pi e^2}{\epsilon_0}\left(\frac{Z^3}{\pi a_0^3}\right)^2 \int dr_1\ r_1^2 e^{-2Zr_1/a_0}\int dr_2\ r_2^2 e^{-2Zr_2/a_0}\ \frac{|r_1-r_2| - |r_1+r_2|}{r_1r_2}\end{aligned}$$

Those modulus signs are a little odd, but easily dealt with. Because the integral is symmetric in r_1 and r_2, the regime $r_1 > r_2$ must give the same result as the regime $r_1 < r_2$. We can then focus on one of these regimes — say $r_1 > r_2$ where $|r_1 - r_2| - |r_1 + r_2| = -2r_2$ — and just double our result. We have

$$\begin{aligned}\Delta E &= \frac{8\pi e^2}{\epsilon_0}\left(\frac{Z^3}{\pi a_0^3}\right)^2 \int_{r_2}^{\infty} dr_1\; r_1\, e^{-2Zr_1/a_0} \int_0^{\infty} dr_2\; r_2^2\, e^{-2Zr_2/a_0} \\ &= \frac{8\pi e^2}{\epsilon_0}\left(\frac{Z^3}{\pi a_0^3}\right)^2 \int_0^{\infty} dr_2\; r_2^2 \left(\frac{a_0 r_2}{2Z} + \frac{a_0^2}{4Z^2}\right) e^{-4Zr_2/a_0} \\ &= \frac{5}{8}\frac{Ze^2}{4\pi\epsilon_0 a_0} = \frac{5Z}{4}\,Ry\end{aligned}$$

Using first order perturbation, we find that the ground state energy of helium is

$$E_0 \approx E + \Delta E = \left(-2Z^2 + \frac{5Z}{4}\right) Ry \approx -74.8\ eV$$

This is much closer to the correct value of $E_0 \approx -79\ eV$. In fact, given that our perturbative expansion parameter is $1/Z = 1/2$, it's much better than we might have anticipated.

The Variational Method

We'll now try again, this time using the variational method. For our trial wavefunction we pick $\Psi(\mathbf{x}_1, \mathbf{x}_2) = \psi(\mathbf{x}_1)\psi(\mathbf{x}_2)$ where

$$\psi(\mathbf{x}; \alpha) = \sqrt{\frac{\alpha^3}{\pi a_0^3}}\, e^{-\alpha r/a_0} \tag{2.5}$$

This is almost the same as the hydrogen ground state (2.4) that we worked with above. The only difference is that we've replaced the atomic number Z with a general parameter α that we will allow to vary. We can tell immediately that this approach must do at least as well at estimating the ground state energy because setting $\alpha = Z$ reproduces the results of first order perturbation theory.

The expectation of the energy using our trial wavefunction is

$$E(\alpha) = \int d^3x_1 d^3x_2\; \psi^\star(\mathbf{x}_1)\psi^\star(\mathbf{x}_2)\, H\psi(\mathbf{x}_1)\psi(\mathbf{x}_2)$$

with H the differential operator given in (2.1). Now we have to evaluate all terms in the Hamiltonian afresh. However, there is trick we can use. We know that (2.5) is the ground state of the Hamiltonian

$$H_\alpha = \frac{\mathbf{p}^2}{2m} - \frac{\alpha e^2}{4\pi\epsilon_0}\frac{1}{r}$$

where we've replaced Z by α in the second term. With this observation, we write the helium Hamiltonian (2.1) as

$$H = H_\alpha(\mathbf{p}_1, \mathbf{r}_1) + H_\alpha(\mathbf{p}_2, \mathbf{r}_2) + \frac{e^2}{4\pi\epsilon_0}\left[(\alpha - Z)\left(\frac{1}{r_1} + \frac{1}{r_2}\right) + \frac{1}{|\mathbf{x}_1 - \mathbf{x}_2|}\right]$$

Written in this way, the expected energy becomes

$$E(\alpha) = -2\alpha^2\, Ry + \frac{e^2}{4\pi\epsilon_0}\left[2(\alpha - Z)\int d^3x\ \frac{|\psi(\mathbf{x})|^2}{r} + \int d^3x_1 d^3x_2\ \frac{|\psi(\mathbf{x}_1)|^2|\psi(\mathbf{x}_2)|^2}{|\mathbf{x}_1 - \mathbf{x}_2|}\right]$$

Here, the first term comes from the fact that our trial wavefunction is the ground state of H_α with ground state energy given by (2.2). We still need to compute the integrals in the second and third term. But both of these are straightforward. The first is

$$\int d^3x\ \frac{|\psi(\mathbf{x})|^2}{r} = 4\pi\frac{\alpha^3}{\pi a_0^3}\int dr\ re^{-2\alpha r/a_0} = \frac{\alpha}{a_0}$$

Meanwhile, the final integral is the same as we computed in our perturbative calculation. It is

$$\int d^3x_1 d^3x_2\ \frac{|\psi(\mathbf{x}_1)|^2|\psi(\mathbf{x}_2)|^2}{|\mathbf{x}_1 - \mathbf{x}_2|} = \frac{5\alpha}{8a_0}$$

Putting this together, we have

$$E(\alpha) = \left(-2\alpha^2 + 4(\alpha - Z)\alpha + \frac{5}{4}\alpha\right)\, Ry$$

This is minimized for $\alpha_\star = Z - 5/16$. The minimum value of the energy is then

$$E(\alpha_\star) = -2\left(Z - \frac{5}{16}\right)^2\, Ry \approx -77.5\, eV \tag{2.6}$$

We see that this is somewhat closer to the true value of $E_0 \approx -79.0\, eV$.

There's one last bit of physics hidden in this calculation. The optimum trial wavefunction that we ended up using was that of an electron orbiting a nucleus with charge $(Z - 5/16)e$, rather than charge Ze. This has a nice interpretation: the charge of the nucleus is screened by the presence of the other electron.

2.1.3 Do Bound States Exist?

There is one kind of question where variational methods can give a definitive answer. This is the question of the existence of bound states.

Consider a particle moving in a localised potential $V(\mathbf{x})$, such that $V(\mathbf{x}) \to 0$ as $x \to \infty$. A bound state is an energy eigenstate with $E < 0$. For some potentials, there exist an infinite number of bound states; the Coulomb potential $V = 1/r$ in three dimensions is a familiar example. For other potentials there will be only a finite number. And for some potentials there will be none. How can we tell what properties a given potential has?

Clearly the variational method can be used to prove the existence of a bound state. All we need to do is exhibit a trial wavefunction which has $E < 0$. This then ensures that the true ground state also has $E_0 < 0$.

An Example: The Hydrogen Anion

A hydrogen anion H^- consists of a single proton, with two electrons in its orbit. But does a bound state of two electrons and a proton exist?

The Hamiltonian for H^- is the same as that for helium, (2.1), but now with $Z = 1$. This means that we can import all the calculations of the previous section. In particular, our variational method gives a minimum energy (2.6) which is negative when we set $Z = 1$. This tells us that a bound state of two electrons and a proton does indeed exist.

An Example: The Yukawa Potential

The Yukawa potential in three-dimensions takes the form

$$V(r) = -A\frac{e^{-\lambda r}}{r} \tag{2.7}$$

For $A > 0$, this is an attractive potential. Note that if we set $\lambda = 0$, this coincides with the Coulomb force. However, for $\lambda \neq 0$ the Yukawa force drops off much more quickly.

The Yukawa potential arises in a number of different places in physics. Here are two examples:

- In a metal, electric charge is *screened*. This was described in Section 7.7 of the lecture notes on Electromagnetism. This causes the Coulomb potential to be replaced by the Yukawa potential.
- The strong nuclear force between a proton and a neutron is complicated. However, at suitably large distances it is well approximated by the Yukawa potential, with r the relative separation of the proton and neutron. Indeed, this is the context in which Yukawa first suggested his potential. Thus the question of whether (2.7) admits a bound state is the question of whether a proton and neutron can bind together.

A spoiler: the hydrogen atom has stable isotope known as deuterium. Its nucleus, known as the deuteron, consists of a proton and neutron. Thus, experiment tells us that a bound state must exist. We'd like to understand this theoretically, if only to be sure that the experiments aren't wrong!

The Hamiltonian is

$$H = -\frac{\hbar^2}{2m}\nabla^2 + V(r)$$

In the context of deuterium, r is the distance between the proton and neutron so m should really be interpreted as the reduced mass $m = m_p m_n/(m_p + m_n) \approx m_p/2$. We will work with a familiar trial wavefunction,

$$\psi(\mathbf{x};\alpha) = \sqrt{\frac{\alpha^3}{\pi}}e^{-\alpha r}$$

This is the ground state of the hydrogen atom. The factor in front ensures that the wavefunction is normalised: $\int d^3x\, |\psi|^2 = 1$. A short calculation shows that the expected energy is

$$E(\alpha) = \frac{\hbar^2\alpha^2}{2m} - \frac{4A\alpha^3}{(\lambda + 2\alpha)^2}$$

It's easy to check that there is a value of α for which $E(\alpha) < 0$ whenever

$$\lambda < \frac{Am}{\hbar^2}$$

This guarantees that the Yukawa potential has a bound state when the parameters lie within this regime. We cannot, however, infer the converse: this method doesn't tell us whether there is a bound state when $\lambda > Am/\hbar^2$.

It turns out that for λ suitably large, bound states do cease to exist. The simple variational method above gets this qualitative bit of physics right, but it does not do so well in estimating the bound. Numerical results tell us that there should be a bound state whenever $\lambda \lesssim 2.4Am/\hbar$.

Bound States and The Virial Theorem

There is a connection between these ideas and the virial theorem. Let's first remind ourselves what the virial theorem is this context. Suppose that we have a particle in d dimensions, moving in the potential

$$V(\mathbf{x}) = Ar^n \tag{2.8}$$

This means that the potential scales as $V(\lambda\mathbf{x}) = \lambda^n V(\mathbf{x})$. We will assume that there is a normalised ground state with wavefunction $\psi_0(\mathbf{x})$.

The ground state energy is

$$E_0 = \int d^d\mathbf{x}\ \frac{\hbar^2}{2m}|\nabla\psi_0(\mathbf{x})|^2 + V(\mathbf{x})|\psi_0(\mathbf{x})|^2 \equiv \langle T\rangle_0 + \langle V\rangle_0$$

Now consider the trial wavefunction $\psi(\mathbf{x}) = \alpha^{d/2}\psi_0(\alpha\mathbf{x})$, where the prefactor ensures that $\psi(\mathbf{x})$ continues to be normalised. From the scaling property of the potential (2.8), it is simple to show that

$$E(\alpha) = \alpha^2\langle T\rangle_0 + \alpha^{-n}\langle V\rangle_0$$

The minimum of $E(\alpha)$ is at

$$\frac{dE}{d\alpha} = 2\alpha\langle T\rangle_0 - n\alpha^{-n+1}\langle V\rangle_0 = 0$$

But this minimum must sit at $\alpha = 1$ since, by construction, this is the true ground state. We learn that for the homogeneous potentials (2.8), we have

$$2\langle T\rangle_0 = n\langle V\rangle_0 \tag{2.9}$$

This is the *virial theorem*.

Let's now apply this to our question of bound states. Here are some examples:

- $V \sim -1/r$: This is the Coulomb potential. The virial theorem tells us that $E_0 = \langle T\rangle_0 + \langle V\rangle_0 = -\langle T\rangle_0 < 0$. In other words, we proved what we already know: the Coulomb potential has bound states.

  There's a subtlety here. Nowhere in our argument of the virial theorem did we state that the potential (2.8) has $A < 0$. Our conclusion above would seem to hold for $A > 0$, yet this is clearly wrong: the repulsive potential $V \sim +1/r$ has no bound states. What did we miss? Well, we assumed right at the beginning of the argument that the ground state ψ_0 was normalisable. For repulsive potentials like $V \sim 1/r$ this is not true: all states are asymptotically plane waves of the form $e^{i\mathbf{k}\cdot\mathbf{x}}$. The virial theorem is not valid for repulsive potentials of this kind.

- $V \sim -1/r^3$: Now the virial theorem tells us that $E_0 = \frac{1}{3}\langle T\rangle_0 > 0$. This is actually a contradiction! In a potential like $V \sim 1/r^3$, any state with $E > 0$ is non-normalisable since it mixes with the asymptotic plane waves. It must be that this potential has no localised states.

This result might seem surprising. Any potential $V \sim -r^n$ with $n \leq -3$ descends steeply at the origin and you might think that this makes it efficient at trapping particles there. The trouble is that it is too efficient. The kinetic energy of the particle is not sufficient to hold it up at some finite distance, and the particle falls towards the origin. Such potentials have no bound states.

Bound States in One Dimension

There is an exact and rather pretty result that holds for particles moving in one-dimension. Consider a particle moving in a potential $V(x)$ such that $V(x) = 0$ for $|x| > L$. However, when $|x| < L$, the potential can do anything you like: it can be positive or negative, oscillate wildly or behave very calmly.

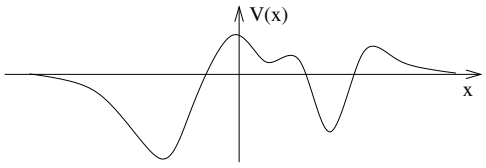


Figure 22: Does a bound state exist?

Theorem: A bound state exists whenever $\int dx\ V(x) < 0$. In other words, a bound state exists whenever the potential is "mostly attractive".

Proof: We use the Gaussian variational ansatz

$$\psi(x;\alpha) = \left(\frac{\alpha}{\pi}\right)^{1/4} e^{-\alpha x^2/2}$$

Then we find

$$E(\alpha) = \frac{\hbar^2\alpha}{4m} + \sqrt{\frac{\alpha}{\pi}}\int_{-\infty}^{\infty} dx\ V(x)e^{-\alpha x^2}$$

where the $\hbar^2\alpha/4m$ term comes from the kinetic energy. The trick is to look at the function

$$\frac{E(\alpha)}{\sqrt{\alpha}} = \frac{\hbar^2\sqrt{\alpha}}{4m} + \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} dx\ V(x)e^{-\alpha x^2}$$

This is a continuous function of α. In the limit $\alpha \to 0$, we have

$$\frac{E(\alpha)}{\sqrt{\alpha}} \ \to\ \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} dx\ V(x)$$

If $\int dx\ V(x) < 0$ then $\lim_{\alpha\to 0} E(\alpha)/\sqrt{\alpha} < 0$ and, by continuity, there must be some small $\alpha > 0$ for which $E(\alpha) < 0$. This ensures that a bound state exists. □

Once again, the converse to this statement does not hold. There are potentials with $\int dx\ V(x) > 0$ which do admit bound states.

You may wonder if we can extend this result to higher dimensions. It turns out that there is an analogous statement in two dimensions[4]. However, in three dimensions or higher there is no such statement. In that case, if the potential is suitably shallow there are no bound states.

2.1.4 An Upper Bound on Excited States

So far, we've focussed only on approximating the energy of the ground state. Can we also use the variational method to give a bound on the energy of excited states?

This is rather more tricky. We can make progress if we know the ground state $|0\rangle$ exactly. In this case, we construct a trial wavefunction $|\psi(\alpha)\rangle$ that is orthogonal to the ground state,

$$\langle\psi(\alpha)|0\rangle = 0 \quad \text{for all } \alpha \tag{2.10}$$

Now we can simply rerun our arguments of Section 2.1.1. The minimum of $E(\alpha) = \langle\psi(\alpha)|H|\psi(\alpha)\rangle$ provides an upper bound on the energy E_1 of the first excited state.

In principle, we could then repeat this argument. Working with a trial wavefunction that is orthogonal to both $|0\rangle$ and $|1\rangle$ will provide an upper bound on the energy E_2 of the second excited state.

In practice, this approach is not much use. Usually, if we're working with the variational method then it's because we don't have an exact expression for the ground state, making it difficult to construct a trial wavefunction obeying (2.10). If all we have is an approximation to the ground state, this is no good at all in providing a bound for excited states.

There is, however, one situation where we can make progress: this is if our Hamiltonian has some symmetry or, equivalently, some other conserved quantity. If we know the quantum number of the ground state under this symmetry then we can guarantee (2.10) by constructing our trial wavefunction to have a different quantum number.

An Example: Parity and the Quartic Potential

For a simple example of this, let's return to the quartic potential of Section 2.1.1. The Hamiltonian is

$$H = -\frac{d^2}{dx^2} + x^4$$

[4]More details can be found in the paper by Barry Simon, "*The bound state of weakly coupled Schrödinger operators in one and two dimensions*", Ann. Phys. 97, 2 (1976), which you can download here.

This Hamiltonian is invariant under parity, mapping $x \to -x$. The true ground state must be even under parity. We can therefore construct a class of trial wavefunctions for the first excited state which are odd under parity. An obvious choice is

$$\psi(x;\alpha) = \left(\frac{4\alpha^3}{\pi}\right)^{1/4} x\, e^{-\alpha x^2/2}$$

Churning through some algebra, one finds that the minimum energy using this wavefunction is

$$E(\alpha_\star) \approx 3.85$$

The true value is $E_1 \approx 3.80$.

3. Band Structure

In this chapter, we start our journey into the world of condensed matter physics. This is the study of the properties of "stuff". Here, our interest lies in a particular and familiar kind of stuff: solids.

Solids are collections of tightly bound atoms. For most solids, these atoms arrange themselves in regular patterns on an underlying crystalline lattice. Some of the electrons of the atom then disassociate themselves from their parent atom and wander through the lattice environment. The properties of these electrons determine many of the properties of the solid, not least its ability to conduct electricity.

One might imagine that the electrons in a solid move in a fairly random fashion, as they bounce from one lattice site to another, like a ball in a pinball machine. However, as we will see, this is not at all the case: the more fluid nature of quantum particles allows them to glide through a regular lattice, almost unimpeded, with a distorted energy spectrum the only memory of the underlying lattice.

In this chapter, we will focus on understanding how the energy of an electron depends on its momentum when it moves in a lattice environment. The usual formula for kinetic energy, $E = \frac{1}{2}mv^2 = p^2/2m$, is one of the first things we learn in theoretical physics as children. As we will see, a lattice changes this in interesting ways, the consequences of which we will explore in chapter 4.

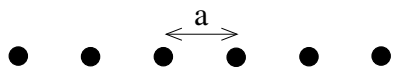
3.1 Electrons Moving in One Dimension

We begin with some particularly simple toy models which capture much of the relevant physics. These toy models describe an electron moving in a one-dimensional lattice. We'll take what lessons we can from this before moving onto more realistic descriptions of electrons moving in higher dimensions.

3.1.1 The Tight-Binding Model

The tight-binding model is a caricature of electron motion in solid in which space is made discrete. The electron can sit only on the locations of atoms in the solid and has some small probability to hop to a neighbouring site due to quantum tunnelling.

To start with our "solid" consists of a one-dimensional lattice of atoms. This is described by N points arranged along a line, each separated by distance a.



Consider a single electron moving on this lattice. We will assume that the electron can only sit on a given lattice point; it's not allowed to roam between lattice points. This is supposed to mimic the idea that electrons are bound to the atoms in a lattice and goes by the name of the *tight-binding approximation*. (We'll see exactly what we're neglecting in this approximation later.)

When the electron sits on the n^{th} atom, we denote the quantum state as $|n\rangle$. These states are considered orthogonal to each other, so

$$\langle n|m\rangle = \delta_{nm}$$

Clearly the total Hilbert space has dimension N, and is spanned by $|n\rangle$ with $n = 1, \ldots, N$.

What kind of Hamiltonian will govern the dynamics of this electron? If the electron just remains on a given atom, an appropriate Hamiltonian would be

$$H_0 = E_0 \sum_n |n\rangle\langle n|$$

Each of the position states $|n\rangle$ is an energy eigenstate of H_0 with energy E_0. The electrons governed by this Hamiltonian don't move. This Hamiltonian is boring.

To make things more interesting, we need to include the possibility that the electron can tunnel from one site to another. How to do this? Well, the Hamiltonian governs time evolution. In some small time increment of time Δt, a state evolves as

$$|\psi\rangle \mapsto |\psi\rangle - \frac{i\Delta t}{\hbar} H|\psi\rangle + \mathcal{O}(\Delta t^2)$$

This means that if we want the possibility for the electron to hop from one site to another, we should include in the Hamiltonian a term of the form $|m\rangle\langle n|$ which takes an electron at site n and moves it to an electron at site m.

There is one last ingredient that we want to feed into our model: locality. We don't want electrons to disappear and reappear many thousands of lattice spacings down the line. We want our model to describe electrons hopping from one atom to neighbouring atoms. This motivates our final form of the Hamiltonian,

$$H = E_0 \sum_n |n\rangle\langle n| - t \sum_n \Big(|n\rangle\langle n+1| + |n+1\rangle\langle n| \Big) \tag{3.1}$$

First a comment on notation: the parameter t is called the *hopping parameter*. It is not time; it is simply a number which determines the probability that a particle will hop to a neighbouring site. (More precisely, the ratio t^2/E_0^2 will determine the probability. of hopping.) It's annoying notation, but unfortunately t is the canonical name for this hopping parameter so it's best we get used to it now.

Now back to the physics encoded in H. We've chosen a Hamiltonian that only includes hopping terms between neighbouring sites. This is the simplest choice; we will describe more general choices later. Moreover, the probability of hopping to the left is the same as the probability of hopping to the right. This is required because H must be a Hermitian operator.

There's one final issue that we have to address before solving for the spectrum of H: what happens at the edges? Again, there are a number of different possibilities but none of the choices affect the physics that we're interested in here. The simplest option is simply to declare that the lattice is periodic. This is best achieved by introducing a new state $|N+1\rangle$, which sits to the right of $|N\rangle$, and is identified with $|N+1\rangle \equiv |1\rangle$.

Solving the Tight-Binding Model

Let's now solve for the energy eigenstates of the Hamiltonian (3.1). A general state can be expanded as

$$|\psi\rangle = \sum_m \psi_m |m\rangle$$

with $\psi_n \in \mathbf{C}$. Substituting this into the Schrödinger equation gives

$$H|\psi\rangle = E|\psi\rangle \quad \Rightarrow \quad E_0 \sum_m \psi_m |m\rangle - t\Big(\sum_m \psi_{m+1}|m\rangle + \psi_m |m+1\rangle\Big) = E\sum_n \psi_m |m\rangle$$

If we now take the overlap with a given state $\langle n|$, we get the set of linear equations for the coefficients ψ_n

$$\langle n|H|\psi\rangle = E\langle n|\psi\rangle \quad \Rightarrow \quad E_0\psi_n - t(\psi_{n+1} + \psi_{n-1}) = E\psi_n \tag{3.2}$$

These kind of equations arise fairly often in physics. (Indeed, they will arise again in Section 5 when we come to discuss the vibrations of a lattice.) They are solved by the ansatz

$$\psi_n = e^{ikna} \tag{3.3}$$

Or, if we want to ensure that the wavefunction is normalised, $\psi_n = e^{ikna}/\sqrt{N}$. The exponent k is called the *wavenumber*. The quantity $p = \hbar k$ plays a role similar to momentum in our discrete model; we will discuss the ways in which it is like momentum in Section 3.1.4. We'll also often be lazy and refer to k as *momentum*.

The wavenumber has a number of properties. First, the set of solutions remain the same if we shift $k \to k + 2\pi/a$ so the wavenumber takes values in

$$k \in \left[-\frac{\pi}{a}, +\frac{\pi}{a}\right) \tag{3.4}$$

This range of k is given the fancy name *Brillouin zone*. We'll see why this is a useful concept that deserves its own name in Section 3.2.

There is also a condition on the allowed values of k coming from the requirement of periodicity. We want $\psi_{N+1} = \psi_1$, which means that $e^{ikNa} = 1$. This requires that k is quantised in units of $2\pi/aN$. In other words, within the Brillouin zone (3.4) there are exactly N quantum states of the form (3.3). But that's what we expect as it's the dimension of our Hilbert space; the states (3.3) form a different basis.

States of the form (3.3) have the property that

$$\psi_{n\pm1} = e^{\pm ika}\psi_n$$

This immediately ensures that equation (3.2) is solved for any value of k, with the energy eigenvalue

$$E = E_0 - 2t\cos(ka) \tag{3.5}$$

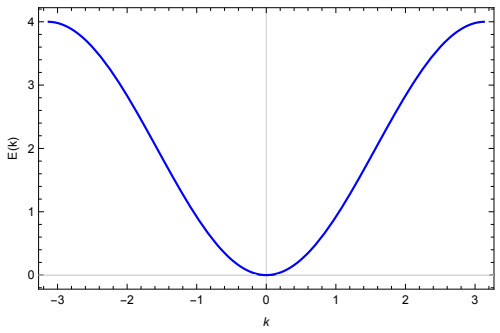


Figure 23:

The spectrum is shown in the figure for $t > 0$. (The plot was made with $a = t = 1$ and $E_0 = 2$.) The states with $k > 0$ describe electrons which move to the right; those with $k < 0$ describe electrons moving to the left.

There is a wealth of physics hiding in this simple result, and much of the following sections will be fleshing out these ideas. Here we highlight a few pertinent points

- The electrons do not like to sit still. The eigenstates $|n\rangle$ of the original Hamiltonian H_0 were localised in space. One might naively think that adding a tiny hopping parameter t would result in eigenstates that were spread over a few sites. But this is wrong. Instead, all energy eigenstates are spread throughout the whole lattice. Arbitrarily small local interactions result in completely delocalised energy eigenstates.

- The energy eigenstates of H_0 were completely degenerate. Adding the hopping term lifts this degeneracy. Instead, the eigenstates are labelled by the wavevector

k and have energies (3.5) that lie in a range $E(k) \in [E_0 - 2t, E_0 + 2t]$. This range of energies is referred to a *band* and the difference between the maximum and minimum energy (which is $4t$ in this case) is called the *band width*. In our simple model, we have just a single energy band. In subsequent models, we will see multiple bands emerging.

- For suitably small momentum, $k \ll \pi/a$, we can Taylor expand the energy (3.5) as

$$E(k) \approx (E_0 - 2t) + ta^2k^2$$

Up to a constant, this takes the same form as a free particle moving in the continuum,

$$E_{\text{free}} = \frac{\hbar^2 k^2}{2m} \tag{3.6}$$

This is telling us that low energy, low momentum particles are unaware that they are moving on an underlying lattice. Instead, they act as if they are moving along a continuous line with *effective mass* $m^\star = \hbar^2/2ta^2$. Notice that in this model the effective mass has nothing to do with the physical mass of the electron; it is inherited from properties of the lattice.

- There is a cute reciprocity between the properties of momentum and position. We know from our first course on quantum mechanics that if space is made finite — for example, a particle in a box, or a particle moving on a circle — then momentum becomes discrete. We also saw this above as the periodic boundary conditions enforced the wavenumber to be quantised in units of $2\pi/Na$.

  However, our tight-binding model also exhibits the converse phenomenon: when we make space discrete, momentum becomes periodic: it has to lie in the Brillouin zone (3.4). More generally, discreteness is the Fourier transform of compactness.

A First Look at Metals and Insulators

There's further physics to uncover if we consider more than one electron moving in the lattice. This section is just to give a flavour of these ideas; we will discuss them in more detail in Section 4.1. For simplicity, we will assume that the electrons do not interact with each other. Now the state of the system is governed by the Pauli exclusion principle: two electrons are not allowed to occupy the same state.

As we have seen, our tight-binding model contains N states. However, each electron has two internal states, spin $|\uparrow\rangle$ and spin $|\downarrow\rangle$. This means that, in total, each electron can be in one of $2N$ different states. Invoking the Pauli exclusion principle, we see that our tight-binding model makes sense as long as the number of electrons is less than or equal to $2N$.

The Pauli exclusion principle means that the ground state of a multi-electron system has interesting properties. The first two electrons that we put in the system can both sit in the lowest energy state with $k = 0$ as long as they have opposite spins. The next electron that we put in finds these states occupied; it must sit in the next available energy state which has $k = \pm 2\pi/Na$. And so this continues, with subsequent electrons sitting in the lowest energy states which have not previously been occupied. The net result is that the electrons fill all states up to some final k_F which is known as the *Fermi momentum*. The boundary between the occupied and unoccupied states in known as the *Fermi surface*. Note that it is a surface in momentum space, rather than in real space. We will describe this in more detail in Section 4.1. (See also the lectures on Statistical Physics.)

How many electrons exist in a real material? Here something nice happens, because the electrons which are hopping around the lattice come from the atoms themselves. One sometimes talks about each atom "donating" an electron. Following our chemist friends, these are called *valence electrons*. Given that our lattice contains N atoms, it's most natural to talk about the situation where the system contains ZN electrons, with Z an integer. The atom is said to have valency Z.

Suppose $Z = 1$, so we have N electrons. Then only half of the states are filled and $k_F = \pi/2a$. This is shown in the figure. Note that there are as many electrons moving to the left (with $k < 0$) as there are electrons moving to the right ($k > 0$). This is the statement that there is no current in the ground state of the system.

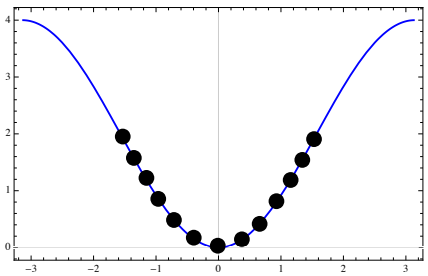


Figure 24:

We can now ask: what are the low-energy excitations of the system? We see that there are many: we can take any electron just below the Fermi surface and promote it to an electron just above the Fermi surface at a relatively small cost in energy. This becomes particularly relevant if we perturb the system slightly. For example, we could ask: what happens if we apply an electric field? As we will describe in more detail in 4.1.1, the ground

state of the system re-arranges itself at just a small cost of energy: some left-moving states below the Fermi surface become unoccupied, while right-moving states above the Fermi surface become occupied. Now, however, there are more electrons with $k > 0$ than with $k < 0$. This results in an electrical current. What we have just described is a *conductor*.

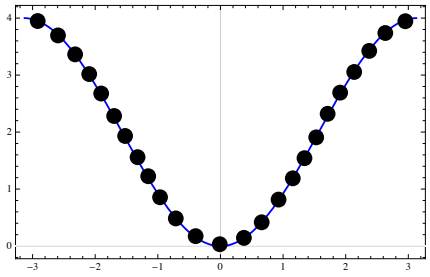


Figure 25:

Let's contrast this with what happens when we have $2N$ electrons in the system. Now we don't get any choice about how to occupy states since all are occupied. Said another way, the *multi-particle* Hilbert space contains just a single state: the fully filled band. This time, if we perturb with an electric field then the electrons can't move anywhere, simply because there's no where for them to go: they are locked in place by the Pauli principle. This means that, despite the presence of the electric field, there is no electric current. This is what we call an *insulator*. (It is sometimes said to be a *band* insulator to distinguish it from other mechanisms that also lead to insulating behaviour.)

The difference between a conductor and an insulator is one of the most striking characterisations of materials, one that we all learn in high school. The rough sketch above is telling us that this distinction arises due to quantum phenomena: the formation of energy bands and the Pauli exclusion principle. We'll explore this more in Section 4.1.

3.1.2 Nearly Free Electrons

The tight-binding model is an extreme cartoon of the real physics in which space is discrete; electrons are stuck on atomic sites with a non-vanishing probability to hop to a neighbouring site. In this section we present another cartoon that is designed to capture the opposite extreme.

We will assume that our electron is free to move anywhere along the line, parameterised by the position x. To mimic the underlying lattice, we add a weak, periodic potential $V(x)$. This means that we consider the Hamiltonian

$$H = \frac{p^2}{2m} + V(x)$$

where $p = -i\hbar d/dx$ is the usual momentum operator. The periodicity of the potential means that it satisfies

$$V(x + a) = V(x) \tag{3.7}$$

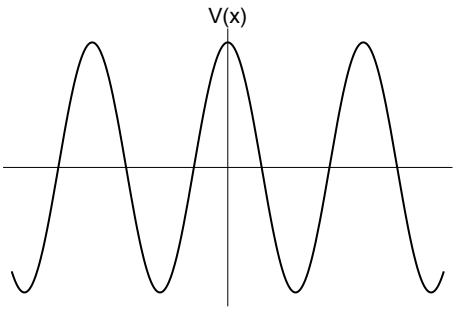


Figure 26: A periodic sine wave.

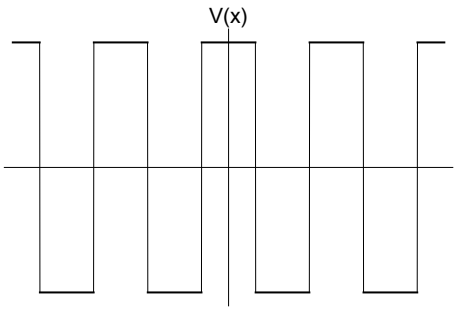


Figure 27: A periodic square wave.

For example, the potential could take the form of a sine wave, or a square wave as shown in the figure, or it could be a an infinite series of delta functions. For much of our discussion we won't need the exact form of the potential.

To avoid discussing edge effects, it's again useful to consider the particle moving on a circle $\mathbf{S}^1$ of length (circumference) L. This is compatible with the periodicity requirement (3.7) only if $L/a = N \in \mathbf{Z}$. The integer N plays the role of the number of atoms in the lattice.

In the absence of the potential, the eigenstates are the familiar plane waves $|k\rangle$, labelled by the momentum $p = \hbar k$. Because we are on a circle, the wavenumber of k is quantised in units of $2\pi/L$. The associated wavefunctions are

$$\psi_k(x) = \langle x|k\rangle = \frac{1}{\sqrt{L}} e^{ikx} \tag{3.8}$$

These states are are orthonormal, with

$$\langle k|k'\rangle = \frac{1}{L}\int dx\ e^{i(k'-k)x} = \delta_{k,k'} \tag{3.9}$$

(Recall that we are living on a circle, so the momenta k are discrete and the Kronecker delta is the appropriate thing to put on the right-hand side.) Meanwhile, the energy of a free particle is given by

$$E_0(k) = \frac{\hbar^2 k^2}{2m} \tag{3.10}$$

Our goal is to understand how the presence of the potential $V(x)$ affects this energy spectrum. To do this, we work perturbatively. However, perturbation theory in the present situation is a little more subtle than usual. Let's see why.

Perturbation Theory

Recall that the first thing we usually do in perturbation theory is decide whether we have non-degenerate or degenerate energy eigenstates. Which do we have in the present case? Well, all states are trivially degenerate because the energy of a free particle moving to the right is the same as the energy of a free particle moving to the left: $E_0(k) = E_0(-k)$. But the fact that the two states $|k\rangle$ and $|-k\rangle$ have the same energy does not necessarily mean that we have to use degenerate perturbation theory. This is only true if the perturbation causes the two states to mix.

To see what happens we will need to compute matrix elements $\langle k|V|k'\rangle$. The key bit of physics is the statement that the potential is periodic (3.7). This ensures that it can be Fourier expanded

$$V(x) = \sum_{n\in\mathbf{Z}} V_n\, e^{2\pi i n x/a} \qquad \text{with} \quad V_n = V^\star_{-n}$$

where the Fourier coefficients follow from the inverse transformation

$$V_n = \frac{1}{a}\int_0^a dx\; V(x)\, e^{-2\pi i n x/a}$$

The matrix elements are then given by

$$\langle k|V|k'\rangle = \frac{1}{L}\int dx\; \sum_{n\in\mathbf{Z}} V_n\, e^{i(k'-k+2\pi n/a)x} = \sum_{n\in\mathbf{Z}} V_n\, \delta_{k-k',2\pi n/a} \tag{3.11}$$

We see that we get mixing only when

$$k = k' + \frac{2\pi n}{a}$$

for some integer n. In particular, we get mixing between degenerate states $|k\rangle$ and $|-k\rangle$ only when

$$k = \frac{\pi n}{a}$$

for some n. The first time that this happens is when $k = \pi/a$. But we've seen this value of momentum before: it is the edge of the Brillouin zone (3.4). This is the first hint that the tight-binding model and nearly free electron model share some common features.

With this background, let's now try to sketch the basic features of the energy spectrum as a function of k.

Low Momentum: With low momentum $|k| \ll \pi/a$, there is no mixing between states at leading order in perturbation theory (and very little mixing at higher order). In this regime we can use our standard results from non-degenerate perturbation theory. Expanding the energy to second order, we have

$$E(k) = \frac{\hbar^2 k^2}{2m} + \langle k|V|k\rangle + \sum_{k' \neq k} \frac{|\langle k|V|k'\rangle|^2}{E_0(k) - E_0(k')} + \dots \tag{3.12}$$

From (3.11), we know that the first order correction is $\langle k|V|k\rangle = V_0$, and so just gives a constant shift to the energy, independent of k. Meanwhile, the second order term only gets contributions from $|k'\rangle = |k + 2\pi n/a\rangle$ for some n. When $|k| \ll \pi/a$, these corrections are small. We learn that, for small momenta, the particle moves as if unaffected by the potential. Intuitively, the de Broglie wavelength $2\pi/k$ of the particle much greater than the wavelength a of the potential, and the particle just glides over it unimpeded.

The formula (3.12) holds for low momenta. It also holds for momenta $\pi n/a \ll k \ll \pi(n+1)/a$ which are far from the special points where mixing occurs. However, the formula knows about its own failings because if we attempt to use it when $k = n\pi/a$ for some n, the the numerator $\langle k|V|-k\rangle$ is finite while the denominator becomes zero. Whenever perturbation theory diverges in this manner it's because we're doing something wrong. In this case it's because we should be working with degenerate perturbation theory.

At the Edge of the Brillouin Zone: Let's consider the momentum eigenstates which sit right at the edge of the Brillouin zone, $k = \pi/a$, or at integer multiples

$$k = \frac{n\pi}{a}$$

As we've seen, these are the values which mix due to the potential perturbation and we must work with degenerate perturbation theory.

Let's recall the basics of degenerate perturbation theory. We focus on the subsector of the Hilbert space formed by the two degenerate states, in our case $|k\rangle$ and $|k'\rangle = |-k\rangle$. To leading order in perturbation theory, the new energy eigenstates will be some linear combination of these original states $\alpha|k\rangle + \beta|k'\rangle$. We would like to figure out what choice of α and β will diagonalise the new Hamiltonian. There will be two such choices since there must, at the end of the day, remain two energy eigenstates. To determine the correct choice of these coefficients, we write the Schrödinger equation, restricted to

this subsector, in matrix form

$$\begin{pmatrix} \langle k|H|k\rangle & \langle k|H|k'\rangle \\ \langle k'|H|k\rangle & \langle k'|H|k'\rangle \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = E \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{3.13}$$

We've computed the individual matrix elements above: using the fact that the states $|k\rangle$ are orthonormal (3.9), the unperturbed energy (3.10) and the potential matrix elements (3.11), our eigenvalue equation becomes

$$\begin{pmatrix} E_0(k) + V_0 & V_n \\ V_n^\star & E_0(k') + V_0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = E \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{3.14}$$

where, for the value $k = -k' = n\pi/a$ of interest, $E_0(k) = E_0(k') = n^2\hbar^2\pi^2/2ma^2$. It's simple to determine the eigenvalues E of this matrix: they are given by the roots of the quadratic equation

$$(E_0(k) + V_0 - E)^2 - |V_n|^2 = 0 \quad \Rightarrow \quad E = \frac{\hbar^2}{2m}\frac{n^2\pi^2}{a^2} + V_0 \pm |V_n| \tag{3.15}$$

This is important. We see that a gap opens up in the spectrum at the values $k = \pm n\pi/a$. The size of the gap is proportional to $2|V_n|$.

It's simple to understand what's going on here. Consider the simple potential

$$V = 2V_1 \cos\left(\frac{2\pi x}{a}\right)$$

which gives rise to a gap only at $k = \pm\pi/a$. The eigenvectors of the matrix are $(\alpha, \beta) = (1, -1)$ and $(\alpha, \beta) = (1, 1)$, corresponding to the wavefunctions

$$\begin{aligned} \psi_+(x) &= \langle x|\Big(|k\rangle + |-k\rangle\Big) \sim \cos\left(\frac{\pi x}{a}\right) \\ \psi_-(x) &= \langle x|\Big(|k\rangle - |-k\rangle\Big) \sim \sin\left(\frac{\pi x}{a}\right) \end{aligned}$$

The density of electrons is proportional to $|\psi_\pm|^2$. Plotting these densities on top of the potential, we see that ψ_+ describes electrons that are gathered around the peaks of the potential, while ψ_- describes electrons gathered around the minima. It is no surprise that the energy of ψ_+ is higher than that of ψ_-.

Close to the Edge of the Brillouin Zone: Now consider an electron with

$$k = \frac{n\pi}{a} + \delta$$

for some small δ. As we've seen, the potential causes plane wave states to mix only if their wavenumbers differ by some multiple of $2\pi/a$. This means that $|k\rangle = |n\pi/a + \delta\rangle$ will mix with $|k'\rangle = |-n\pi/a + \delta\rangle$. These states don't quite have the same kinetic energy, but they have very *nearly* the same kinetic energy. And, as we will see, the perturbation due to the potential V will mean that these states still mix strongly.

To see this mixing, we need once again to solve the eigenvalue equation (3.13) or, equivalently, (3.14). The eigenvalues are given by solutions to the quadratic equation

$$\Big(E_0(k) + V_0 - E\Big)\Big(E_0(k') + V_0 - E\Big) - |V_n|^2 = 0 \tag{3.16}$$

The only difference from our previous discussion is that $E(k)$ and $E(k')$ are now given by

$$E(k) = \frac{\hbar^2}{2m}\left(\frac{n\pi}{a} + \delta\right)^2 \quad \text{and} \quad E(k') = \frac{\hbar^2}{2m}\left(\frac{n\pi}{a} - \delta\right)^2$$

and the quadratic equation (3.16) becomes

$$\left(\frac{\hbar^2}{2m}\left(\frac{n^2\pi^2}{a^2} + \delta^2\right) + V_0 - E\right)^2 - \left(\frac{\hbar^2}{2m}\frac{2n\pi\delta}{a}\right)^2 - |V_n|^2 = 0$$

This equation has two solutions, $E = E_\pm$, given by

$$E_\pm = \frac{\hbar^2}{2m}\left(\frac{n^2\pi^2}{a^2} + \delta^2\right) + V_0 \pm \sqrt{|V_n|^2 + \left(\frac{\hbar^2}{2m}\frac{2n\pi\delta}{a}\right)^2}$$

We're ultimately interested in this expression when δ is small, where we anticipate that the effect of mixing will be important. But, as a sanity check, let's first expand it in the opposite regime, when we're far from the edge of the Brillouin zone and δ is large compared to the gap V_n. In this case, a little bit of algebra shows that the eigenvalues can be written as

$$E_\pm = E_0(n\pi/a \pm \delta) + V_0 \pm \frac{|V_n|^2}{E_0(n\pi/a + \delta) - E_0(n\pi/a - \delta)}$$

But this coincides with the the expression that we got from second-order, non-degenerate perturbation theory (3.12). (Or, more precisely, because we have kept just a single mixing term in our discussion above we get just a single term in the sum in (3.12); for some choice of potentials, keeping further terms may be important.)

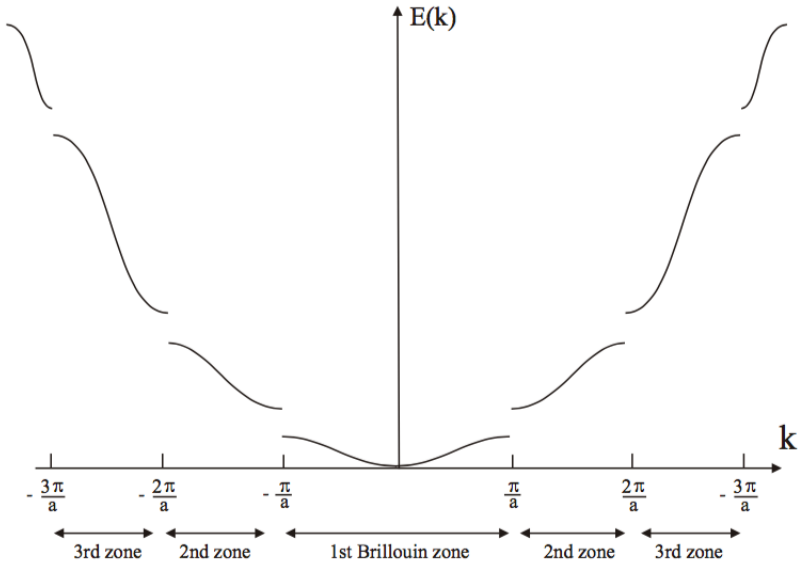


Figure 28: Energy dispersion for the free electron model.

Our real interest is what happens close to the edge of the Brillouin zone when δ is small compared to the gap V_n. In this case we can expand the square-root to give

$$E_\pm \approx \frac{\hbar^2}{2m}\frac{n^2\pi^2}{a^2} + V_0 \pm |V_n| + \frac{\hbar^2}{2m}\left(1 \pm \frac{1}{|V_n|}\frac{n^2\hbar^2\pi^2}{ma^2}\right)\delta^2$$

The first collection of terms coincide with the energy at the edge of the Brillouin zone (3.15), as indeed it must. For us, the important new point is in the second term which tells us that as we approach the gaps, the energy is quadratic in the momentum δ.

Band Structure

We now have all we need to sketch the rough form of the energy spectrum $E(k)$. The original quadratic spectrum is deformed with a number of striking features:

- For small momenta, $k \ll \pi/a$, the spectrum remains roughly unchanged.
- The energy spectrum splits into distinct bands, with gaps arising at $k = n\pi/a$ with $n \in \mathbf{Z}$. The size of these gaps is given by $2|V_n|$, where V_n is the appropriate Fourier mode of the potential.

The region of momentum space corresponding to the n^{th} energy band is called the n^{th} *Brillouin zone*. However, we usually call the 1^{st} Brillouin zone simply *the* Brillouin zone.

- As we approach the edge of a band, the spectrum is quadratic. In particular, $dE/dk \to 0$ at the end of a band.

The relationship $E(k)$ between energy and momentum is usually called the *dispersion relation*. In the present case, it is best summarised in a figure.

Note that the spectrum within the first Brillouin zone $|k| \leq \pi/a$, looks very similar to what we saw in the tight-binding model . The qualitative differences in the two models arise because the tight-binding model has a finite number of states, all contained in the first Brillouin zone, while the nearly-free electron model has an infinite number of states which continue for $|k| > \pi/a$.

3.1.3 The Floquet Matrix

One of the main lessons that we learned above is that there are gaps in the energy spectrum. It's hard to overstate the importance of these gaps. Indeed, as we saw briefly above, and will describe in more detail in 4.1.1, the gaps are responsible for some of the most prominent properties of materials, such as the distinction between conductors and insulators.

Because of the important role they play, we will here describe another way to see the emergence of gaps in the spectrum that does not rely on perturbation theory. Consider a general, periodic potential $V(x) = V(x + a)$. We are interested in solutions to the Schrödinger equation

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + V(x)\psi(x) = E\psi(x) \tag{3.17}$$

Since this is a second order differential equation, we know that there must be two solutions $\psi_1(x)$ and $\psi_2(x)$. However, because the potential is periodic, it must be the case that $\psi_1(x + a)$ and $\psi_2(x + a)$ are also solutions. These two sets of solutions are therefore related by some linear transformation

$$\begin{pmatrix} \psi_1(x+a) \\ \psi_2(x+a) \end{pmatrix} = F(E) \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \end{pmatrix} \tag{3.18}$$

where $F(E)$ is a 2×2 matrix which, as the notation suggests, depends on the energy of the solution E. It is known as the *Floquet matrix* and has a number of nice properties.

Claim: $\det(F) = 1$.

Proof: First some gymnastics. We differentiate (3.18) to get

$$\begin{pmatrix} \psi_1'(x+a) \\ \psi_2'(x+a) \end{pmatrix} = F(E) \begin{pmatrix} \psi_1'(x) \\ \psi_2'(x) \end{pmatrix}$$

We can combine this with our previous equation by introducing the 2×2 matrix

$$W(x) = \begin{pmatrix} \psi_1(x) & \psi_1'(x) \\ \psi_2(x) & \psi_2'(x) \end{pmatrix}$$

which obeys the matrix equation

$$W(x+a) = F(E)W(x) \tag{3.19}$$

Consider $\det W = \psi_1\psi_2' - \psi_1'\psi_2$. You might recognise this from the earlier course on *Differential Equations* as the *Wronskian*. It's simple to show, using the Schrödinger equation (3.17), that $(\det W)' = 0$. This means that $\det W$ is independent of x so, in particular, $\det W(x+a) = \det W(x)$. Taking the determinant of (3.19) then tells us that $\det F = 1$ as claimed. □

Claim: $\mathrm{Tr}\, F$ is real.

Proof: We always have the choice pick the original wavefunctions $\psi_1(x)$ and $\psi_2(x)$ to be entirely real for all x. (If they're not, simply take the real part and this is also a solution to the Schrodinger equation). With this choice, the Floquet matrix itself has real elements, and so its trace is obviously real. But the trace is independent of our choice of basis of wavefunctions. Any other choice is related by a transformation $F \to AFA^{-1}$, for some invertible matrix A and this leaves the trace invariant. Hence, even if the components of $F(E)$ are complex, its trace remains real. □

To understand the structure of solutions to (3.18), we look at the eigenvalues, λ_+ and λ_- of $F(E)$. Of course, these too depend on the energy E of the solutions. Because $\det F = 1$, they obey $\lambda_+\lambda_- = 1$. They obey the characteristic equation

$$\lambda^2 - (\mathrm{Tr}\, F(E))\lambda + 1 = 0$$

The kind of solution that we get depends on whether $(\mathrm{Tr}\, F(E))^2 < 4$ or $(\mathrm{Tr}\, F(E))^2 > 4$.

$\underline{(\mathrm{Tr}\, F(E))^2 < 4}$: In this case, the roots are complex and of equal magnitude. We can write

$$\lambda_+ = e^{ika} \quad \text{and} \quad \lambda_- = e^{-ika}$$

for some k which, assuming that the roots are distinct, lies in the range $|k| < \pi/a$. To see what this means for solutions to (3.18), we introduce the left-eigenvector of $(\alpha_\pm, \beta_\pm)F = \lambda_\pm(\alpha_\pm, \beta_\pm)$. Then the linear combinations $\psi_\pm = \alpha_\pm\psi_1 + \beta_\pm\psi_2$ obey

$$\psi_\pm(x+a) = e^{\pm ika}\psi_\pm(x)$$

These are extended states, spread (on average) equally throughout the lattice. They corresponds to the bands in the spectrum.

$\underline{(\mathrm{Tr}\, F(E))^2 > 4}$: Now the eigenvalues take the form

$$\lambda_1 = e^{\mu a} \quad \text{and} \quad \lambda_2 = e^{-\mu a}$$

for some μ. The corresponding eigenstates now obey

$$\psi_\pm(x+a) = e^{\pm\mu a}\psi_\pm(x)$$

States of this form are not allowed: they are unbounded either as $x \to +\infty$ or as $x \to -\infty$. These values of energy E are where the gaps occur in the spectrum.

We have to work a little harder when $F(E) = 4$ and the two eigenvalues are degenerate, either both $+1$ or both -1. This situations corresponds to the edge of the band. Consider the case when both eigenvalues are $+1$. Recall from your first course on *Vectors and Matrices* that attempting to diagonalise such a 2×2 matrix can result in two different canonical forms

$$PF(E)P^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{or} \quad PF(E)P^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

In the former case, there are two allowed solutions. In the latter case, you can check that one solution is allowed, while the other grows linearly in x.

3.1.4 Bloch's Theorem in One Dimension

In both models described above, we ended up labelling states by momentum $\hbar k$. It's worth pausing to ask: why did we do this? And how should we think of k?

Before we get to this, let's back up and ask an even more basic question: why do we label the states of a free particle by momentum? Here, the answer is because momentum is conserved. In the quantum theory, this means that the momentum operator commutes with the Hamiltonian: $[p, H] = 0$, so that we can simultaneously label states by both energy and momentum. Ultimately, Noether's theorem tells us that this conservation law arises because of translational invariance of the system.

Now let's look at our system with a lattice. We no longer have translational invariance. Correspondingly, in the nearly-free electron model, $[p, H] \neq 0$. Hopefully this now makes our original question sharper: why do we get to label states by k?!

While we don't have full, continuous translational invariance, both the models that we discussed do have a discrete version of translational invariance

$$x \to x + a$$

As we now show, this is sufficient to ensure that we can label states by something very similar to "momentum". However, the values of this momentum are restricted. This result is known as *Bloch's Theorem*. Here we prove the theorem for our one-dimensional system; we will revisit it in Section 3.3.1 in higher dimensions.

The Translation Operator

For concreteness, let's work with continuous space where states are described by a wavefunction $\psi(x)$. (There is a simple generalisation to discrete situations such as the tight-binding model that we describe below.) We introduce the translation operator T_l as

$$T_l\,\psi(x) = \psi(x + l)$$

First note that T_l is a unitary operator. To see this, we just need to look at the overlap

$$\begin{aligned}\langle\phi|T_l|\psi\rangle = \int dx\ \phi(x)^\star\, T_l\psi(x) &= \int dx\ \phi(x)^\star\psi(x+l) \\ &= \int dx\ \phi(x-l)^\star\psi(x) = \int dx\ [T_{-l}\,\phi(x)]^\star\psi(x)\end{aligned}$$

where, in the step to the second line, we've simply shifted the origin. This tells us that $T_l^\dagger = T_{-l}$. But clearly $T_l^{-1} = T_{-l}$ as well, so $T_l^\dagger = T_l^{-1}$ and the translation operator is unitary as claimed.

Next note that the set of translation operators form an Abelian group,

$$T_{l_1} T_{l_2} = T_{l_1+l_2} \tag{3.20}$$

with $[T_{l_1}, T_{l_2}] = 0$.

The translation operator is a close cousin of the familiar momentum operator

$$p = -i\hbar \frac{d}{dx}$$

The relationship between the two is as follows: the unitary translation operator is the exponentiation of the Hermitian momentum operator

$$T_l = e^{ilp/\hbar}$$

To see this, we expand the exponent and observe that $T_l\psi(x) = \psi(x+l)$ is just a compact way of expressing the Taylor expansion of a function

$$\begin{aligned} T_l\psi(x) &= \left(1 + \frac{ilp}{\hbar} + \frac{1}{2}\left(\frac{ilp}{\hbar}\right)^2 + \ldots\right)\psi(x) \\ &= \left(1 + l\frac{d}{dx} + \frac{l^2}{2}\frac{d^2}{dx^2} + \ldots\right)\psi(x) = \psi(x+l) \end{aligned}$$

We say that the momentum operator is the "generator" of infinitesimal translations.

A quantum system is said to be invariant under translations by l if

$$[H, T_l] = 0 \tag{3.21}$$

Phrased in this way, we can describe both continuous translational symmetry and discrete translational symmetry. A system has continuous translational invariance if (3.21) holds for all l. In this case, we may equivalently say that $[p, H] = 0$. Alternatively, a system may have discrete translational invariance if (3.21) holds only when l is an integer multiple of the lattice spacing a. Now p does not commute with H.

Let's look at the case of discrete symmetry. Now we can't simultaneously diagonalise p and H, but we can simultaneously diagonalise T_a and H. In other words, energy eigenstates can be labelled by the eigenvalues of T_a. But T_a is a unitary operator and its eigenvalues are simply a phase, $e^{i\theta}$ for some θ. Moreover, we want the eigenvalues to respect the group structure (3.20). This is achieved if we write the eigenvalue of T_l

as $e^{i\theta} = e^{ikl}$ for some k, so that the eigenvalue of T_{na} coincides with the eigenvalue of T_a^n. The upshot is that eigenstates are labelled by some k, such that

$$T_a\psi_k(x) = \psi_k(x+a) = e^{ika}\psi_k(x)$$

Now comes the rub. Because the eigenvalue is a phase, there is an arbitrariness in this labelling: states labelled by k have the same eigenvalue under T_a as states labelled by $k + 2\pi/a$. To remedy this, we will simply require that k lies in the range

$$k \in \left[-\frac{\pi}{a}, \frac{\pi}{a}\right) \tag{3.22}$$

We recognise this as the first Brillouin zone.

This, then, is the essence of physics on a lattice. We can still label states by k, but it now lies in a finite range. Note that we can approximate a system with continuous translational symmetry by taking a arbitrarily small; in this limit we get the usual result $k \in \mathbf{R}$.

This discussion leads us directly to:

Bloch's Theorem in One Dimension: In a periodic potential, $V(x) = V(x+a)$, there exists a basis of energy eigenstates that can be written as

$$\psi_k(x) = e^{ikx}u_k(x)$$

where $u_k(x) = u_k(x+a)$ is a periodic function and k lies in the Brillouin zone (3.22).

Proof: We take ψ_k to be an eigenstate of the translation operator T_a, so that $\psi_k(x+a) = e^{ika}\psi_k(x)$. Then $u_k(x+a) = e^{-ik(x+a)}\psi_k(x+a) = e^{-ikx}\psi_k(x) = u_k(x)$. □

Bloch's theorem is rather surprising. One might think that the presence of a periodic potential would dramatically alter the energy eigenstates, perhaps localising them in some region of space. Bloch's theorem is telling us that this doesn't happen: instead the plane wave states e^{ikx} are altered only by a periodic function $u(x)$, sometimes referred to as a *Bloch function*, and the fact that the wavenumber is restricted to the first Brillouin zone.

Finally, note that we've couched the above discussion in terms of wavefunctions $\psi(x)$, but everything works equally well for the tight-binding model with the translation operator defined by $T_a|n\rangle = |n+1\rangle$.

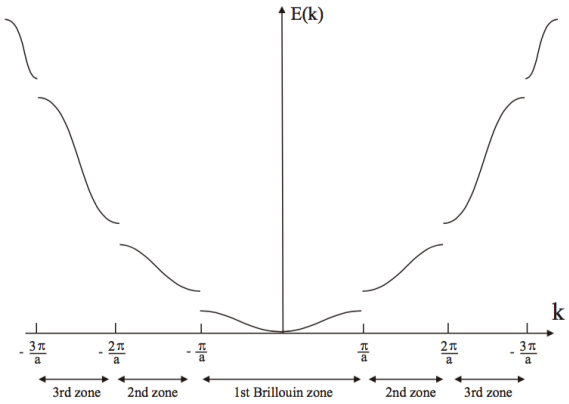


Figure 29: The extended zone scheme.

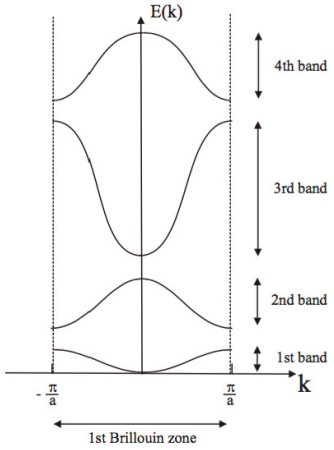


Figure 30: The reduced zone scheme.

Crystal Momentum

The quantity $p = \hbar k$ is the quantity that replaces momentum in the presence of a lattice. It is called the *crystal momentum*. Note, however, that it doesn't have the simple interpretation of "mass $\times$ velocity". (We will describe how to compute the velocity of a particle in terms of the crystal momentum in Section 4.2.1.)

Crystal momentum is conserved. This becomes particularly important when we consider multiple particles moving in a lattice and their interactions. This, of course, sounds the same as the usual story of momentum. Except there's a twist: crystal momentum is conserved only mod $2\pi/a$. It is perfectly possible for two particles to collide in a lattice environment and their final crystal momentum to differ from their initial crystal momentum by some multiple of $2\pi/a$. Roughly speaking, the lattice absorbs the excess momentum.

This motivates us to re-think how we draw the energy spectrum. Those parts of the spectrum that lie outside the first Brillouin zone should really be viewed as having the same crystal momentum. To show this, we draw the energy spectrum as a multi-valued function of $k \in [-\pi/a, \pi/a)$. The spectrum that we previously saw in Figure 28 then looks like

The original way of drawing the spectrum is known as the *extended zone scheme*. The new way is known as the *reduced zone scheme*. Both have their uses. Note that edges of the Brillouin zone are identified: $k = \pi/a$ is the same as $k = -\pi/a$. In other words, the Brillouin zone is topologically a circle.

In the reduced zone scheme, states are labelled by both $k \in [-\pi/a, \pi/a)$ and an integer $n = 1, 2, \ldots$ which tells us which band we are talking about.

3.2 Lattices

The ideas that we described above all go over to higher dimensions. The key difference is that lattices in higher dimensions are somewhat more complicated than a row of points. In this section, we introduce the terminology needed to describe different kinds of lattices. In Section 3.3, we'll return to look at what happens to electrons moving in these lattice environments.

3.2.1 Bravais Lattices

The simplest kind of lattice is called a *Bravais lattice*. This is a periodic array of points defined by integer sums of linearly independent basis vectors $\mathbf{a}_i$. In two-dimensions, a Bravais lattice Λ is defined by

$$\Lambda = \{\mathbf{r} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 \quad , \quad n_i \in \mathbf{Z}\}$$

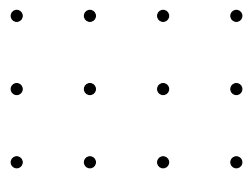


Figure 31:

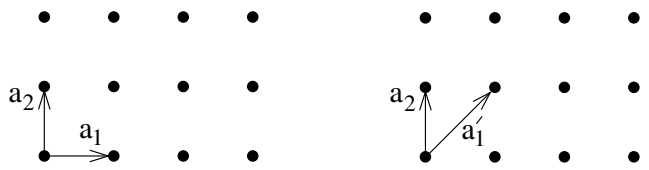
An obvious example is the square lattice shown to the right. We will see further examples shortly.

In three dimensions, a Bravais lattice is defined by

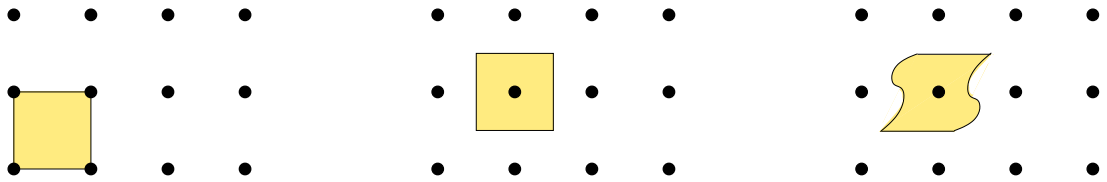
$$\Lambda = \{\mathbf{r} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \quad , \quad n_i \in \mathbf{Z}\}$$

These lattices have the property that any point looks just the same as any other point. In mathematics, such an object would simply be called a lattice. Here we add the word *Bravais* to distinguish these from more general kinds of lattices that we will meet shortly.

The basis vectors $\mathbf{a}_i$ are called *primitive lattice vectors*. They are not unique. As an example, look at the 2-dimensional square lattice below. We could choose basis vectors ($\mathbf{a}_1$,$\mathbf{a}_2$) or ($\mathbf{a}'_1$,$\mathbf{a}_2$). Both will do the job.



A *primitive unit cell* is a region of space which, when translated by the primitive lattice vectors $\mathbf{a}_i$, tessellates the space. This means that the cells fit together, without overlapping and without leaving any gaps. These primitive unit cells are not unique. As an example, let's look again at the 2-dimensional square lattice. Each of the three possibilities shown below is a good unit cell.



Each primitive unit cell contains a single lattice point. This is obvious in the second and third examples above. In the first example, there are four lattice points associated to the corners of the primitive unit cell, but each is shared by four other cells. Counting these as a 1/4 each, we see that there is again just a single lattice point in the primitive unit cell.

Although the primitive unit cells are not unique, each has the same volume. It is given by

$$V = |\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)| \tag{3.23}$$

Because each primitive unit cell is associated to a single lattice point, $V = 1/n$ where n is the density of lattice points.

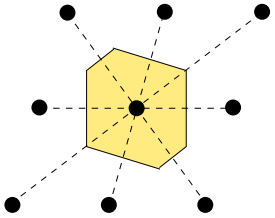
Note finally that the primitive unit cell need not have the full symmetry of the lattice. For example, the third possible unit cell shown above for the square lattice is not invariant under 90° rotations.

For any lattice, there is a canonical choice of primitive unit cell that does inherit the symmetry of the underlying lattice. This is called the *Wigner-Seitz* cell, Γ. (It sometimes goes by the name of the *Voronoi cell*.) Pick a lattice point which we choose to be at the origin. The Wigner-Seitz cell is defined is defined to be the region of space around such that the origin is the closest lattice point. In equations,

$$\Gamma = \{\mathbf{x} : |\mathbf{x}| < |\mathbf{x} - \mathbf{r}| \quad \forall\ \mathbf{r} \in \Lambda \text{ s.t. } \mathbf{r} \neq 0\ \}$$

The Wigner-Seitz cells for square and triangular lattices are given by

There is a simple way to construct the Wigner-Seitz cell. Draw lines from the origin to all other lattice points. For each of these lines, construct the perpendicular bi-sectors; these are lines in 2d and planes in 3d. The Wigner-Seitz cell is the inner area bounded by these bi-sectors. Here's another example.



Examples of Bravais Lattices in 2d

Let's look at some examples. In two dimensions, a Bravais lattice is defined by two non-parallel vectors $\mathbf{a}_1$ and $\mathbf{a}_2$, with angle $\theta \neq 0$ between them. However, some of these lattices are more special than others. For example, when $|\mathbf{a}_1| = |\mathbf{a}_2|$ and $\theta = \pi/2$, the lattice is square and enjoys an extra rotational symmetry.

We will consider two Bravais lattices to be equivalent if they share the same symmetry group. With this definition, there are five possible Bravais lattices in two dimensions. They are

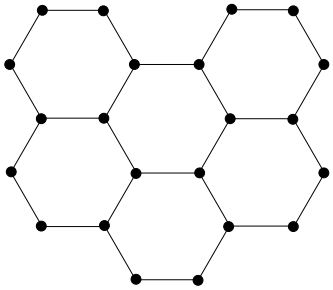
- **Square:** $|\mathbf{a}_1| = |\mathbf{a}_2|$ and $\theta = \pi/2$. It has four-fold rotation symmetry and reflection symmetry.
- **Triangular:** $|\mathbf{a}_1| = |\mathbf{a}_2|$ and $\theta = \pi/3$ or $\theta = 2\pi/3$. This is also sometimes called a hexagonal lattice. It has six-fold rotation symmetry.
- **Rectangular:** $|\mathbf{a}_1| \neq |\mathbf{a}_2|$ and $\theta = \pi/2$. This has reflection symmetry.
- **Centred Rectangular:** $|\mathbf{a}_1| \neq |\mathbf{a}_2|$ and $\theta \neq \pi/2$, but the primitive basis vectors should obey $(2\mathbf{a}_2 - \mathbf{a}_1) \cdot \mathbf{a}_1 = 0$. This means that the lattice looks like a rectangle with an extra point in the middle.

- **Oblique:** $|\mathbf{a}_1| \neq |\mathbf{a}_2|$ and nothing special. This contains all other cases.

The square, triangular and oblique lattices were shown on the previous page where we also drew their Wigner-Seitz cells.

Not all Lattices are Bravais

Not all lattices of interest are Bravais lattices. One particularly important lattice in two dimensions has the shape of a honeycomb and is shown below.



This lattice describes a material called *graphene* that we will describe in more detail in Section 4.1.3. The lattice is not Bravais because not all points are the same. To see this, consider a single hexagon from the lattice as drawn below.

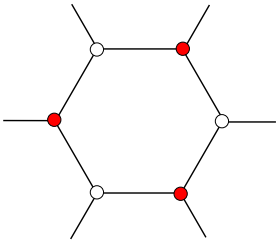


Figure 32:

Each of the red points is the same: each has a neighbour directly to the left of them, and two neighbours diagonally to the right. But the white points are different. Each of them has a neighbour directly to the right, and two neighbours diagonally to the left.

Lattices like this are best thought by decomposing them into groups of atoms, where some element of each group sits on the vertices of a Bravais lattice. For the honeycomb lattice, we can consider the group of atoms ○——●. The red vertices form a triangular lattice, with primitive lattice vectors

$$\mathbf{a}_1 = \frac{\sqrt{3}a}{2}(\sqrt{3}, 1) \quad , \quad \mathbf{a}_2 = \frac{\sqrt{3}a}{2}(\sqrt{3}, -1)$$

Meanwhile, each red vertex is accompanied by a white vertex which is displaced by

$$\mathbf{d} = (-a, 0)$$

This way we build our honeycomb lattice.

This kind of construction generalises. We can describe any lattice as a repeating group of atoms, where each group sits on an underlying Bravais lattice Λ. Each atom in the group is displaced from the vertex of the Bravais lattice by a vector $\mathbf{d}_i$. Each group of atoms, labelled by their positions $\mathbf{d}_i$ is called the *basis*. For example, for the honeycomb lattice we chose the basis $\mathbf{d}_1 = 0$ for red atoms and $\mathbf{d}_2 = \mathbf{d}$ for white atoms, since the red atoms sat at the positions of the underlying triangular lattice. In general there's no requirement that any atom sits on the vertex of the underlying Bravais lattice. The whole lattice is then described by the union of the Bravais lattice and the basis, $\cup_i\{\Lambda + \mathbf{d}_i\}$.

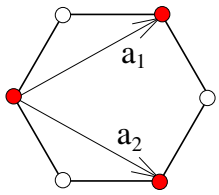


Figure 33:

Examples of Bravais Lattices in 3d

It turns out that there are 14 different Bravais lattices in three dimensions. Fortunately we won't need all of them. In fact, we will describe only the three that arise most frequently in Nature. These are:

- **Cubic:** This is the simplest lattice. The primitive lattice vectors are aligned with the Euclidean axes

$$\mathbf{a}_1 = a\hat{\mathbf{x}} \quad , \quad \mathbf{a}_2 = a\hat{\mathbf{y}} \quad , \quad \mathbf{a}_3 = a\hat{\mathbf{z}}$$

  And the primitive cell has volume $V = a^3$. The Wigner-Seitz cell is also a cube, centered around one of the lattice points.

- **Body Centered Cubic (BCC):** This is a cubic lattice, with an extra point placed at the centre of each cube. We could take the primitive lattice vectors to be

$$\mathbf{a}_1 = a\hat{\mathbf{x}} \quad , \quad \mathbf{a}_2 = a\hat{\mathbf{y}} \quad , \quad \mathbf{a}_3 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$$

  However, a more symmetric choice is

$$\mathbf{a}_1 = \frac{a}{2}(-\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}) \quad , \quad \mathbf{a}_2 = \frac{a}{2}(\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}}) \quad , \quad \mathbf{a}_3 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}})$$

  The primitive unit cell has volume $V = a^3/2$.

  The BCC lattice can also be thought of as a cubic lattice, with a basis of two atoms with $\mathbf{d}_1 = 0$ and $\mathbf{d}_2 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$. However, this doesn't affect the fact that the BCC lattice is itself Bravais.

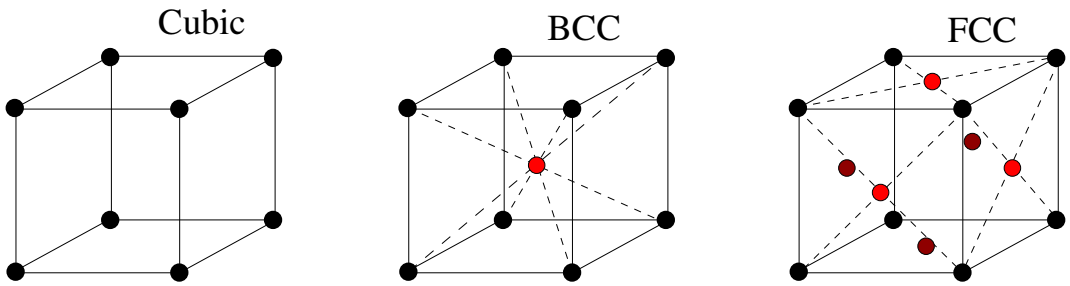


Figure 34: Three Bravais lattices. The different coloured atoms are there in an attempt to make the diagrams less confusing; they do not denote different types of atoms.

The Alkali metals (Li, Na, K, Rb, Cs) all have a BCC structure, as do the Vanadium group (V, Nb, Ta) and Chromium group (Cr, Mo, W) and Iron (Fe). In each case, the lattice constant is roughly $a \approx 3$ to $6 \times 10^{-10}\ m$.

- **Face Centered Cubic (FCC):** This is again built from the cubic lattice, now with an extra point added to the centre of each face. The primitive lattice vectors are

$$\mathbf{a}_1 = \frac{a}{2}(\hat{\mathbf{y}} + \hat{\mathbf{z}}) \quad , \quad \mathbf{a}_2 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{z}}) \quad , \quad \mathbf{a}_3 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}})$$

  The primitive unit cell has volume $V = a^3/4$.

  The FCC lattice can also be thought of as a cubic lattice, now with a basis of four atoms sitting at $\mathbf{d}_1 = 0$, $\mathbf{d}_2 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}})$, $\mathbf{d}_3 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{z}})$ and $\mathbf{d}_4 = \frac{a}{2}(\hat{\mathbf{y}} + \hat{\mathbf{z}})$. Nonetheless, it is also a Bravais lattice in its own right.

  Examples of FCC structures include several of the Alkaline earth metals (Be, Ca, Sr), many of the transition metals (Sc, Ni, Pd, Pt, Rh, Ir, Cu, Ag, Au) and the Noble gases (Ne, Ar, Kr, Xe) when in solid form, again with $a \approx 3$ to $6 \times 10^{-10}\ m$ in each case.

The Wigner-Seitz cells for the BCC and FCC lattices are polyhedra, sitting inside a cube. For example, the Wigner-Seitz cell for the BCC lattice is shown in the left-hand figure.

Examples of non-Bravais Lattices in 3d

As in the 2d examples above, we can describe non-Bravais crystals in terms of a basis of atoms sitting on an underlying Bravais lattice. Here are two particularly simple examples.

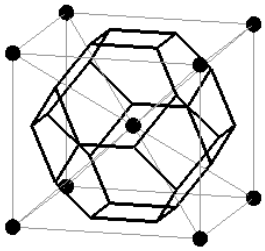


Figure 35: Wigner-Seitz cell for BCC

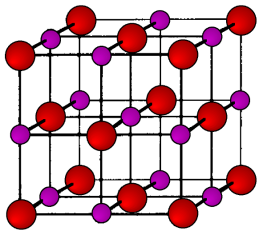


Figure 36: Salt.

Diamond is made up of two, interlaced FCC lattices, with carbon atoms sitting at the basis points $\mathbf{d}_1 = 0$ and $\mathbf{d}_2 = \frac{a}{4}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$. Silicon and germanium also adopt this structure.

Another example is salt ($NaCl$). Here, the basic structure is a cubic lattice, but with Na and Cl atoms sitting at alternate sites. It's best to think of this as two, interlaced FCC lattices, but shifted differently from diamond. The basis consists of a Na atom at $\mathbf{d} = 0$ and a Cl atom at $\mathbf{d}_2 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$. This basis then sits on top of an FCC lattice.

3.2.2 The Reciprical Lattice

Given a Bravais lattice Λ, defined by primitive vectors $\mathbf{a}_i$, the *reciprocal lattice* $\Lambda^\star$ is defined by the set of points

$$\Lambda^\star = \{\mathbf{k} = \sum_i n_i \mathbf{b}_i \ , \ n_i \in \mathbf{Z}\}$$

where the new primitive vectors $\mathbf{b}_i$ obey

$$\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\, \delta_{ij} \tag{3.24}$$

$\Lambda^\star$ is sometimes referred to as the *dual lattice*. In three dimensions, we can simply construct the lattice vectors $\mathbf{b}_i$ by

$$\mathbf{b}_i = \frac{2\pi}{V}\frac{1}{2}\epsilon_{ijk}\, \mathbf{a}_j \times \mathbf{a}_k$$

where V is the volume of unit cell of Λ (3.23). We can also invert this relation to get

$$\mathbf{a}_i = \frac{2\pi}{V^\star}\frac{1}{2}\epsilon_{ijk}\, \mathbf{b}_j \times \mathbf{b}_k$$

where $V^\star = |\mathbf{b}_1 \cdot (\mathbf{b}_2 \times \mathbf{b}_3)| = (2\pi)^3/V$ is the volume of $\Gamma^\star$, the unit cell of $\Lambda^\star$. Note that this shows that the reciprocal of the reciprocal lattice gives you back the original.

The condition (3.24) can also be stated as the requirement that

$$e^{i\mathbf{k}\cdot\mathbf{r}} = 1 \quad \forall \;\; \mathbf{r} \in \Lambda \;,\; \mathbf{k} \in \Lambda^\star \tag{3.25}$$

which provides an alternative definition of the reciprocal lattice.

Here are some examples:

- The cubic lattice has $\mathbf{a}_1 = a\hat{\mathbf{x}}$, $\mathbf{a}_2 = a\hat{\mathbf{y}}$ and $\mathbf{a}_3 = a\hat{\mathbf{z}}$. The reciprocal lattice is also cubic, with primitive vectors $\mathbf{b}_1 = (2\pi/a)\hat{\mathbf{x}}$, $\mathbf{b}_2 = (2\pi/a)\hat{\mathbf{y}}$ and $\mathbf{b}_3 = (2\pi/a)\hat{\mathbf{z}}$
- The BCC lattice has $\mathbf{a}_1 = \frac{a}{2}(-\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}})$, $\mathbf{a}_2 = \frac{a}{2}(\hat{\mathbf{x}}-\hat{\mathbf{y}}+\hat{\mathbf{z}})$ and $\mathbf{a}_3 = \frac{a}{2}(\hat{\mathbf{x}}+\hat{\mathbf{y}}-\hat{\mathbf{z}})$. The reciprocal lattice vectors are $\mathbf{b}_1 = (2\pi/a)(\hat{\mathbf{y}}+\hat{\mathbf{z}})$, $\mathbf{b}_2 = (2\pi/a)(\hat{\mathbf{x}}+\hat{\mathbf{z}})$ and $\mathbf{b}_3 = (2\pi/a)(\hat{\mathbf{x}}+\hat{\mathbf{y}})$. But we've seen these before: they are the lattice vectors for a FCC lattice with the sides of the cubic cell of length $4\pi/a$.

  We see that the reciprocal of a BCC lattice is an FCC lattice and vice versa.

The Reciprocal Lattice and Fourier Transforms

The reciprocal lattice should not be thought of as sitting in the same space as the original. This follows on dimensional grounds. The original lattice vectors $\mathbf{a}_i$ have the dimension of length, $[\mathbf{a}_i] = L$. The definition (3.24) then requires the dual lattice vectors $\mathbf{b}_i$ to have dimension $[\mathbf{b}_i] = 1/L$. The reciprocal lattice should be thought of as living in Fourier space which, in physics language, is the same thing as momentum space. As we'll now see, the reciprocal lattice plays an important role in the Fourier transform.

Consider a function $f(\mathbf{x})$ where, for definiteness, we'll take $\mathbf{x} \in \mathbf{R}^3$. Suppose that this function has the periodicity of the lattice Λ, which means that $f(\mathbf{x}) = f(\mathbf{x}+\mathbf{r})$ for all $\mathbf{r} \in \Lambda$. The Fourier transform is

$$\begin{aligned}\tilde{f}(\mathbf{k}) = \int d^3x \; e^{-i\mathbf{k}\cdot\mathbf{x}} \, f(\mathbf{x}) &= \sum_{\mathbf{r}\in\Lambda} \int_\Gamma d^3x \; e^{-i\mathbf{k}\cdot(\mathbf{x}+\mathbf{r})} f(\mathbf{x}+\mathbf{r}) \\ &= \sum_{\mathbf{r}\in\Lambda} e^{-i\mathbf{k}\cdot\mathbf{r}} \int_\Gamma d^3x \; e^{-i\mathbf{k}\cdot\mathbf{x}} f(\mathbf{x})\end{aligned} \tag{3.26}$$

In the second equality, we have replaced the integral over $\mathbf{R}^3$ with a sum over lattice points, together with an integral over the Wigner-Seitz cell Γ. In going to the second line, we have used the periodicity of $f(\mathbf{x})$. We see that the Fourier transform comes with the overall factor

$$\Delta(\mathbf{k}) = \sum_{\mathbf{r}\in\Lambda} e^{-i\mathbf{k}\cdot\mathbf{r}} \tag{3.27}$$

This is an interesting quantity. It has the following property:

Claim: $\Delta(\mathbf{k}) = \mathbf{0}$ unless $\mathbf{k} \in \Lambda^\star$.

Proof: Since we're summing over all lattice sites, we could equally well write $\Delta(\mathbf{k}) = \sum_{\mathbf{r}\in\mathbf{\Lambda}} \mathbf{e}^{-\mathbf{ik}\cdot(\mathbf{r}-\mathbf{r_0})}$ for any $\mathbf{r}_0 \in \Lambda$. This tells us that $\Delta(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{r}_0}\,\Delta(\mathbf{k})$ for any $\mathbf{r}_0 \in \Lambda$. This means that $\Delta(\mathbf{k}) = 0$ unless $e^{i\mathbf{k}\cdot\mathbf{r}_0} = 1$ for all $\mathbf{r}_0 \in \Lambda$. But this is equivalent to saying that $\Delta(\mathbf{k}) = 0$ unless $\mathbf{k} \in \Lambda^\star$. □

In fact, we can get a better handle on the function (strictly, a distribution) $\Delta(\mathbf{k})$. We have

Claim: $\Delta(\mathbf{k}) = V^\star \sum_{\mathbf{q}\in\Lambda^\star} \delta(\mathbf{k}-\mathbf{q})$.

Proof: We can expand $\mathbf{k} = \sum_i k_i\mathbf{b}_i$, with $k_i \in \mathbf{R}$, and $\mathbf{r} = \sum_i n_i\mathbf{a}_i$ with $n_i \in \mathbf{Z}$. Then, using (3.24), we have

$$\Delta(\mathbf{k}) = \sigma(k_1)\sigma(k_2)\sigma(k_3) \quad \text{where} \quad \sigma(k) = \sum_{n=-\infty}^{\infty} e^{-2\pi ikn}$$

The range of the sum in $\sigma(k)$ is appropriate for an infinite lattice. If, instead, we had a finite lattice with, say, $N+1$ points in each direction, (assume, for convenience, that N is even), we would replace $\sigma(k)$ with

$$\sigma_N(k) = \sum_{n=-N/2}^{N/2} e^{-2\pi ikn} = \frac{e^{-2\pi ik(N/2+1)} - e^{2\pi ikN/2}}{e^{-2\pi ik} - 1} = \frac{\sin(N+1)\pi k}{\sin \pi k}$$

This function is plotted on the right for $-1/2 < k < 1/2$. We have chosen a measly $N = 10$ in this plot, but already we see that the function is heavily peaked near the origin: when $k \sim \mathcal{O}(1/N)$, then $\sigma_N(k) \sim \mathcal{O}(N)$. As $N \to \infty$, this peak becomes narrower and taller and the area under it tends towards 1. To see this last point, replace $\sin(\pi k) \approx \pi k$ and use the fact that $\int_{-\infty}^{+\infty} \sin(x)/x = \pi$. This shows that the peak near the origin tends towards a delta function.

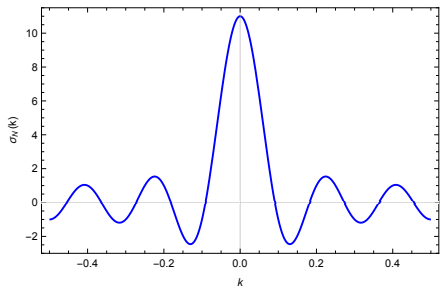


Figure 37:

The function $\sigma_N(k)$ is periodic. We learn that, for large N, $\sigma_N(k)$ just becomes a series of delta functions, restricting k to be integer valued

$$\lim_{N\to\infty} \sigma_N(k) = \sum_{n=-\infty}^{\infty} \delta(k-n)$$

Looking back at (3.27), we see that these delta functions mean that the Fourier transform is only non-vanishing when $\mathbf{k} = \sum_i k_i \mathbf{b}_i$ with $k_i \in \mathbf{Z}$. But this is precisely the condition that $\mathbf{k}$ lies in the reciprocal lattice. We have

$$\Delta(\mathbf{k}) = \sum_{\mathbf{r}\in\Lambda} e^{-i\mathbf{k}\cdot\mathbf{r}} = V^\star \sum_{\mathbf{q}\in\Lambda^\star} \delta(\mathbf{k}-\mathbf{q}) \tag{3.28}$$

We can understand this formula as follows: if $\mathbf{k} \in \Lambda^\star$, then $e^{-i\mathbf{k}\cdot\mathbf{r}} = 1$ for all $\mathbf{r} \in \Lambda$ and summing over all lattice points gives us infinity. In contrast, if $\mathbf{k} \notin \Lambda^\star$, then the phases $e^{-i\mathbf{k}\cdot\mathbf{r}}$ oscillate wildly for different $\mathbf{r}$ and cancel each other out. □

The upshot is that if we start with a continuous function $f(\mathbf{x})$ with periodicity Λ, then the Fourier transform (3.26) has support only at discrete points $\Lambda^\star$,

$$\tilde{f}(\mathbf{k}) = \Delta(\mathbf{k})S(\mathbf{k}) \quad \text{with} \quad S(\mathbf{k}) = \int_\Gamma d^3x\; e^{-i\mathbf{k}\cdot\mathbf{x}} f(\mathbf{x})$$

Here $S(\mathbf{k})$ is known as the *structure factor*. Alternatively, inverting the Fourier transform, we have

$$f(\mathbf{x}) = \frac{1}{(2\pi)^3}\int d^3k\; e^{i\mathbf{k}\cdot\mathbf{x}}\,\tilde{f}(\mathbf{k}) = \frac{V^\star}{(2\pi)^3}\sum_{\mathbf{q}\in\Lambda^\star} e^{i\mathbf{q}\cdot\mathbf{x}}\,S(\mathbf{q}) \tag{3.29}$$

This tells us that any periodic function is a sum of plane waves whose wavevectors lie on the reciprocal lattice, We'll revisit these ideas in Section 3.4 when we discuss x-ray scattering from a lattice.

3.2.3 The Brillouin Zone

The Wigner-Seitz cell of the reciprocal lattice is called the *Brillouin zone*.

We already saw the concept of the Brillouin zone in our one-dimensional lattice. Let's check that this coincides with the definition given above. The one-dimensional lattice is defined by a single number, a, which determines the lattice spacing. The Wigner-Seitz cell is defined as those points which lie closer to the origin than any other lattice point, namely $r \in [-a/2, a/2)$. The reciprocal lattice is defined by (3.24) which, in this context, gives the lattice spacing $b = 2\pi/a$. The Wigner-Seitz cell of this reciprocal lattice consists of those points which lie between $[-b/2, b/2) = [-\pi/a, \pi/a)$. This coincides with what we called the Brillouin zone in Section 3.1.

The Brillouin zone is also called the *first Brillouin zone*. As it is the Wigner-Seitz cell, it is defined as all points in reciprocal space that are closest to a given lattice point, say the origin. The n^{th} *Brillouin zone* is defined as all points in reciprocal space that are n^{th} closest to the origin. All these higher Brillouin zones have the same volume as the first.

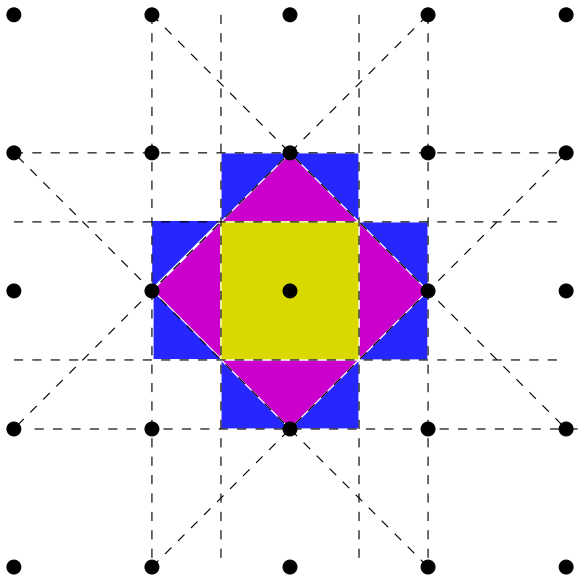


Figure 38: The Brillouin zones for a 2d square lattice. The first is shown in yellow, the second in pink, the third in blue.

We can construct the Brillouin zone boundaries by drawing the perpendicular bisectors between the origin and each other point in $\Lambda^\star$. The region enclosing the origin is the first Brillouin zone. The region you can reach by crossing just a single bisector is the second Brillouin zone, and so on. In fact, this definition generalises the Brillouin zone beyond the simple Bravais lattices.

As an example, consider the square lattice in 2d. The reciprocal lattice is also square. The first few Brillouin zones on this square lattice are shown in Figure 38.

For the one-dimensional lattice that we looked at in Section 3.1, we saw that the conserved momentum lies within the first Brillouin zone. This will also be true in higher dimensions. This motivates us to work in the *reduced zone scheme*, in which these higher Brillouin zones are mapped back into the first. This is achieved by translating them by some lattice vector. The higher Brillouin zones of the square lattice in the reduced zone scheme are shown in Figure 39.

Finally, note that the edges of the Brillouin zone should be identified; they label the same momentum state $\mathbf{k}$. For one-dimensional lattices, this results in the Brillouin zone having the topology of a circle. For d-dimensional lattices, the Brillouin zone is topologically a torus $\mathbf{T}^d$.

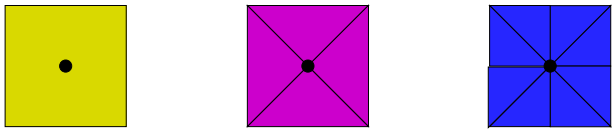


Figure 39: The first three Brillouin zones for a square lattice in the reduced zone scheme.

Crystallographic Notation

The Brillouin zone of real materials is a three-dimensional space. We often want to describe how certain quantities – such as the energy of the electrons – vary as we move around the Brillouin zone. To display this information graphically, we need to find a way to depict the underlying Brillouin zone as a two-dimensional, or even one-dimensional space. Crystallographers have developed a notation for this. Certain, highly symmetric points in the Brillouin zone are labelled by letters. From the letter, you're also supposed to remember what underlying lattice we're talking about.

For example, all Brillouin zones have an origin. The concept of an "origin" occurs in many different parts of maths and physics and almost everyone has agreed to label it as "0". Almost everyone. But not our crystallographer friends. Instead, they call the origin Γ.

From hereon, it gets more bewildering although if you stare at enough of these you get used to it. For example, for a cubic lattice, the centre of each face is called X, the centre of each edge is M while each corner is R. Various labels for BCC and FCC lattices are shown in Figure 40

3.3 Band Structure

> "When I started to think about it, I felt that the main problem was to explain how the electrons could sneak by all the ions in a metal.... I found to my delight that the wave differed from a plane wave of free electron only by a periodic modulation.This was so simple that I didn't think it could be much of a discovery, but when I showed it to Heisenberg he said right away, 'That's it.'."
>
> *Felix Bloch*

Now that we've developed the language to describe lattices in higher dimensions, it's time to understand how electrons behave when they move in the background of a fixed lattice. We already saw many of the main ideas in the context of a one-dimensional lattice in Section 3.1. Here we will describe the generalisation to higher dimensions.

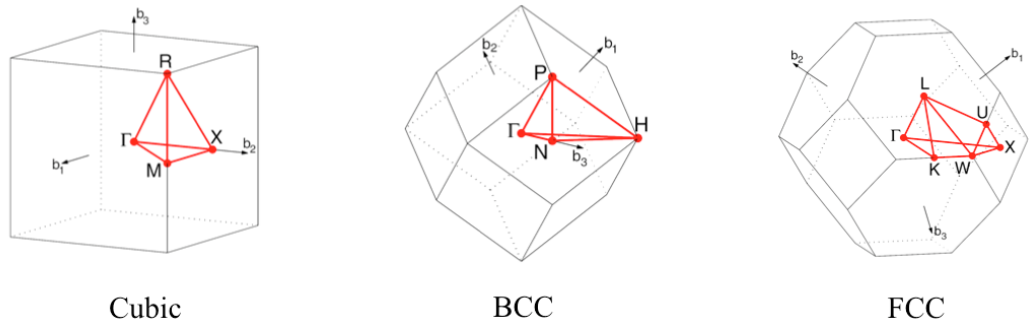


Figure 40: The labels for various special points on the Brillouin zone.

3.3.1 Bloch's Theorem

Consider an electron moving in a potential $V(\mathbf{x})$ which has the periodicity of a Bravais lattice Λ,

$$V(\mathbf{x}+\mathbf{r}) = V(\mathbf{x}) \quad \text{for all } \mathbf{r} \in \Lambda$$

Bloch's theorem states that the energy eigenstates take the form

$$\psi_{\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}\, u_{\mathbf{k}}(\mathbf{x})$$

where $u_{\mathbf{k}}(\mathbf{x})$ has the same periodicity as the lattice, $u_{\mathbf{k}}(\mathbf{x}+\mathbf{r}) = u_{\mathbf{k}}(\mathbf{x})$ for all $\mathbf{r} \in \Lambda$.

There are different ways to prove Bloch's theorem. Here we will give a simple proof using the ideas of translation operators, analogous to the one-dimensional proof that we saw in Section 3.1.4. Later, in Section 3.3.2, we will provide a more direct proof by decomposing the Schrödinger equation into Fourier modes.

Our starting point is that the Hamiltonian is invariant under discrete translations by the lattice vectors $\mathbf{r} \in \Lambda$. As we explained in Section 3.1.4, these translations are implemented by unitary operators $T_{\mathbf{r}}$. These operators form an Abelian group,

$$T_{\mathbf{r}} T_{\mathbf{r}'} = T_{\mathbf{r}+\mathbf{r}'} \tag{3.30}$$

and commute with the Hamiltonian: $[H, T_{\mathbf{r}}] = 0$. This means that we can simultaneously diagonalise H and $T_{\mathbf{r}}$, so that energy eigenstates are labelled also labelled by the eigenvalue of each $T_{\mathbf{r}}$. Because $T_{\mathbf{r}}$ is unitary, this is simply a phase. But we also have the group structure (3.30) that must be respected. Suppose that translation of a given eigenstate by a basis element $\mathbf{a}_i$ gives eigenvalue

$$T_{\mathbf{a}_i}\psi(\mathbf{x}) = \psi(\mathbf{x}+\mathbf{a}_i) = e^{i\theta_i}\psi(\mathbf{x})$$

Then translation by a general lattice vector $\mathbf{r} = \sum_i n_i \mathbf{a}_i$ must give

$$T_{\mathbf{r}}\psi(\mathbf{x}) = \psi(\mathbf{x}+\mathbf{r}) = e^{i\sum_i n_i\theta_i}\psi(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{r}}\psi(\mathbf{x})$$

where the vector $\mathbf{k}$ is defined by $\mathbf{k}\cdot\mathbf{a}_i = \theta_i$. In other words, we can label eigenstates of $T_{\mathbf{r}}$ by a vector $\mathbf{k}$. They obey

$$T_{\mathbf{r}}\psi_{\mathbf{k}}(\mathbf{x}) = \psi_{\mathbf{k}}(\mathbf{x}+\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}\psi_{\mathbf{k}}(\mathbf{x})$$

Now we simply need to look at the function $u_{\mathbf{k}}(\mathbf{x}) = e^{-i\mathbf{k}\cdot\mathbf{x}}\psi_{\mathbf{k}}(\mathbf{x})$. The statement of Bloch's theorem is that $u_{\mathbf{k}}(\mathbf{x})$ has the periodicity of Λ which is indeed true, since $u_{\mathbf{k}}(\mathbf{x}+\mathbf{r}) = e^{-i\mathbf{k}\cdot\mathbf{x}}e^{-i\mathbf{k}\cdot\mathbf{r}}\psi_{\mathbf{k}}(\mathbf{x}+\mathbf{r}) = e^{-i\mathbf{k}\cdot\mathbf{x}}\psi_{\mathbf{k}}(\mathbf{x}) = u_{\mathbf{k}}(\mathbf{x})$.

Crystal Momentum

The energy eigenstates are labelled by the wavevector $\mathbf{k}$, called the *crystal momentum*. There is an ambiguity in the definition of this crystal momentum. This is not the same as the true momentum. The energy eigenstates do not have a well defined momentum because they are not eigenstates of the momentum operator $\mathbf{p} = -i\hbar\nabla$ unless $u_{\mathbf{k}}(\mathbf{x})$ is constant. Nonetheless, we will see as we go along that the crystal momentum plays a role similar to the true momentum. For this reason, we will often refer to $\mathbf{k}$ simply as "momentum".

There is an ambiguity in the definition of the crystal momentum. Consider a state with a crystal momentum $\mathbf{k}' = \mathbf{k} + \mathbf{q}$, with $\mathbf{q} \in \Lambda^\star$ a reciprocal lattice vector. Then

$$\psi_{\mathbf{k}'}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}e^{i\mathbf{q}\cdot\mathbf{x}}u_{\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}\tilde{u}_{\mathbf{k}}(\mathbf{x})$$

where $\tilde{u}_{\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{q}\cdot\mathbf{x}}u_{\mathbf{k}}(\mathbf{x})$ also has the periodicity of Λ by virtue of the definition of the reciprocal lattice (3.25).

As in the one-dimensional example, we have different options. We could choose to label states by $\mathbf{k}$ which lie in the first Brillouin zone. In this case, there will typically be many states with the same $\mathbf{k}$ and different energies. This is the *reduced zone scheme*. In this case, the energy eigenstates are labelled by two indices, $\psi_{\mathbf{k},n}$, where $\mathbf{k}$ is the crystal momentum and n is referred to as the *band index*. (We will see examples shortly.)

Alternatively, we can label states by taking any $\mathbf{k} \in \mathbf{R}^d$ where d is the dimension of the problem. This is the *extended zone scheme*. In this case that states labelled by $\mathbf{k}$ which differ by $\Lambda^\star$ have the same crystal momenta.

3.3.2 Nearly Free Electrons in Three Dimensions

Consider an electron moving in $\mathbf{R}^3$ in the presence of a weak potential $V(\mathbf{x})$. We'll assume that this potential has the periodicity of a Bravais lattice Λ, so

$$V(\mathbf{x}) = V(\mathbf{x} + \mathbf{r}) \quad \text{for all } \mathbf{r} \in \Lambda$$

We treat this potential as a perturbation on the free electron. This means that we start with plane wave states $|\mathbf{k}\rangle$ with wavefunctions

$$\langle \mathbf{x}|\mathbf{k}\rangle \sim e^{i\mathbf{k}\cdot\mathbf{x}}$$

with energy $E_0(\mathbf{k}) = \hbar k^2/2m$. We want to see how these states and their energy levels are affected by the presence of the potential. The discussion will follow closely the one-dimensional case that we saw in Section 3.1.2 and we only highlight the differences.

When performing perturbation theory, we're going to have to consider the potential $V(\mathbf{x})$ sandwiched between plane-wave states,

$$\langle \mathbf{k}|V(\mathbf{x})|\mathbf{k}'\rangle = \frac{1}{\text{Volume}} \int d^3x \; e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{x}} V(\mathbf{x})$$

However, we've already seen in (3.29) that the Fourier transform of a periodic function can be written as a sum over wavevectors that lie in the reciprocal lattice $\Lambda^\star$,

$$V(\mathbf{x}) = \sum_{\mathbf{q}\in\Lambda^\star} e^{i\mathbf{q}\cdot\mathbf{x}} V_{\mathbf{q}}$$

(Note: here $V_{\mathbf{q}}$ is the Fourier component of the potential and should not be confused with the volumes of unit cells which were denoted as V and $V^\star$ in Section 3.2.) This means that $\langle \mathbf{k}|V(\mathbf{x})|\mathbf{k}'\rangle$ is non-vanishing only when the two momenta differ by

$$\mathbf{k} - \mathbf{k}' = \mathbf{q} \quad \mathbf{q} \in \Lambda^\star$$

This has a simple physical interpretation: a plane wave state $|\mathbf{k}\rangle$ can scatter into another plane wave state $|\mathbf{k}'\rangle$ only if they differ by a reciprocal lattice vector. In other words, only momenta $\mathbf{q}$, with $\mathbf{q} \in \Lambda^\star$, can be absorbed by the lattice.

Another Perspective on Bloch's Theorem

The fact that that a plane wave state $|\mathbf{k}\rangle$ can only scatter into states $|\mathbf{k} - \mathbf{q}\rangle$, with $\mathbf{q} \in \Lambda^\star$, provides a simple viewpoint on Bloch's theorem, one that reconciles the quantum state with the naive picture of the particle bouncing off lattice sites like a ball in a pinball machine. Suppose that the particle starts in some state $|\mathbf{k}\rangle$. After scattering,

we might expect it to be some superposition of all the possible scattering states $|\mathbf{k}-\mathbf{q}\rangle$. In other words,

$$\psi_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{q}\in\Lambda^\star} e^{i(\mathbf{k}-\mathbf{q})\cdot\mathbf{x}}\, c_{\mathbf{k}-\mathbf{q}}$$

for some coefficients $c_{\mathbf{k}-\mathbf{q}}$. We can write this as

$$\psi_{\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}} \sum_{\mathbf{q}\in\Lambda^\star} e^{-i\mathbf{q}\cdot\mathbf{x}}\, c_{\mathbf{k}-\mathbf{q}} = e^{i\mathbf{k}\cdot\mathbf{x}}\, u_{\mathbf{k}}(\mathbf{x})$$

where, by construction, $u_{\mathbf{k}}(\mathbf{x}+\mathbf{r}) = u_{\mathbf{k}}(\mathbf{x})$ for all $\mathbf{r}\in\Lambda$. But this is precisely the form guaranteed by Bloch's theorem.

Although the discussion here holds at first order in perturbation theory, it is not hard to extend this argument to give an alternative proof of Bloch's theorem, which essentially comes down to analysing the different Fourier modes of the Schrödinger equation.

Band Structure

Let's now look at what becomes of the energy levels after we include the perturbation. We will see that, as in the 1d example, they form bands. The resulting eigenstates $\psi_{\mathbf{k},n}(\mathbf{x})$ and their associated energy levels $E_n(\mathbf{k})$ are referred to as the *band structure* of the system.

Low Momentum: Far from the edge of the Brillouin zone, the states $|\mathbf{k}\rangle$ can only scatter into states $|\mathbf{k}+\mathbf{q}\rangle$ with greatly different energy. In this case, we can work with non-degenerate perturbation theory to compute the corrections to the energy levels.

On the Boundary of the Brillouin zone: Things get more interesting when we have to use degenerate perturbation theory. This occurs whenever the state $|\mathbf{k}\rangle$ has the same energy as another state $|\mathbf{k}+\mathbf{q}\rangle$ with $\mathbf{q}\in\Lambda^\star$,

$$E_0(\mathbf{k}) = E_0(\mathbf{k}+\mathbf{q}) \quad\Rightarrow\quad k^2 = (\mathbf{k}+\mathbf{q})^2 \quad\Rightarrow\quad 2\mathbf{k}\cdot\mathbf{q} + q^2 = 0$$

This condition is satisfied whenever we can write

$$\mathbf{k} = -\frac{1}{2}\mathbf{q} + \mathbf{k}_\perp$$

where $\mathbf{q}\cdot\mathbf{k}_\perp = 0$. This is the condition that we sit on the perpendicular bisector of the origin and the lattice point $-\mathbf{q}\in\Lambda^\star$. But, as we explained in Section 3.2.3, these bisectors form the boundaries of the Brillouin zones. We learn something important: momentum states are degenerate only when they lie on the boundary of a Brillouin zone. This agrees with what we found in our one-dimensional example in Section 3.1.2.

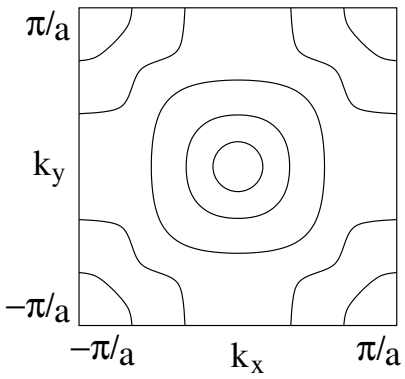


Figure 41: Energy contours for nearly-free electrons in the first Brillouin zone.

We know from experience what the effect of the perturbation $V(\mathbf{x})$ will be: it will lift the degeneracy. This means that a gap opens at the boundary of the Brillouin zone. For example, the energy of states just inside the first Brillouin zone will be pushed down, while the energy of those states just outside the first Brillouin zone will be pushed up. Note that the size of this gap will vary as we move around the boundary..

There is one further subtlety that we should mention. At a generic point on the boundary of the Brillouin zone, the degeneracy will usually be two-fold. However, at special points — such as edges, or corners — it is often higher. In this case, we must work with all degenerate states when computing the gap.

All of this is well illustrated with an example. However, it's illustrated even better if you do the example yourself! The problem of nearly free electrons in a two-dimensional square lattice is on the problem sheet. The resulting energy contours are shown in Figure 41.

Plotting Band Structures in Three Dimensions

For three-dimensional lattice, we run into the problem of depicting the bands. For this, we need the crystallographer's notation we described previously. The spectrum of free particles (i.e. with no lattice) is plotted in the Brillouin zone of BCC and FCC lattices in Figure 42[5].

We can then compare this to the band structure of real materials. The dispersion relation for silicon is also shown in Figure 42. This has a diamond lattice structure, which is plotted as FCC. Note that you can clearly see the energy gap of around 1.1 eV between the bands.

[5]Images plotted by Jan-Rens Reitsma, from Wikimedia commons.

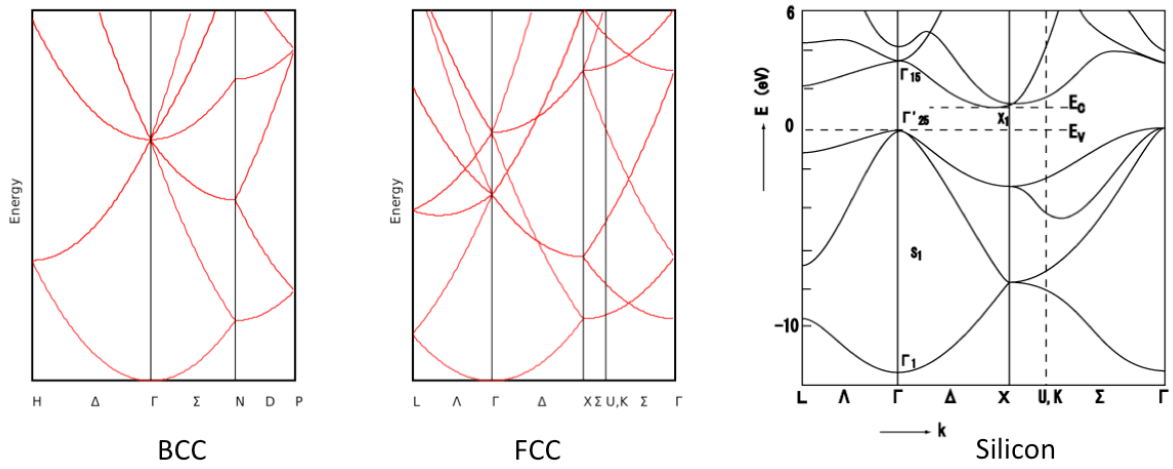


Figure 42: Free band structure (in red) for BCC and FCC, together with the band structure for silicon, exhibiting a gap.

How Many States in the Brillouin Zone?

The Brillouin zone consists of all wavevectors $\mathbf{k}$ that lie within the Wigner-Seitz cell of the reciprocal lattice $\Lambda^\star$. How many quantum states does it hold? Well, if the spatial lattice Λ is infinite in extent then $\mathbf{k}$ can take any continuous value and there are an infinite number of states in the Brillouin zone. But what if the spatial lattice is finite in size?

In this section we will count the number of quantum states in the Brillouin zone of a finite spatial lattice Λ. We will find a lovely answer: the number of states is equal to N, the number of lattice sites.

Recall that the lattice Λ consists of all vectors $\mathbf{r} = \sum_i n_i \mathbf{a}_i$ where $\mathbf{a}_i$ are the primitive lattice vectors and $n_i \in \mathbf{Z}$. For a finite lattice, we simply restrict the value of these integers to be

$$0 \leq n_i < N_i$$

for some N_i. The total number of lattice sites is then $N = N_1 N_2 N_3$ (assuming a three-dimensional lattice). The total volume of the lattice is VN where $V = |\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)|$ is the volume of the unit cell.

The basic physics is something that we've met before: if we put a particle in a box, then the momentum $\hbar\mathbf{k}$ becomes quantised. This arises because of the boundary

conditions that we place on the wavefunction. It's simplest to think about a finite, periodic lattice where we require that the wavefunction inherits this periodicity, so that

$$\psi(\mathbf{x} + N_i\mathbf{a}_i) = \psi(\mathbf{x}) \quad \text{for each } i = 1, 2, 3 \tag{3.31}$$

But we know from Bloch's theorem that energy eigenstates take the form $\psi_{\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}u_{\mathbf{k}}(\mathbf{x})$ where $u_{\mathbf{k}}(\mathbf{x} + \mathbf{a}_i) = u_{\mathbf{k}}(\mathbf{x})$. This means that the periodicity condition (3.31) becomes

$$e^{iN_i\mathbf{k}\cdot\mathbf{a}_i} = 1 \quad \Rightarrow \quad \mathbf{k} = \sum_i \frac{m_i}{N_i}\mathbf{b}_i$$

where $m_i \in \mathbf{Z}$ and $\mathbf{b}_i$ are the primitive vectors of the reciprocal lattice defined in (3.24). This is sometimes called the *Born-von Karmen* boundary condition.

This is the quantisation of momentum that we would expect in a finite system. The states are now labelled by integers $m_i \in \mathbf{Z}$. Each state can be thought of as occupying a volume in $\mathbf{k}$-space, given by

$$\frac{|\mathbf{b}_1 \cdot (\mathbf{b}_2 \times \mathbf{b}_3)|}{N_1N_2N_3} = \frac{V^\star}{N}$$

where $V^\star$ is the volume of the Brillouin zone. We see that the number of states that live inside the Brillouin zone is precisely N, the number of sites in the spatial lattice.

3.3.3 Wannier Functions

Bloch's theorem tells that the energy eigenstates can be written in the form

$$\psi_{\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}u_{\mathbf{k}}(\mathbf{x})$$

with $\mathbf{k}$ lying in the first Brillouin zone and $u_{\mathbf{k}}(\mathbf{x})$ a periodic function. Clearly these are delocalised throughout the crystal. For some purposes, it's useful to think about these Bloch waves as arising from the sum of states, each of which is localised at a given lattice site. These states are called *Wannier functions*; they are defined as

$$w_{\mathbf{r}}(\mathbf{x}) = \frac{1}{\sqrt{N}}\sum_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{r}}\,\psi_{\mathbf{k}}(\mathbf{x}) \tag{3.32}$$

where the sum is over all $\mathbf{k}$ in the first Brillouin zone.

The basic idea is that the Wannier wavefunction $w_{\mathbf{r}}(\mathbf{x})$ is localised around the lattice site $\mathbf{r} \in \Lambda$. Indeed, using the periodicity properties of the Bloch wavefunction, it's simple to show that $w_{\mathbf{r}+\mathbf{r}'}(\mathbf{x} + \mathbf{r}') = w_{\mathbf{r}}(x)$, which means that we can write $w_{\mathbf{r}}(\mathbf{x}) = w(\mathbf{x} - \mathbf{r})$.

The Wannier functions aren't unique. We can always do a phase rotation $\psi_{\mathbf{k}}(\mathbf{x}) \to e^{i\chi(\mathbf{k})}\psi_{\mathbf{k}}(\mathbf{k})$ in the definition (3.32). Different choices of $\chi(\mathbf{k})$ result in differing amounts of localisation of the state $w_{\mathbf{r}}(\mathbf{x})$ around the lattice site $\mathbf{r}$.

We can invert the definition of the Wannier function to write the original Bloch wavefunction as

$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{r}\in\Lambda} e^{i\mathbf{k}\cdot\mathbf{r}}\, w(\mathbf{x}-\mathbf{r}) \tag{3.33}$$

which follows from (3.28).

The Wannier functions have one final, nice property: they are orthonormal in the sense that

$$\begin{aligned}\int d^3x\; w^\star(\mathbf{x}-\mathbf{r}')w(\mathbf{x}-\mathbf{r}) &= \frac{1}{N}\int d^3x \sum_{\mathbf{k},\mathbf{k}'} e^{i\mathbf{k}'\cdot\mathbf{r}'-i\mathbf{k}\cdot\mathbf{r}}\psi^\star_{\mathbf{k}'}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{x}) \\ &= \frac{1}{N}\sum_{\mathbf{k}} e^{i\mathbf{k}\cdot(\mathbf{r}'-\mathbf{r})} = \delta(\mathbf{r}-\mathbf{r}')\end{aligned}$$

where, in going to the second line, we have used the orthogonality of Bloch wavefunctions for different $\mathbf{k}$ (which, in turn, follows because they are eigenstates of the Hamiltonian with different energies).

3.3.4 Tight-Binding in Three Dimensions

We started our discussion of band structure in Section 3.1.1 with the one-dimensional tight binding model. This is a toy Hamiltonian describing electrons hopping from one lattice site to another. Here we'll look at this same class of models in higher dimensional lattices.

We assume that the electron can only sit on a site of the lattice $\mathbf{r}\in\Lambda$. The Hilbert space is then spanned by the states $|\mathbf{r}\rangle$ with $\mathbf{r}\in\Lambda$. We want to write down a Hamiltonian which describes a particle hopping between these sites. There are many different ways to do this; the simplest is

$$H = \sum_{\mathbf{r}\in\Lambda} E_0|\mathbf{r}\rangle\langle\mathbf{r}| - \sum_{<\mathbf{r}\mathbf{r}'>} t_{\mathbf{r}'-\mathbf{r}}\Big(|\mathbf{r}\rangle\langle\mathbf{r}'| + |\mathbf{r}'\rangle\langle\mathbf{r}|\Big)$$

where the label $<\mathbf{r}\mathbf{r}'>$ means that we only sum over pairs of sites $\mathbf{r}$ and $\mathbf{r}'$ which are nearest neighbours in the lattice. Alternatively, if these nearest neighbours are connected by a set of lattice vectors $\mathbf{a}$, then we can write this as

$$H = \sum_{\mathbf{r}\in\Lambda}\left[E_0|\mathbf{r}\rangle\langle\mathbf{r}| - \sum_{\mathbf{a}} t_{\mathbf{a}}\,|\mathbf{r}\rangle\langle\mathbf{r}+\mathbf{a}|\right] \tag{3.34}$$

Note that we've just got one term here, since if $|\mathbf{r}+\mathbf{a}\rangle$ is a nearest neighbour, then so is $|\mathbf{r}-\mathbf{a}\rangle$. The Hamiltonian is Hermitian provided $t_{\mathbf{a}} = t_{-\mathbf{a}}$ This Hamiltonian is easily solved. The eigenstates take the form

$$|\psi(\mathbf{k})\rangle = \frac{1}{\sqrt{N}} \sum_{\mathbf{r}\in\Lambda} e^{i\mathbf{k}\cdot\mathbf{r}} |\mathbf{r}\rangle \tag{3.35}$$

where N is the total number of lattice sites. It's simple to check that these states satisfy $H|\psi(\mathbf{k})\rangle = E(\mathbf{k})|\psi(\mathbf{k})\rangle$ with

$$E(\mathbf{k}) = E_0 - \frac{1}{2}\sum_{\mathbf{a}} 2t_{\mathbf{a}} \cos(\mathbf{k}\cdot\mathbf{a}) \tag{3.36}$$

where the factor of $1/2$ is there because we are still summing over all nearest neighbours, including $\pm\mathbf{a}$. This exhibits all the properties of that we saw in the tight-binding model. The energy eigenstates (3.35) are no longer localised, but are instead spread throughout the lattice. The states form just a single band labelled, as usual, but by crystal momentum $\mathbf{k}$ lying in the first Brillouin zone. This is to be expected in the tight-binding model as we start with N states, one per lattice site, and we know that each Brillouin zone accommodates precisely N states.

As a specific example, consider a cubic lattice. The nearest neighbour lattice sites are $\mathbf{a} \in \Big\{(\pm a, 0, 0), (0, \pm a, 0), (0, 0, \pm a)\Big\}$ and the hopping parameters are the same in all directions: $t_{\mathbf{a}} = t$. The dispersion relation is then given by

$$E(\mathbf{k}) = E_0 - 2t\Big(\cos(k_x a) + \cos(k_y a) + \cos(k_z a)\Big) \tag{3.37}$$

The width of this band is $\Delta E = E_{\max} - E_{\min} = 12t$.

Note that for small k, the dispersion relation takes the form of a free particle

$$E(\mathbf{k}) = \text{constant} + \frac{\hbar^2\mathbf{k}^2}{2m^\star} + \ldots$$

where the effective mass $m^\star$ is determined by various parameters of the underlying lattice, $m^\star = \hbar^2/2ta^2$. However, at higher k the energy is distorted away from the that of a free particle. For example, you can check that $k_x \pm k_y = \mp\pi/a$ (with $k_z = 0$) is a line of constant energy.

3.3.5 Deriving the Tight-Binding Model

Above, we have simply written down the tight-binding model. But it's interesting to ask how we can derive it from first principles. In particular, this will tell us what physics it captures and what physics it misses.

To do this, we start by considering a single atom which we place at the origin. The Hamiltonian for a single electron orbiting this atom takes the familiar form

$$H_{\rm atom} = \frac{\mathbf{p}^2}{2m} + V_{\rm atom}(\mathbf{x})$$

The electrons will bind to the atom with eigenstates $\phi_n(\mathbf{x})$ and discrete energies $\epsilon_n < 0$, which obey

$$H_{\rm atom}\phi_n(\mathbf{x}) = \epsilon_n\phi_n(\mathbf{x})$$

A sketch of a typical potential $V_{\rm atom}(\mathbf{x})$ and the binding energies ϵ_n is shown on the right. There will also be scattering states, with energies $\epsilon > 0$, which are not bound to the atom.

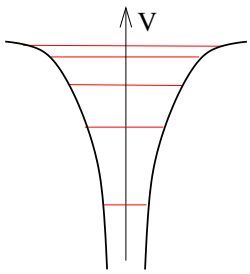


Figure 43:

Our real interest lies in a lattice of these atoms. The resulting potential is

$$V_{\rm lattice}(\mathbf{x}) = \sum_{\mathbf{r}\in\Lambda} V_{\rm atom}(\mathbf{x}-\mathbf{r})$$

This is shown in Figure 44 for a one-dimensional lattice. What happens to the energy levels? Roughly speaking, we expect those electrons with large binding energies — those shown at the bottom of the spectrum — to remain close to their host atoms. But those that are bound more weakly become free to move. This happens because the tails of their wavefunctions have substantial overlap with electrons on neighbouring atoms, causing these states to mix. This is the physics captured by the tight-binding model.

The weakly bound electrons which become dislodged from their host atoms are called *valence electrons*. (These are the same electrons which typically sit in outer shells and give rise to bonding in chemistry.) As we've seen previously, these electrons will form a band of extended states.

Let's see how to translate this intuition into equations. We want to solve the Hamiltonian

$$H = \frac{\mathbf{p}^2}{2m} + V_{\rm lattice}(\mathbf{x}) \tag{3.38}$$

Our goal is to write the energy eigenstates in terms of the localised atomic states $\phi_n(\mathbf{x})$. Getting an exact solution is hard; instead, we're going to guess an approximate solution.

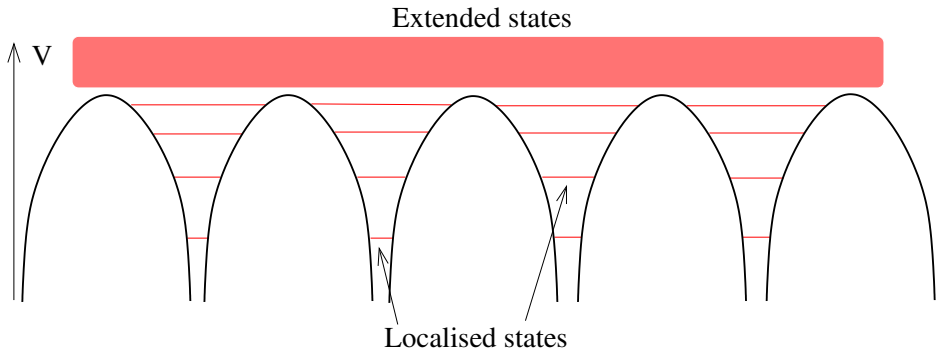


Figure 44: Extended and localised states in a lattice potential.

First, let's assume that there is just a single valence electron with localised wavefunction $\phi(\mathbf{x})$ with energy ϵ . We know that the eigenstates of (3.38) must have Bloch form. We can build such a Bloch state from the localised state $\phi(\mathbf{x})$ by writing

$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{r}\in\Lambda} e^{i\mathbf{k}\cdot\mathbf{r}}\, \phi(\mathbf{x}-\mathbf{r}) \tag{3.39}$$

where N is the number of lattice sites. This is a Bloch state because for any $\mathbf{a} \in \Lambda$, we have $\psi_{\mathbf{k}}(\mathbf{x}+\mathbf{a}) = e^{i\mathbf{k}\cdot\mathbf{a}}\,\psi_{\mathbf{k}}(\mathbf{x})$. Note that this is the same kind of state (3.35) that solved our original tight-binding model. Note also that this ansatz takes the same form as the expansion in terms of Wannier functions (3.33). However, in contrast to Wannier functions, the wavefunctions $\phi(\mathbf{x})$ localised around different lattice sites are not orthogonal. This difference will be important below.

The expected energy for the state (3.39) is

$$E(\mathbf{k}) = \frac{\langle\psi_{\mathbf{k}}|H|\psi_{\mathbf{k}}\rangle}{\langle\psi_{\mathbf{k}}|\psi_{\mathbf{k}}\rangle}$$

First, the denominator.

$$\begin{aligned}
\langle\psi_{\mathbf{k}}|\psi_{\mathbf{k}}\rangle &= \frac{1}{N}\sum_{\mathbf{r},\mathbf{r}'\in\Lambda} e^{i\mathbf{k}\cdot(\mathbf{r}'-\mathbf{r})}\int d^3x\ \phi^\star(\mathbf{x}-\mathbf{r})\phi(\mathbf{x}-\mathbf{r}') \\
&= \sum_{\mathbf{r}\in\Lambda} e^{-i\mathbf{k}\cdot\mathbf{r}}\int d^3x\ \phi^\star(\mathbf{x}-\mathbf{r})\phi(\mathbf{x}) \\
&\equiv 1 + \sum_{\mathbf{r}\neq 0} e^{-i\mathbf{k}\cdot\mathbf{r}}\alpha(\mathbf{r})
\end{aligned}$$

where, in going to the second line, we've used the translational invariance of the lattice. The function $\alpha(\mathbf{r})$ measures the overlap of the wavefuntions localised at lattice sites separated by $\mathbf{r}$.

Next the numerator. To compute this, we write $H = H_{\rm atom} + \Delta V(\mathbf{x})$ where

$$\Delta V(\mathbf{x}) = V_{\rm lattice}(\mathbf{x}) - V_{\rm atom}(\mathbf{x}) = \sum_{\mathbf{r}\in\Lambda, \mathbf{r}\neq 0} V_{\rm atom}(\mathbf{x}-\mathbf{r})$$

We then have

$$\begin{aligned}\langle\psi_{\mathbf{k}}|H|\psi_{\mathbf{k}}\rangle &= \frac{1}{N}\sum_{\mathbf{r},\mathbf{r}'\in\Lambda} e^{i\mathbf{k}\cdot(\mathbf{r}'-\mathbf{r})}\int d^3x\ \phi^\star(\mathbf{x}-\mathbf{r})(H_{\rm atom}+\Delta V)\phi(\mathbf{x}-\mathbf{r}') \\ &= \sum_{\mathbf{r}\in\Lambda} e^{-i\mathbf{k}\cdot\mathbf{r}}\int d^3x\ \phi^\star(\mathbf{x}-\mathbf{r})(H_{\rm atom}+\Delta V)\phi(\mathbf{x}) \\ &\equiv \epsilon\langle\psi_{\mathbf{k}}|\psi_{\mathbf{k}}\rangle + \Delta\epsilon + \sum_{\mathbf{r}\neq 0} e^{-i\mathbf{k}\cdot\mathbf{r}}\gamma(\mathbf{r})\end{aligned}$$

Here $\Delta\epsilon$ is the shift in the energy of the bound state $\phi(\mathbf{x})$ due to the potential ΔV,

$$\Delta\epsilon = \int d^3x\ \phi^\star(\mathbf{x})\Delta V(\mathbf{x})\phi(\mathbf{x})$$

Meanwhile, the last term arises from the overlap of localised atoms on different sites

$$\gamma(\mathbf{r}) = \int d^3x\ \phi^\star(\mathbf{x}-\mathbf{r})\,\Delta V(\mathbf{x})\,\phi(\mathbf{x})$$

The upshot of this is an expression for the expected energy of the Bloch wave (3.39)

$$E(\mathbf{k}) = \epsilon + \frac{\Delta\epsilon + \sum_{\mathbf{r}\neq 0} e^{-i\mathbf{k}\cdot\mathbf{r}}\,\gamma(\mathbf{r})}{1 + \sum_{\mathbf{r}\neq 0} e^{-i\mathbf{k}\cdot\mathbf{r}}\,\alpha(\mathbf{r})}$$

Under the assumption that $\alpha(\mathbf{r}) \ll 1$, we can expand out the denominator $(1+x)^{-1} \approx 1-x$, and write

$$E(\mathbf{k}) = \epsilon + \Delta\epsilon + \sum_{\mathbf{r}\neq 0} e^{-i\mathbf{k}\cdot\mathbf{r}}\left(\gamma(\mathbf{r}) - \alpha(\mathbf{r})\,\Delta\epsilon\right) \tag{3.40}$$

This still looks rather complicated. However, the expression simplifies because the overlap functions $\alpha(\mathbf{r})$ and $\gamma(\mathbf{r})$ both drop off quickly with separation. Very often, it's sufficient to take these to be non-zero only when $\mathbf{r}$ are the nearest neighbour lattice sites. Sometimes we need to go to next-to-nearest neighbours.

An Example: s-Orbitals

Let's assume that $\alpha(\mathbf{r})$ and $\gamma(\mathbf{r})$ are important only for $\mathbf{r}$ connecting nearest neighbour lattice sites; all others will be taken to vanish. We'll further take the valence electron

to sit in the s-orbital. This has two consequences: first, the localised wavefunction is rotationally invariant, so that $\phi(\mathbf{r}) = \phi(r)$. Second, the wavefunction can be taken to be real, so $\phi^\star(\mathbf{x}) = \phi(\mathbf{x})$. With these restrictions, we have

$$\alpha(\mathbf{r}) = \int d^3x \phi(\mathbf{x}-\mathbf{r})\phi(\mathbf{x}) = \alpha(-\mathbf{r})$$

We want a similar expression for $\gamma(\mathbf{r})$. For this, we need to make one further assumption: we want the crystal to have *inversion symmetry*. This means that $V(\mathbf{x}) = V(-\mathbf{x})$ or, more pertinently for us, $\Delta V(\mathbf{x}) = \Delta V(-\mathbf{x})$. We can then write

$$\begin{aligned}\gamma(\mathbf{r}) &= \int d^3x\ \phi(\mathbf{x}-\mathbf{r})\Delta V(\mathbf{x})\phi(\mathbf{x}) \\ &= \int d^3x'\ \phi(-\mathbf{x}'-\mathbf{r})\Delta V(-\mathbf{x}')\phi(-\mathbf{x}') \\ &= \int d^3x' \phi(|\mathbf{x}'+r|)\Delta V(\mathbf{x}')\phi(|\mathbf{x}'|) \\ &= \gamma(-\mathbf{r})\end{aligned}$$

where we have defined $\mathbf{x}' = -\mathbf{x}$ in the second line and used both the inversion symmetry and rotational invariance of the s-orbital in the third. Now we can write the energy (3.40) in a slightly nicer form. We need to remember that the vectors $\mathbf{r}$ span a lattice which ensures that if $\mathbf{r}$ is a nearest neighbour site then $-\mathbf{r}$ is too. We then have

$$E(\mathbf{k}) = \epsilon + \Delta\epsilon + \sum_{\mathbf{a}} \cos(\mathbf{k}\cdot\mathbf{a})\Big(\gamma(\mathbf{a}) - \Delta\epsilon\,\alpha(\mathbf{a})\Big) \tag{3.41}$$

where $\mathbf{a}$ are the nearest neighbour lattice sites. We recognise this as the dispersion relation that we found in our original tight-binding model (3.36), with $E_0 = \epsilon + \Delta\epsilon$ and $t_{\mathbf{a}} = \gamma(\mathbf{a}) - \Delta\epsilon\,\alpha(\mathbf{a})$.

So far we've shown that the state (3.39) has the same energy as eigenstates of the tight-binding Hamiltonian. But we haven't yet understood when the state (3.39) is a good approximation to the true eigenstate of the Hamiltonian (3.38).

We can intuit the answer to this question by looking in more detail at (3.41). We see that the localised eigenstates $\phi(\mathbf{x})$, each of which had energy ϵ, have spread into a band with energies $E(\mathbf{k})$. For this calculation to be valid, it's important that this band doesn't mix with other states. This means that the energies $E(\mathbf{k})$ shouldn't be too low, so that it has overlap with the energies of more deeply bound states. Nor should $E(\mathbf{k})$ be too high, so that it overlaps with the energies of the scattering states which will give rise to higher bands. If the various lattice parameters are chosen so that it

sits between these two values, our ansatz (3.39) will be a good approximation to the true wavefunction. Another way of saying this is that if we focus on states in the first band, we can approximate the Hamiltonian (3.38) describing a lattice of atoms by the tight-binding Hamiltonian (3.34).

A Linear Combination of Atomic Orbitals

What should we do if the band of interest does overlap with bands from more deeply bound states? The answer is that we should go back to our original ansatz (3.39) and replace it with something more general, namely

$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{r}\in\Lambda} e^{i\mathbf{k}\cdot\mathbf{r}} \sum_n c_n \phi_n(\mathbf{x}-\mathbf{r}) \tag{3.42}$$

where this time we sum over all localised states of interest, $\phi_n(\mathbf{x})$ with energies ϵ_n. These are now weighted with coefficients c_n which we will determine shortly. This kind of ansatz is known as a *linear combination of atomic orbitals*. Among people who play these kind of games, it is common enough to have its own acronym (*LCAO* obviously).

The wavefunction (3.42) should be viewed as a variational ansatz for the eigenstates, where we get to vary the parameters c_n. The expected energy is again

$$E(\mathbf{k}) = \frac{\langle\psi_{\mathbf{k}}|H|\psi_{\mathbf{k}}\rangle}{\langle\psi_{\mathbf{k}}|\psi_{\mathbf{k}}\rangle}$$

where, repeating the calculations that we just saw, we have

$$\begin{aligned}\langle\psi_{\mathbf{k}}|\psi_{\mathbf{k}}\rangle &= \sum_{\mathbf{r}\in\Lambda}\sum_{n,n'} c^\star_{n'} c_n\, e^{-i\mathbf{k}\cdot\mathbf{r}} \int d^3x\ \phi^\star_{n'}(\mathbf{x}-\mathbf{r})\phi_n(\mathbf{x}) \\ &\equiv \sum_{\mathbf{r}\in\Lambda}\sum_{n,n'} c^\star_{n'} c_n\, e^{-i\mathbf{k}\cdot\mathbf{r}} \alpha_{n,n'}(\mathbf{r})\end{aligned} \tag{3.43}$$

and

$$\begin{aligned}\langle\psi_{\mathbf{k}}|H|\psi_{\mathbf{k}}\rangle &= \sum_{\mathbf{r}\in\Lambda}\sum_{n,n'} c^\star_{n'} c_n\, e^{-i\mathbf{k}\cdot\mathbf{r}} \int d^3x\ \phi^\star_{n'}(\mathbf{x}-\mathbf{r})(H_{\rm atom}+\Delta V)\phi_n(\mathbf{x}) \\ &\equiv \sum_{\mathbf{r}\in\Lambda}\sum_{n,n'} c^\star_{n'} c_n\, e^{-i\mathbf{k}\cdot\mathbf{r}} \Big(\epsilon_n \alpha_{n,n'}(\mathbf{r}) + \gamma_{n,n'}(\mathbf{r})\Big)\end{aligned} \tag{3.44}$$

Note that we've used slightly different notation from before. We haven't isolated the piece $\alpha_{n,n'}(\mathbf{r}=0) = \delta_{n,n'}$, nor the analogous $\Delta\epsilon$ piece corresponding to $\gamma_{n,n'}(\mathbf{r}=0)$. Instead, we continue to sum over all lattice points $\mathbf{r}\in\Lambda$, including the origin.

The variational principle says that we should minimise the expected energy over all c_n. This means we should solve

$$\frac{\partial E(\mathbf{k})}{\partial c^\star_{n'}} = \frac{1}{\langle\psi_\mathbf{k}|\psi_\mathbf{k}\rangle}\frac{\partial}{\partial c^\star_{n'}}\langle\psi_\mathbf{k}|H|\psi_\mathbf{k}\rangle - \frac{\langle\psi_\mathbf{k}|H|\psi_\mathbf{k}\rangle}{\langle\psi_\mathbf{k}|\psi_\mathbf{k}\rangle^2}\frac{\partial}{\partial c^\star_{n'}}\langle\psi_\mathbf{k}|\psi_\mathbf{k}\rangle = 0$$
$$\Rightarrow \qquad \frac{\partial}{\partial c^\star_{n'}}\langle\psi_\mathbf{k}|H|\psi_\mathbf{k}\rangle - E(\mathbf{k})\frac{\partial}{\partial c^\star_{n'}}\langle\psi_\mathbf{k}|\psi_\mathbf{k}\rangle = 0$$

Using our expressions (3.43) and (3.44), we can write the resulting expression as the matrix equation

$$\sum_n M_{n,n'}(\mathbf{k})c_n = 0 \tag{3.45}$$

where $M_{n,n'}(\mathbf{k})$ is the Hermitian matrix

$$M_{n,n'}(\mathbf{k}) = \sum_{\mathbf{r}\in\Lambda} e^{-i\mathbf{k}\cdot\mathbf{r}}\Big(\tilde{\gamma}_{n,n'}(\mathbf{r}) - (E(\mathbf{k}) - \epsilon_n)\alpha_{n,n'}(\mathbf{r})\Big)$$

The requirement (3.45) that $M_{n,n'}(\mathbf{k})$ has a zero eigenvalue can be equivalently written as

$$\det\, M_{n,n'}(\mathbf{k}) = 0$$

Let's think about how to view this equation. The matrix $M_{n,n'}(\mathbf{k})$ is a function of the various parameters which encode the underlying lattice dynamics as well as $E(\mathbf{k})$. But what we want to figure out is the dispersion relation $E(\mathbf{k})$. We should view the condition $\det M_{n,n'}(\mathbf{k}) = 0$ as an equation for $E(\mathbf{k})$.

Suppose that we include p localised states at each site, so $M_{n,n'}(\mathbf{k})$ is a $p \times p$ matrix. Then $\det M_{n,n'}(\mathbf{k}) = 0$ is a polynomial in $E(\mathbf{k})$ of degree p. This polynomial will have p roots; these are the energies $E_n(\mathbf{k})$ of p bands. In each case, the corresponding null eigenvector is c_n which tells us how the atomic orbitals mix in the Bloch state (3.42).

3.4 Scattering Off a Lattice

Finally, we come to an important question: how do we know that solids are made of lattices? The answer, of course, is scattering. Firing a beam of particles — whether neutrons, electrons or photons in the X-ray spectrum — at the solid reveals a characteristic diffraction pattern. Our goal here is to understand this within the general context of scattering theory that we met in Section 1.

Our starting point is the standard asymptotic expression describing a wave scattering off a central potential, localised around the origin,

$$\psi(\mathbf{r}) \sim e^{i\mathbf{k}\cdot\mathbf{r}} + f(\mathbf{k};\mathbf{k}')\frac{e^{ikr}}{r} \tag{3.46}$$

Here we're using the notation, introduced in earlier sections, of the scattered momentum

$$\mathbf{k}' = k\hat{\mathbf{r}}$$

The idea here is that if you sit far away in the direction $\hat{\mathbf{r}}$, you will effectively see a wave with momentum $\mathbf{k}'$. We therefore write $f(\mathbf{k},\mathbf{k}')$ to mean the same thing as $f(k;\theta,\phi)$.

Suppose now that the wave scatters off a potential which is localised at some other position, $\mathbf{r} = R$. Then the equation (3.46) becomes

$$\psi(\mathbf{r}) \sim e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})} + f(\mathbf{k},\mathbf{k}')\frac{e^{ik|\mathbf{r}-\mathbf{R}|}}{|\mathbf{r}-\mathbf{R}|}$$

For $r \to \infty$, we can expand

$$|\mathbf{r}-\mathbf{R}| = \sqrt{r^2 + R^2 - 2\mathbf{r}\cdot\mathbf{R}} \approx r\sqrt{1 - 2\mathbf{r}\cdot\mathbf{R}/r^2} \approx r - \hat{\mathbf{r}}\cdot\mathbf{R}$$

We then have

$$\psi(\mathbf{r}) \sim e^{-i\mathbf{k}\cdot\mathbf{R}}\left[e^{i\mathbf{k}\cdot\mathbf{r}} + f(\mathbf{k},\mathbf{k}')e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}}\,\frac{e^{ikr}}{r}\right] \tag{3.47}$$

The overall factor is unimportant, since our interest lies in the phase shift between the incident wave and the scattered wave. We see that we get an effective scattering amplitude

$$f_{\mathbf{R}}(\mathbf{k};\hat{\mathbf{r}}) = f(\mathbf{k},\mathbf{k}')\,e^{i\mathbf{q}\cdot\mathbf{R}}$$

where we have defined the transferred momentum

$$\mathbf{q} = \mathbf{k} - \mathbf{k}'$$

Now let's turn to a lattice of points Λ. Ignoring multiple scatterings, the amplitude is simply the sum of the amplitudes from each lattice point

$$f_\Lambda(\mathbf{k},\mathbf{k}') = f(\mathbf{k},\mathbf{k}')\sum_{\mathbf{R}\in\Lambda} e^{i\mathbf{q}\cdot\mathbf{R}} \tag{3.48}$$

However, we already discussed the sum $\Delta(\mathbf{q}) = \sum_{\mathbf{R}\in\Lambda} e^{i\mathbf{q}\cdot\mathbf{R}}$ in Section 3.2.2. The sum has the nice property that it vanishes unless $\mathbf{q}$ lies in the reciprocal lattice $\Lambda^\star$. This is simple to see: since we have an infinite lattice it must be true that, for any vector $\mathbf{R}_0 \in \Lambda$,

$$\Delta(\mathbf{q}) \equiv \sum_{\mathbf{R}\in\Lambda} e^{i\mathbf{q}\cdot\mathbf{R}} = \sum_{\mathbf{R}\in\Lambda} e^{i\mathbf{q}\cdot(\mathbf{R}-\mathbf{R}_0)} = e^{-i\mathbf{q}\cdot\mathbf{R}_0}\Delta(\mathbf{q})$$

This means that either $e^{-i\mathbf{q}\cdot\mathbf{R}_0} = 1$ or $\Delta(\mathbf{q}) = 0$. The former result is equivalent to the statement that $\mathbf{q} \in \Lambda^\star$. More generally,

$$\sum_{\mathbf{R}\in\Lambda} e^{i\mathbf{q}\cdot\mathbf{R}} \equiv \Delta(\mathbf{q}) = V^\star \sum_{\mathbf{Q}\in\Lambda^\star} \delta(\mathbf{q}-\mathbf{Q}) \tag{3.49}$$

where $V^\star$ is the volume of the unit cell of $\Lambda^\star$. We see that $\Delta(\mathbf{q})$ is very strongly (formally, infinitely) peaked on the reciprocal lattice.

The upshot of this discussion is a lovely result: there is scattering from a lattice if and only if

$$\mathbf{k} - \mathbf{k}' \in \Lambda^\star \tag{3.50}$$

This is known as the *Laue condition*. If the scattered momentum does not satisfy this condition, then the interference between all the different scattering sites results in a vanishing wave. Only when the Laue condition is obeyed is this interference constructive.

Alternatively, the Laue condition can be viewed as momentum conservation, with the intuition — garnered from Section 3 — that the lattice can only absorb momentum in $\Lambda^\star$.

Solutions to the Laue condition are not generic. If you take a lattice with a fixed orientation and fire a beam with fixed $\mathbf{k}$, chances are that there are no solutions to (3.50). To see this, consider the reciprocal lattice as shown in the left-hand panel of the figure. From the tip of $\mathbf{k}$ draw a sphere of radius k. This is sometimes known as the *Ewald sphere* and its surface gives the possible transferred momenta $\mathbf{q} = \mathbf{k} - \mathbf{k}'$. There is scattering only if this surface passes through a point on the reciprocal lattice.

To get scattering, we must therefore either find a way to vary the incoming momentum $\mathbf{k}$, or find a way to vary the orientation of the lattice. But when this is achieved, the outgoing photons $\mathbf{k}' = k\hat{\mathbf{r}}$ sit only at very specific positions. In this way, we get to literally take a photograph of the reciprocal lattice! The resulting diffraction pattern for salt ($NaCl$) which has a cubic lattice structure is shown in the right-hand panel. The four-fold symmetry of the reciprocal lattice is clearly visible.

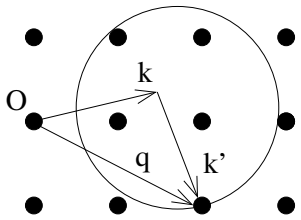


Figure 45: The Ewald sphere, drawn in the reciprocal lattice.

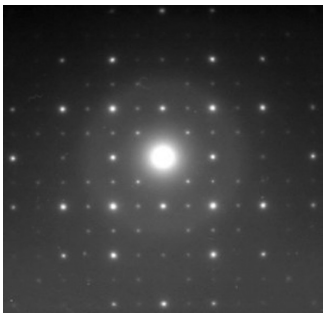


Figure 46: Salt.

3.4.1 The Bragg Condition

There is an equivalent phrasing of the Laue condition in real space. Suppose that the momentum vectors obey

$$\mathbf{k} - \mathbf{k}' = \mathbf{Q} \in \Lambda^\star$$

Since $\mathbf{Q}$ is a lattice vector, so too is $n\mathbf{Q}$ for all $n \in \mathbf{Z}$. Suppose that $\mathbf{Q}$ is minimal, so that $n\mathbf{Q}$ is not a lattice a vector for any $n < 1$. Defining the angle θ by $\mathbf{k} \cdot \mathbf{k}' = k^2 \cos\theta$, we can take the square of the equation above to get

$$2k^2(1 - \cos\theta) = 4k^2 \sin^2(\theta/2) = Q^2 \quad \Rightarrow \quad 2k \sin(\theta/2) = Q$$

We can massage this further. The vector $\mathbf{Q} \in \Lambda^\star$ defines a set of parallel planes in Λ. Known as *Bragg planes*, these are labelled by an integer n and defined by those $\mathbf{a} \in \Lambda$ which obey $\mathbf{a} \cdot \mathbf{Q} = 2\pi n$. The distance between successive planes is

$$d = \frac{2\pi}{Q}$$

Furthermore, the wavevector k corresponds to a wavelength $\lambda = 2\pi/k$. We learn that the Laue condition can be written as the requirement that

$$\lambda = 2d \sin(\theta/2)$$

Repeating this argument for vectors $n\mathbf{Q}$ with $n \in \mathbf{Z}$, we get

$$n\lambda = 2d \sin(\theta/2)$$

Figure 48: A quasi-crystal.

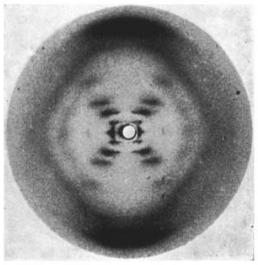


Figure 49: DNA, Photograph 51.

This is the Bragg condition. It has a simple interpretation. For $n = 1$, we assume that the wave scatters off two consecutive planes of the lattice, as shown figure. The wave which hits the lower plane travels an extra distance of $2x = 2d\sin(\theta/2)$. The Bragg condition requires this extra distance to coincide with the wavelength of light. In other words, it is the statement that waves reflecting off consecutive planes interfere constructively.

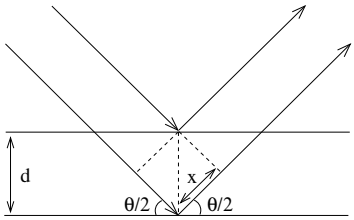


Figure 47:

The Bragg condition gives us licence to think about scattering of light off planes in the lattice, rather than individual lattice sites. Moreover, it tells us that the wavelength of light should be comparable to the atomic separation in the crystal. This means x-rays. The technique of x-ray crystallography was pioneered by Max von Laue, who won the 1914 Nobel prize. The Bragg law was developed by William Bragg, a fellow of Trinity and director of the Cavendish. He shared the 1915 Nobel prize in physics with his father, also William Bragg, for their development of crystallographic techniques.

X-ray crystallography remains the most important technique to determine the structure of materials. Two examples of historical interest are shown in the figures. The picture on the left is something of an enigma since it has five-fold symmetry. Yet there are no Bravais lattices with this symmetry! The diffraction pictures is revealing a *quasi-crystal*, an ordered but non-periodic crystal. The image on the right was taken by Rosalind Franklin and is known as "photograph 51". It provided a major, and somewhat controversial, hint to Crick and Watson in their discovery of the structure of DNA.

3.4.2 The Structure Factor

Many crystals are described by a repeating group of atoms, in which each group sits on an underlying Bravais lattice Λ. The atoms in the group are displaced from the vertex

of the Bravais lattice by a vector $\mathbf{d}_i$. We saw several examples of this in Section 3. In such a situation, the scattering amplitude (3.48) is replaced by

$$f_{\text{lattice}}(\mathbf{k}, \mathbf{k}') = \Delta(\mathbf{q})\, S(\mathbf{q})$$

where

$$S(\mathbf{q}) = \sum_i f_i(\mathbf{k}, \mathbf{k}') e^{i\mathbf{q}\cdot\mathbf{d}_i}$$

We have allowed for the possibility that each atom in the basis has a different scattering amplitude $f_i(\mathbf{k}, \mathbf{k}')$. The function $S(\mathbf{q})$ is called the *geometric structure factor*.

An Example: BCC Lattice

As an example, consider the BCC lattice viewed as a simple cubic lattice of size a, with two basis vectors sitting at $\mathbf{d}_1 = 0$ and $\mathbf{d}_2 = \frac{a}{2}(1, 1, 1)$. If we take the atoms on the points $\mathbf{d}_1$ and $\mathbf{d}_2$ to be identical, then the associated scattering amplitudes are also equal: $f_1 = f_2 = f$.

We know that the scattering amplitude is non-vanishing only if the transferred momentum $\mathbf{q}$ lies on the reciprocal lattice, meaning

$$\mathbf{q} = \frac{2\pi}{a}(n_1, n_2, n_3) \qquad n_i \in \mathbf{Z}$$

This then gives the structure factor

$$\begin{aligned} S(\mathbf{q}) &= f\left(e^{i\mathbf{q}\cdot\mathbf{d}_1} + e^{i\mathbf{q}\cdot\mathbf{d}_2}\right) \\ &= f\left(1 + e^{i\pi\sum_i n_i}\right) = \begin{cases} 2 & \sum n_i \text{ even} \\ 0 & \sum n_i \text{ odd} \end{cases} \end{aligned}$$

We see that not all points in the reciprocal lattice $\Lambda^\star$ contribute. If we draw the reciprocal, simple cubic lattice and delete the odd points, as shown in the right-hand figure, we find ourselves left with a FCC lattice. (Admittedly, the perspective in the figure isn't great.) But this is exactly what we expect since it is the reciprocal of the BCC lattice.

Another Example: Diamond

A diamond lattice consists of two, interlaced FCC lattices with basis vectors $\mathbf{d}_1 = 0$ and $\mathbf{d}_2 = \frac{a}{4}(1, 1, 1)$. An FCC lattice has reciprocal lattice vectors $\mathbf{b}_1 = \frac{2\pi}{a}(-1, 1, 1)$,

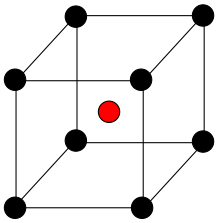


Figure 50: A BCC lattice as cubic lattice + basis.

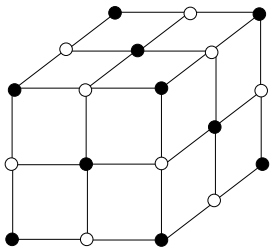


Figure 51: The reciprocal as a cubic lattice minus a basis.

$\mathbf{b}_2 = \frac{2\pi}{a}(1,-1,1)$ and $\mathbf{b}_3 = \frac{2\pi}{a}(1,1,-1)$. For $\mathbf{q} = \sum_i n_i \mathbf{b}_i$, the structure factor is

$$S(\mathbf{q}) = f\left(1 + e^{i(\pi/2)\sum_i n_i}\right) = \begin{cases} 2 & \sum n_i = 0 \text{ mod } 4 \\ 1+i & \sum n_i = 1 \text{ mod } 4 \\ 0 & \sum n_i = 2 \text{ mod } 4 \\ 1-i & \sum n_i = 3 \text{ mod } 4 \end{cases}$$

3.4.3 The Debye-Waller Factor

So far, we've treated the lattice as a fixed, unmoving object. But we know from our discussion in Section 5 that this is not realistic. The underlying atoms can move. We would like to know what effect this has on the scattering off a lattice.

Let's return to our result (3.48) for the scattering amplitude off a Bravais lattice Λ,

$$f_\Lambda(\mathbf{k},\mathbf{k}') = f(\mathbf{k},\mathbf{k}') \sum_n e^{i\mathbf{q}\cdot\mathbf{R}_n}$$

where $f(\mathbf{k},\mathbf{k}')$ is the amplitude for scattering from each site, $\mathbf{q} = \mathbf{k} - \mathbf{k}'$, and $\mathbf{R}_n \in \Lambda$. Since the atoms can move, the position R_n are no longer fixed. We should replace

$$R_n \to R_n + \mathbf{u}_n(t)$$

where, as in Section 5, $\mathbf{u}_n$ describes the deviation of the lattice from equilibrium. In general, this deviation could arise from either thermal effects or quantum effects. In keeping with the theme of these lectures, we will restrict to the latter. But this is conceptually interesting: it means that the scattering amplitude includes the factor

$$\tilde{\Delta}(\mathbf{q}) = \sum_n e^{i\mathbf{q}\cdot\mathbf{R}_n}\, e^{i\mathbf{q}\cdot\mathbf{u}_n}$$

which is now a quantum operator. This is telling us something important. When a particle – whether photon or neutron – scatters off the lattice, it can now excite a phonon mode. The scattering amplitude is a quantum operator because it includes all possible end-states of the lattice.

This opens up a whole slew of new physics. We could, for example, now start to compute *inelastic scattering*, in which the particle deposits some energy in the lattice. Here, however, we will content ourselves with *elastic scattering*, which means that the the lattice sits in its ground state $|0\rangle$ both before and after the scattering. For this, we need to compute

$$\tilde{\Delta}(\mathbf{q}) = \sum_n e^{i\mathbf{q}\cdot\mathbf{R}_n} \langle 0|e^{i\mathbf{q}\cdot\mathbf{u}_n(t)}|0\rangle$$

To proceed, we need the results of Section 5.1.4 in which we treated lattice vibrations quantum mechanically. For simplicity, let's consider a simple cubic lattice so that the the matrix element above factorises into terms in the x, y and z direction. For each of these, we can use the formalism that we developed for the one-dimensional lattice.

The matrix element $\langle 0|e^{i\mathbf{q}\cdot\mathbf{u}_n}|0\rangle$ is independent of time and is also translationally invariant. This means that we can evaluate it at $t = 0$ and at the lattice site $n = 0$. For a one-dimensional lattice with N sites, the expansion (5.11) gives

$$u_0 = \sum_{k\neq 0} \sqrt{\frac{\hbar}{2mN\omega(k)}}\Big(a(k) + a^\dagger(k)\Big) \equiv A + A^\dagger$$

Here we've used the rescaling (5.14) so that the creation and annihilation operators obey the usual commutation relations $[a(k), a^\dagger(k')] = \delta_{k,k'}$. The operators $a^\dagger(k)$ create a phonon with momentum k and energy $\omega(k)$. The operators A and $A^\dagger$ then obey

$$[A, A^\dagger] = \sum_{k\neq 0} \frac{\hbar}{2mN\omega(k)}$$

Our goal now is to compute $\langle 0|e^{iq(A+A^\dagger)}|0\rangle$. For this we use the BCH formula,

$$e^{iq(A+A^\dagger)} = e^{iqA^\dagger}\, e^{iqA}\, e^{\frac{1}{2}q^2[A^\dagger,A]}$$

But the ground state of the lattice is defined to obey $a_l|0\rangle = 0$ for all l. This means that $e^{iqA}|0\rangle = |0\rangle$. We end up with the result

$$\langle 0|e^{i\mathbf{q}\cdot\mathbf{u}_0}|0\rangle = e^{-W(\mathbf{q})} \quad \text{where} \quad W(\mathbf{q}) = \sum_{\mathbf{k}} \frac{\hbar\mathbf{q}^2}{4mN\omega(\mathbf{k})}$$

This is called the *Debye-Waller factor*. We see that the scattering amplitude becomes

$$f_\Lambda(\mathbf{k}, \mathbf{k}') = e^{-W(\mathbf{q})}\, f(\mathbf{k}, \mathbf{k}')\Delta(\mathbf{q})$$

Note that, perhaps surprisingly, the atomic vibrations do not broaden the Bragg peaks away from $\mathbf{q} \in \Lambda^\star$. Instead, they only diminish their intensity.

4. Electron Dynamics in Solids

In the previous chapter we have seen how the single-electron energy states form a band structure in the presence of a lattice. Our goal now is to understand the consequences of this, so that we can start to get a feel for some of the basic properties of materials.

There is one feature in particular that will be important: materials don't just have one electron sitting in them. They have lots. A large part of condensed matter physics is concerned with in understanding the collective behaviour of this swarm of electrons. This can often involve the interactions between electrons giving rise to subtle and surprising effects. However, for our initial foray into this problem, we will make a fairly brutal simplification: we will ignore the interactions between electrons. Ultimately, much of the basic physics that we describe below is unchanged if we turn on interactions, although the reason for this turns out to be rather deep.

4.1 Fermi Surfaces

Even in the absence of any interactions, electrons still are still affected by the presence of others. This is because electrons are fermions, and so subject to the *Pauli exclusion principle*. This is the statement that only one electron can sit in any given state. As we will see below, the Pauli exclusion principle, coupled with the general features of band structure, goes some way towards explaining the main properties of materials.

Free Electrons

As a simple example, suppose that we have no lattice. We take a cubic box, with sides of length L, and throw in some large number of electrons. What is the lowest energy state of this system? Free electrons sit in eigenstates with momentum $\hbar\mathbf{k}$ and energy $E = \hbar^2k^2/2m$. Because we have a system of finite size, momenta are quantised as $k_i = 2\pi n_i/L$. Further, they also carry one of two spin states, $|\uparrow\rangle$ or $|\downarrow\rangle$.

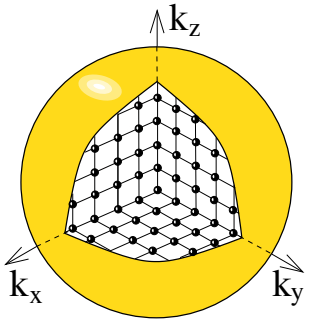


Figure 52: The Fermi surface

The first electron can sit in the state $\mathbf{k} = 0$ with, say, spin $|\uparrow\rangle$. The second electron can also have $\mathbf{k} = 0$, but must have spin $|\downarrow\rangle$, opposite to the first. Neither of these electrons costs any energy. However, the next electron is not so lucky. The minimum energy state it can sit in has $n_i = (1, 0, 0)$. Including spin and momentum there are a total of six electrons which can carry momentum $|\mathbf{k}| = 2\pi/L$. As we go on, we fill out a ball in momentum space. This ball is called the *Fermi sea* and the boundary of the ball is called the *Fermi surface*. The states on the Fermi surface are said to have

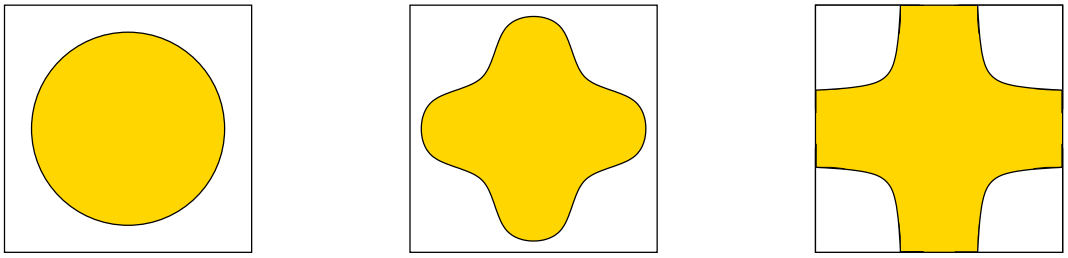


Figure 53: Fermi surfaces for valence $Z = 1$ with increasing lattice strength.

Fermi momentum $\hbar k_F$ and *Fermi energy* $E_F = \hbar^2 k_F^2/2m$. Various properties of the free Fermi sea are explored in the lectures on Statistical Physics.

4.1.1 Metals vs Insulators

Here we would like to understand what becomes of the Fermi sea and, more importantly, the Fermi surface in the presence of a lattice. Let's recapitulate some important facts that we'll need to proceed:

- A lattice causes the energy spectrum to splits into bands. We saw in Section 3.3.2 that a Bravais lattice with N sites results in each band having N momentum states. These are either labelled by momenta in the first Brillouin zone (in the reduced zone scheme) or by momentum in successive Brillouin zones (in the extended zone scheme).

- Because each electron carries one of two spin states, each band can accommodate $2N$ electrons.

- Each atom of the lattice provides an integer number of electrons, Z, which are free to roam the material. These are called *valence electrons* and the atom is said to have *valence* Z.

From this, we can piece the rest of the story together. We'll discuss the situation for two-dimensional square lattices because it's simple to draw the Brillouin zones. But everything we say carries over for more complicated lattices in three-dimensions.

Suppose that our atoms have valence $Z = 1$. There are then N electrons, which can be comfortably housed inside the first Brillouin zone. In the left-hand of Figure 53 we have drawn the Fermi surface for free electrons inside the first Brillouin zone. However, we know that the effect of the lattice is to reduce the energy at the edges of the Brillouin zone. We expect, therefore, that the Fermi surface — which is the

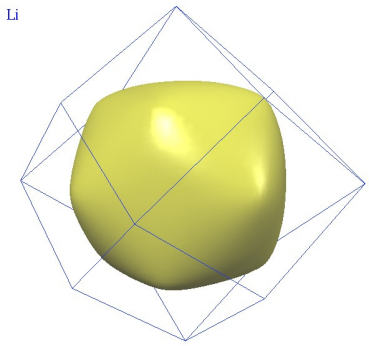


Figure 54: Lithium.

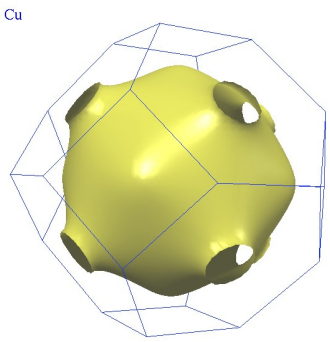


Figure 55: Copper.

equipotential E_F — will be distorted as shown in the middle figure, with states closer to the edge of the Brillouin zone filled preferentially. Note that the area inside the Fermi surface remains the same.

If the effects of the lattice get very strong, it may be that the Fermi surface touches the edge of the Brillouin zone as shown in the right-hand drawing in Figure 53. Because the Brillouin zone is a torus, if the Fermi surface is to be smooth then it must hit the edge of the Brillouin zone at right-angles.

This same physics can be seen in real Fermi surfaces. Lithium has valence $Z = 1$. It forms a BCC lattice, and so the Brillouin zone is FCC. Its Fermi surface is shown above, plotted within its Brillouin zone[6]. Copper also has valency $Z = 1$, with a FCC lattice and hence BCC Brillouin zone. Here the effects of the lattice are somewhat stronger, and the Fermi surface touches the Brillouin zone.

In all of these cases, there are unoccupied states with arbitrarily small energy above E_F. (Strictly speaking, this statement holds only in the limit $L \to \infty$ of an infinitely large lattice.) This means that if we perturb the system in any way, the electrons will easily be able to respond. Note, however, that only those electrons close to the Fermi surface can respond; those that lie deep within the Fermi sea are locked there by the Pauli exclusion principle and require much larger amounts of energy if they wish to escape.

This is an important point, so I'll say it again. In most situations, only those electrons which lie on the Fermi surface can actually do anything. This is why Fermi surfaces play such a crucial role in our understanding of materials.

[6]This, and other pictures of Fermi surfaces, are taken from http://www.phys.ufl.edu/fermisurface/.

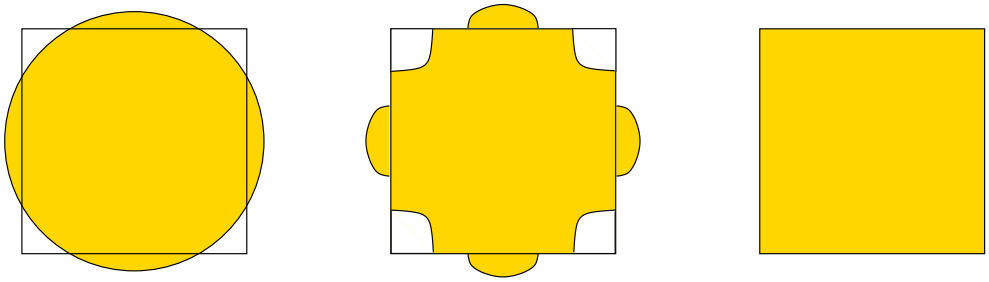


Figure 56: Fermi surfaces for valence $Z = 2$ with increasing lattice strength, moving from a metal to an insulator.

Materials with a Fermi surface are called *metals*. Suppose, for example, that we apply a small electric field to the sample. The electrons that lie at the Fermi surface can move to different available states in order to minimize their energy in the presence of the electric field. This results in a current that flows, the key characteristic of a metal. We'll discuss more about how electrons in lattices respond to outside influences in Section 4.2

Before we move on, a couple of comments:

- The Fermi energy of metals is huge, corresponding to a temperature of $E_F/k_B \sim 10^4\ K$, much higher than the melting temperature. For this reason, the zero temperature analysis is a good starting point for thinking about real materials.
- Metals have a very large number of low-energy excitations, proportional to the area of the Fermi surface. This makes metals a particularly interesting theoretical challenge.

Let's now consider atoms with valency $Z = 2$. These have $2N$ mobile electrons, exactly the right number to fill the first band. However, in the free electron picture, this is not what happens. Instead, they partially fill the first Brillouin zone and then spill over into the second Brillouin zone. The resulting Fermi surface, drawn in the extended zone scheme, is shown in left-hand picture of Figure 56

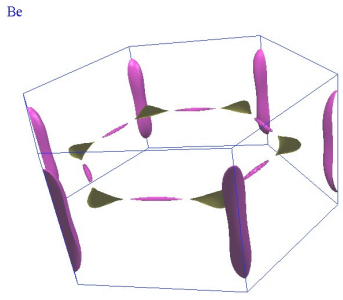


Figure 57: Beryllium

If the effects of the lattice are weak, this will not be greatly changed. Both the first and second Brillouin zones will have available states close to the Fermi surface as shown in the middle picture. These materials remain metals. We sometimes talk

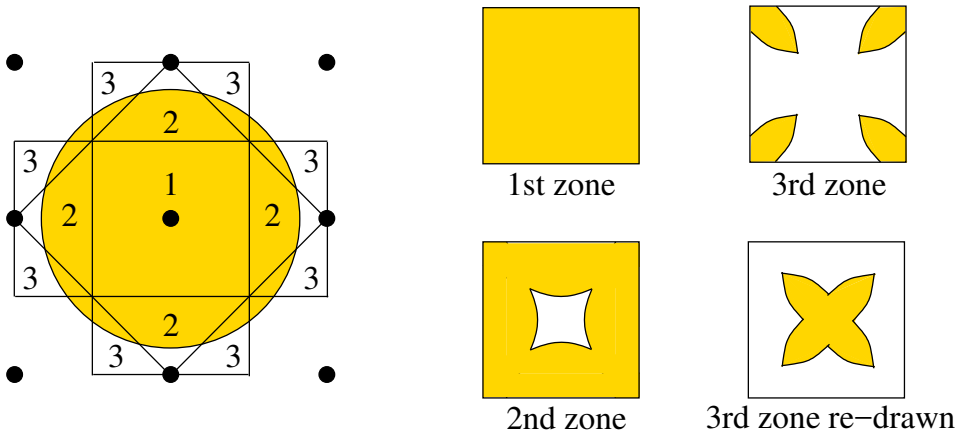


Figure 58: Fermi surfaces for valence $Z = 3$.

of electrons in the second band, and holes (i.e. absence of electrons) in the first band. We will discuss this further in Section 4.2. Beryllium provides an example of a metal with $Z = 2$; its Fermi surface is shown in the figure, now plotted in the reduced zone scheme. It includes both an electron Fermi surface (the cigar-like shapes around the edge) and a hole Fermi surface (the crown in the middle).

Finally, if the effects of the lattice become very strong, the gap between the two bands is large enough to overcome the original difference in kinetic energies. This occurs when the lowest lying state in the second band is higher than the highest state in the first. Now the electrons fill the first band. The second band is empty. The Fermi sea looks like the right-hand picture in Figure 56. This is qualitatively different from previous situations. There is no Fermi surface and, correspondingly, no low-energy excitations. Any electron that wishes to change its state can only do so by jumping to the next band. But that costs a finite amount of energy, equal to the gap between bands. This means that all the electrons are now locked in place and cannot respond to arbitrarily small outside influences. We call such materials *insulators*. (Sometimes they are referred to as *band insulators* to highlight the fact that it is the band structure which prevents the electrons from moving.)

This basic characterisation remains for higher valency Z. Systems with partially filled bands are metals; systems with only fully-filled bands are insulators. Note that a metal may well have several fully-filled bands, before we get to a partially filled band. In such circumstances, we usually differentiate between the fully-filled lower bands — which are called *valence bands* — and the partially filled *conduction band*.

The Fermi surfaces may exist in several different bands. An example of a Fermi surface for $Z = 3$ is shown in Figure 58, the first three Brillouin zones are shown separately in the reduced zone scheme. At first glance, it appears that the Fermi surface in the 3rd Brillouin zone is disconnected. However, we have to remember that the edges of the Brillouin zone are identified. Re-drawn, with the origin taken to be $\mathbf{k} = (\pi/a, \pi/a)$, we see the Fermi surface is connected, taking the rosette shape shown.

Looking Forwards

We have seen how band structure allows us to classify all materials as metals or insulators. This, however, is just the beginning, the first chapter in a long and detailed story which extends from physics into materials science. To whet the appetite, here are three twists that we can add to this basic classification.

- For insulators, the energy required to reach the first excited state is set by the band gap Δ which, in turn, is determined by microscopic considerations. Materials whose band gap is smaller than $\Delta \lesssim 2\ eV$ or so behave as insulators at small temperature, but starts to conduct at higher temperatures as electrons are thermally excited from the valence band to the conduction band. Such materials are called *semiconductors*. They have the property that their conductivity increases as the temperature increases. (This is in contrast to metals whose conductivity decreases as temperature increases.) John Bardeen, Walter Brattain and William Shockley won the 1956 Nobel prize for developing their understanding of semiconductors into a working transistor. This, then, changed the world.

- There are some materials which have $Z = 1$ but are, nonetheless, insulators. An example is nickel oxide NiO. This contradicts our predictions using elementary band structure. The reason is that, for these materials, we cannot ignore the interactions between electrons. Roughly speaking, the repulsive force dominates the physics and effectively prohibits two electrons from sitting on the same site, even if they have different spins. But with only one spin state allowed per site, each band houses only N electrons. Materials with this property are referred to as *Mott insulators*. Nevill Mott, Cavendish professor and master of Caius, won the 1977 Nobel prize, in part for this discovery.

- For a long time band insulators were considered boring. The gap to the first excited state means that they can't do anything when prodded gently. This attitude changed relatively recently when it was realised that you can be boring in different ways. There is a topological classification of how the phase of the quantum states winds as you move around the Brillouin zone. Materials in which

this winding is non-trivial are called *topological insulators*. They have wonderful and surprising properties, most notably on their edges where they come alive with interesting and novel physics. David Thouless and Duncan Haldane won the 2016 Nobel prize for their early, pioneering work on this topic.

More generally, there is a lesson above that holds in a much wider context. Our classification of materials into metals and insulators hinges on whether or not we can excite a multi-electron system with an arbitrarily small cost in energy. For insulators, this is not possible: we require a finite injection of energy to reach the excited states. Such systems are referred to as *gapped*, meaning that there is finite energy gap between the ground state and first excited state. Meanwhile, systems like metals are called *gapless*. Deciding whether any given quantum system is gapped or gapless is one of the most basic questions we can ask. It can also be one of the hardest. For example, the question of whether a quantum system known as *Yang-Mills theory* has a gap is one of the six unsolved milllenium maths problems.

4.1.2 The Discovery of Band Structure

Much of the basic theory of band structure was laid down by Felix Bloch in 1928 as part of his doctoral thesis. As we have seen, Bloch's name is attached to large swathes of the subject. He had an extremely successful career, winning the Nobel prize in 1952, working as the first director-general of CERN, and building the fledgling physics department at Stanford University.

However, Bloch missed the key insight that band structure explains the difference between metals and insulators. This was made by Alan Wilson, a name less well known to physicists. Wilson was a student of Ralph Fowler in Cambridge. In 1931, he took up a research position with Heisenberg and it was here that he made his important breakthrough. He returned on a visit to Cambridge to spread the joy of his newfound discovery, only to find that no one very much cared. At the time, Cambridge was in the thrall of Rutherford and his motto: "There are two kinds of science, physics and stamp collecting". And when Rutherford said "physics", he meant "nuclear physics".

This, from Nevill Mott,

> "I first heard of [Wilson's discovery] when Fowler was explaining it to Charles Ellis, one of Rutherford's closest collaborators, who said 'very interesting' in a tone which implied that he was not interested at all. Neither was I."

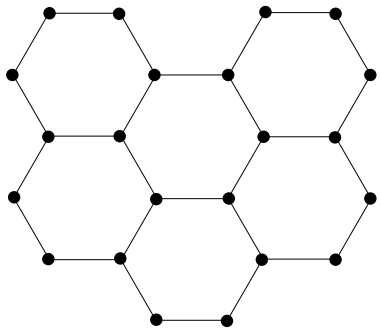


Figure 59: The honeycomb lattice.

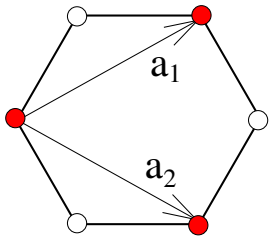


Figure 60: And its basis vectors.

Nevill Mott went on to win the Nobel prize for generalising Wilson's ideas. Wilson himself didn't do so badly either. He left academia and moved to industry, rising to become chairman of Glaxo.

4.1.3 Graphene

Graphene is a two-dimensional lattice of carbon atoms, arranged in a honeycomb structure as shown in the figure. Although it is straightforward to build many layers of these lattices — a substance known as graphite — it was long thought that a purely two-dimensional lattice would be unstable to thermal fluctuations and impossible to create. This changed in 2004 when Andre Geim and Konstantin Novoselov at the University of Manchester succeeded in isolating two-dimensional graphene. For this, they won the 2010 Nobel prize. As we now show, the band structure of graphene is particularly interesting.

First, some basic lattice facts. We described the honeycomb lattice in Section 3.2.1. It is not Bravais. Instead, it is best thought of as two triangular sublattices. We define the primitive lattice vectors

$$\mathbf{a}_1 = \frac{\sqrt{3}a}{2}(\sqrt{3}, 1) \quad \text{and} \quad \mathbf{a}_2 = \frac{\sqrt{3}a}{2}(\sqrt{3}, -1)$$

where a the distance between neighbouring atoms, which in graphene is about $a \approx 1.4 \times 10^{-10}\ m$. These lattice vectors are shown in the figure.

Sublattice A is defined as all the points $\mathbf{r} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2$ with $n_i \in \mathbf{Z}$. These are the red dots in the figure. Sublattice B is defined as all points $\mathbf{r} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + \mathbf{d}$ with $\mathbf{d} = (-a, 0)$. These are the white dots.

The reciprocal lattice is generated by vectors $\mathbf{b}_j$ satisfying $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$. These are

$$\mathbf{b}_1 = \frac{2\pi}{3a}(1, \sqrt{3}) \quad \text{and} \quad \mathbf{b}_2 = \frac{2\pi}{3a}(1, -\sqrt{3})$$

This reciprocal lattice is also triangular, rotated 90° from the original. The Brillouin zone is constructed in the usual manner by drawing perpendicular boundaries between the origin and each other point in the reciprocal lattice. This is shown in the figure. We shortly see that the corners of the Brillouin zone carry particular interest. It naively appears that there are 6 corners, but this should really be viewed as two sets of three. This follows because any points in the Brillouin zone which are connected by a reciprocal lattice vector are identified. Representatives of the two, inequivalent corners of the Brillouin zone are given by

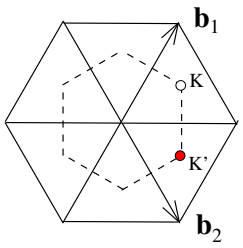


Figure 61:

$$\mathbf{K} = \frac{1}{3}(2\mathbf{b}_1 + \mathbf{b}_2) = \frac{2\pi}{3a}\left(1, \frac{1}{\sqrt{3}}\right) \quad \text{and} \quad \mathbf{K}' = \frac{1}{3}(\mathbf{b}_1 + 2\mathbf{b}_2) = \frac{2\pi}{3a}\left(1, -\frac{1}{\sqrt{3}}\right) \tag{4.1}$$

These are shown in the figure above.

Tight Binding for Graphene

The carbon atoms in graphene have valency $Z = 1$, with the p_z-atomic orbital abandoned by their parent ions and free to roam the lattice. In this context, it is usually called the *π-orbital*. We therefore write down a tight-binding model in which this electron can hop from one atomic site to another. We will work only with nearest neighbour interactions which, for the honeycomb lattice, means that the Hamiltonian admits hopping from a site of the A-lattice to the three nearest neighbours on the B-lattice, and vice versa. The Hamiltonian is given by

$$H = -t\sum_{\mathbf{r}\in\mathbf{\Lambda}} \Big[|\mathbf{r}; A\rangle\langle\mathbf{r}; B| + |\mathbf{r}; A\rangle\langle\mathbf{r} + \mathbf{a}_1; B| + |\mathbf{r}; A\rangle\langle\mathbf{r} + \mathbf{a}_2; B| + \text{h.c.}\Big] \tag{4.2}$$

where we're using the notation

$$|\mathbf{r}; A\rangle = |\mathbf{r}\rangle \quad \text{and} \quad |\mathbf{r}; B\rangle = |\mathbf{r} + \mathbf{d}\rangle \quad \text{with} \quad \mathbf{d} = (-a, 0)$$

Comparing to (3.34), we have set $E_0 = 0$, on the grounds that it doesn't change any of the physics. For what it's worth, $t \approx 2.8\ eV$ in graphene, although we won't need the precise value to get at the key physics.

The energy eigenstates are again plane waves, but now with a suitable mixture of A and B sublattices. We make the ansatz

$$|\psi(\mathbf{k})\rangle = \frac{1}{\sqrt{2N}}\sum_{\mathbf{r}\in\Lambda} e^{i\mathbf{k}\cdot\mathbf{r}}\Big(c_A|\mathbf{r}; A\rangle + c_B|\mathbf{r}; B\rangle\Big)$$

Plugging this into the Schrödinger equation, we find that c_A and c_B must satisfy the eigenvalue equation

$$\begin{pmatrix} 0 & \gamma(\mathbf{k}) \\ \gamma^\star(\mathbf{k}) & 0 \end{pmatrix} \begin{pmatrix} c_A \\ c_B \end{pmatrix} = E(\mathbf{k}) \begin{pmatrix} c_A \\ c_B \end{pmatrix} \tag{4.3}$$

where

$$\gamma(\mathbf{k}) = -t\Big(1 + e^{i\mathbf{k}\cdot\mathbf{a}_1} + e^{i\mathbf{k}\cdot\mathbf{a}_2}\Big)$$

The energy eigenvalues of (4.3) are simply

$$E(\mathbf{k}) = \pm|\gamma(\mathbf{k})|$$

We can write this as

$$E(\mathbf{k})^2 = t^2\Big|1 + e^{i\mathbf{k}\cdot\mathbf{a}_1} + e^{i\mathbf{k}\cdot\mathbf{a}_2}\Big|^2 = t^2\left|1 + 2e^{3ik_xa/2}\cos\left(\frac{\sqrt{3}k_ya}{2}\right)\right|^2$$

Expanding this out, we get the energy eigenvalues

$$E(\mathbf{k}) = \pm t\sqrt{1 + 4\cos\left(\frac{3k_xa}{2}\right)\cos\left(\frac{\sqrt{3}k_ya}{2}\right) + 4\cos^2\left(\frac{\sqrt{3}k_ya}{2}\right)}$$

Note that the energy spectrum is a double cover of the first Brillouin zone, symmetric about $E = 0$. This doubling can be traced to the fact the the honeycomb lattice consists of two intertwined Bravais lattices. Because the carbon atoms have valency $Z = 1$, only the lower band with $E(\mathbf{k}) < 0$ will be filled.

The surprise of graphene is that these two bands meet at special points. These occur on the corners $\mathbf{k} = \mathbf{K}$ and $\mathbf{k} = \mathbf{K}'$ (4.1), where $\cos(3k_xa/2) = -1$ and $\cos(\sqrt{3}k_ya/2) = 1/2$. The resulting band structure is shown in Figure 62[7]. Because the lower band is filled, the Fermi surface in graphene consists of just two points, $\mathbf{K}$ and $\mathbf{K}'$ where the bands meet. It is an example of a *semi-metal*.

Emergent Relativistic Physics

The points $\mathbf{k} = \mathbf{K}$ and $\mathbf{K}'$ where the bands meet are known as *Dirac points*. To see why, we linearise about these points. Write

$$\mathbf{k} = \mathbf{K} + \mathbf{q}$$

[7]The image is taken from the exciting-code website.

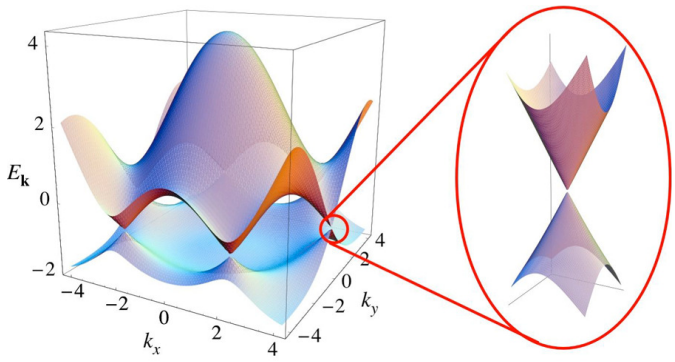


Figure 62: The band structure of graphene.

A little Taylor expansion shows that in the vicinity of the Dirac points, the dispersion relation is linear

$$E(\mathbf{k}) \approx \pm\frac{3ta}{2}|\mathbf{q}|$$

But this is the same kind of energy-momentum relation that we meet in relativistic physics for massless particles! In that case, we have $E = |\mathbf{p}|c$ where p is the momentum and c is the speed of light. For graphene, we have

$$E(\mathbf{k}) \approx \hbar v_F|\mathbf{q}|$$

where $\hbar\mathbf{q}$ is the momentum measured with respect to the Dirac point and $v_F = 3ta/2\hbar$ is the speed at which the excitations propagate. In graphene, v_F is about 300 times smaller than the speed of light. Nonetheless, it remains true that the low-energy excitations of graphene are governed by the same equations that we meet in relativistic quantum field theory. This was part of the reason for the excitement about graphene: we get to test ideas from quantum field theory in a simple desktop experiment.

We can tease out more of the relativistic structure by returning to the Hamiltonian (4.2). Close to the Dirac point $\mathbf{k} = \mathbf{K}$ we have

$$\begin{aligned}
\gamma(\mathbf{k}) &= -t\left(1 - 2e^{3iq_xa/2}\cos\left(\frac{\pi}{3} + \frac{\sqrt{3}q_ya}{2}\right)\right) \\
&= -t\left(1 - 2e^{3iq_xa/2}\left[\frac{1}{2}\cos\left(\frac{\sqrt{3}q_ya}{2}\right) - \frac{\sqrt{3}}{2}\sin\left(\frac{\sqrt{3}q_ya}{2}\right)\right]\right) \\
&\approx -t\left(1 - 2\left(1 + \frac{3iq_xa}{2} + \ldots\right)\left(\frac{1}{2} - \frac{3q_ya}{4} + \ldots\right)\right) \\
&\approx v_F\hbar(iq_x - q_y)
\end{aligned}$$

This means that the Hamiltonian in the vicinity of the Dirac point $\mathbf{k} = \mathbf{K}$ takes the form

$$H = v_F\hbar \begin{pmatrix} 0 & iq_x - q_y \\ -iq_x - q_y & 0 \end{pmatrix} = -v_F\hbar(q_x\sigma^y + q_y\sigma^x) \tag{4.4}$$

where σ^x and σ^y are the Pauli matrices. But this is the *Dirac equation* for a massless particle moving in two-dimensions, sometimes referred to as the *Pauli equation*. (Note: our original choice of orientation of the honeycomb lattice has resulted in a slightly annoying expression for the Hamiltonian. Had we rotated by 90° to begin with, we would be left with the nicer $H = \hbar v_F\, \mathbf{q}\cdot\boldsymbol{\sigma}$ where $\boldsymbol{\sigma} = (\sigma^x, \sigma^y)$.)

There's something of an irony here. In the original Dirac equation, the 2×2 matrix structure comes about because the electron carries spin. But that's not the origin of the matrix structure in (4.4). Indeed, we've not mentioned spin anywhere in our discussion. Instead, in graphene the emergent "spin" degree of freedom arises from the existence of the two A and B sublattices.

We get a very similar equation in the vicinity of the other Dirac point. Expanding $\mathbf{k} = \mathbf{K}' + \mathbf{q}'$, we get the resulting Hamiltonian

$$H = -v_F\hbar(q_x\sigma^y - q_y\sigma^x)$$

The difference in minus sign is sometimes said to be a different *handedness* or *helicity*. You will learn more about this in the context of high energy physics in the lectures on Quantum Field Theory.

As we mentioned above, we have not yet included the spin of the electron. This is trivial: the discussion above is simply repeated twice, once for spin $|\uparrow\rangle$ and once for spin $|\downarrow\rangle$. The upshot is that the low-energy excitations of graphene are described by four massless Dirac fermions. One pair comes from the spin degeneracy of the electrons; the other from the existence of two Dirac points $\mathbf{K}$ and $\mathbf{K}'$, sometimes referred to as the *valley degeneracy*.

4.2 Dynamics of Bloch Electrons

In this section, we look more closely at how electrons moving in a lattice environment react to external forces. We call these electrons *Bloch electrons*. We start by describing how some familiar quantities are redefined for Bloch electrons.

For simplicity, consider an insulator and throw in one further electron. This solitary electron sits all alone in an otherwise unoccupied band. The possible states available to it have energy $E(\mathbf{k})$ where $\mathbf{k}$ lies in the first Brillouin zone. (The energy should also have a further discrete index which labels the particular band the electron is sitting in, but we'll suppress this in what follows). Despite its environment, we can still assign some standard properties to this electron.

4.2.1 Velocity

The average velocity $\mathbf{v}$ of the electron is

$$\mathbf{v} = \frac{1}{\hbar}\frac{\partial E}{\partial \mathbf{k}} \tag{4.5}$$

First note that this is simply the *group velocity* of a wavepacket (a concept that we've met previously in the lectures on Electromagnetism). However, the "average velocity" means something specific in quantum mechanics, and to prove (4.5) we should directly compute $\mathbf{v} = \frac{1}{m}\langle\psi| - i\hbar\nabla|\psi\rangle$.

Bloch's theorem ensures that the electron eigenstates take the form

$$\psi_{\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}\, u_{\mathbf{k}}(\mathbf{x})$$

with $\mathbf{k}$ in the Brillouin zone. As with the energy, we've suppressed the discrete band index on the wavefunction. The full wavefunction satisfies $H\psi_{\mathbf{k}}(\mathbf{x}) = E(\mathbf{k})\psi_{\mathbf{k}}(\mathbf{x})$, so that $u_{\mathbf{k}}(\mathbf{x})$ obeys

$$H_{\mathbf{k}}u_{\mathbf{k}}(\mathbf{x}) = E(\mathbf{k})u_{\mathbf{k}}(\mathbf{x}) \quad \text{with} \quad H_{\mathbf{k}} = \frac{\hbar^2}{2m}(-i\nabla + \mathbf{k})^2 + V(\mathbf{x}) \tag{4.6}$$

We'll use a slick trick. Consider the Hamiltonian $H_{\mathbf{k}+\mathbf{q}}$ which we expand as

$$H_{\mathbf{k}+\mathbf{q}} = H_{\mathbf{k}} + \frac{\partial H_{\mathbf{k}}}{\partial \mathbf{k}}\cdot\mathbf{q} + \frac{1}{2}\frac{\partial^2 H_{\mathbf{k}}}{\partial k^i \partial k^j}\, q^i q^j \tag{4.7}$$

For small $\mathbf{q}$, we view this as a perturbation of $H_{\mathbf{k}}$. From our results of first order perturbation theory, we know that the shift of the energy eigenvalues is

$$\Delta E = \langle u_{\mathbf{k}}|\frac{\partial H_{\mathbf{k}}}{\partial \mathbf{k}}\cdot\mathbf{q}|u_{\mathbf{k}}\rangle$$

But we also know the exact result: it is simply $E(\mathbf{k}+\mathbf{q})$. Expanding this to first order in $\mathbf{q}$, we have the result

$$\langle u_{\mathbf{k}}|\frac{\partial H_{\mathbf{k}}}{\partial \mathbf{k}}|u_{\mathbf{k}}\rangle = \frac{\partial E}{\partial \mathbf{k}}$$

But this is exactly what we need. Using the expression (4.6) for $H_{\mathbf{k}}$, the left-hand side is

$$\frac{\hbar^2}{m}\langle u_{\mathbf{k}}|(-i\nabla + \mathbf{k})|u_{\mathbf{k}}\rangle = \frac{\hbar}{m}\langle \psi_{\mathbf{k}}| - i\hbar\nabla|\psi_{\mathbf{k}}\rangle = \hbar\mathbf{v}$$

This gives our desired result (4.5).

It is perhaps surprising that eigenstates in a crystal have a fixed, average velocity. One might naively expect that the particle would collide with the crystal, bouncing all over the place with a corresponding vanishing average velocity. Yet the beauty of Bloch's theorem is that this is not what happens. The electrons can quite happily glide through the crystal structure.

A Filled Band Carries Neither Current nor Heat

Before we go on, we can use the above result to prove a simple result: a completely filled band does not contribute to the current. This is true whether the filled band is part of an insulator, or part of a metal. (In the latter case, there will also be a partially filled band which will contribute to the current.)

The current carried by each electron is $\mathbf{j} = -e\mathbf{v}$ where $-e$ is the electron charge. From (4.5), the total current of a filled band is then

$$\mathbf{j} = -\frac{2e}{\hbar}\int_{\rm BZ}\frac{d^3k}{(2\pi)^3}\frac{\partial E}{\partial \mathbf{k}} \tag{4.8}$$

where the overall factor of 2 counts the spin degeneracy. This integral vanishes. This follows because $E(\mathbf{k})$ is a periodic function over the Brillouin zone and the total derivative of any periodic function always integrates to zero.

Alternatively, if the crystal has an *inversion symmetry* then there is a more direct proof. The energy satisfies $E(\mathbf{k}) = E(-\mathbf{k})$, which means that $\partial E(\mathbf{k})/\partial\mathbf{k} = -\partial E(-\mathbf{k})/\partial\mathbf{k}$ and the contributions to the integral cancel between the two halves of the Brillouin zone.

The same argument shows that a filled band cannot transport energy in the form of heat. The heat current is defined as

$$\mathbf{j}_E = 2\int_{\rm BZ}\frac{d^3k}{(2\pi)^3}\,E\mathbf{v} = \frac{1}{\hbar}\int_{\rm BZ}\frac{d^3k}{(2\pi)^3}\frac{\partial(E^2)}{\partial\mathbf{k}}$$

which again vanishes when integrated over a filled band. This means that the electrons trapped in insulators can conduct neither electricity nor heat. Note, however, that while there is nothing else charged that can conduct electricity, there are other degrees of freedom – in particular, phonons – which can conduct heat.

4.2.2 The Effective Mass

We define the effective mass tensor to be

$$m^\star_{ij} = \hbar^2 \left(\frac{\partial^2 E}{\partial k^i \partial k^j} \right)^{-1}$$

where we should view the right-hand side as the inverse of a matrix.

For simplicity, we will mostly consider isotropic systems, for which $m^\star_{ij} = m^\star \delta_{ij}$ and the effective mass of the electron is given by

$$m^\star = \hbar^2 \left(\frac{\partial^2 E}{\partial k^2} \right)^{-1} \tag{4.9}$$

where the derivative is now taken in any direction. This definition reduces to something very familiar when the electron sits at the bottom of the band, where we can Taylor expand to find

$$E = E_{\rm min} + \frac{\hbar^2}{2m^\star} |\mathbf{k} - \mathbf{k}_{\rm min}|^2 + \ldots$$

This is the usual dispersion relation or a non-relativistic particle.

The effective mass $m^\star$ has more unusual properties higher up the band. For a typical band structure, $m^\star$ becomes infinite at some point in the middle, and is negative close to the top of the band. We'll see how to interpret this negative effective mass in Section 4.2.4.

In most materials, the effective mass $m^\star$ near the bottom of the band is somewhere between 0.01 and 10 times the actual mass of the electron. But there are exceptions. Near the Dirac point, graphene has an infinite effective mass by the definition (4.9), although this is more because we've used a non-relativistic definition of mass which is rather daft when applied to graphene. More pertinently, there are substances known, appropriately, as *heavy fermion materials* where the effective electron mass is around a 1000 times heavier than the actual mass.

A Microscopic View on the Effective Mass

We can get an explicit expression for the effective mass tensor m_{ij} in terms of the microscopic electron states. This follows by continuing the slick trick we used above, now thinking about the Hamiltonian (4.7) at second order in perturbation theory. This time, we find the inverse mass matrix is given by

$$(m^\star)^{-1}_{ij} = \frac{\delta_{ij}}{m} + \frac{1}{m^2} \sum_{n \neq n'} \frac{\langle \psi_{n,\mathbf{k}} | p_i | \psi_{n',\mathbf{k}} \rangle \langle \psi_{n,\mathbf{k}} | p_j | \psi_{n',\mathbf{k}} \rangle - \text{h.c.}}{E_n(\mathbf{k}) - E_{n'}(\mathbf{k})}$$

where n labels the band of each state. Note that the second term takes the familiar form that arises in second order perturbation theory. We see that, microscopically, the additional contributions to the effective mass come from matrix elements between different bands. Nearby bands of a higher energy give a negative contribution to the effective mass; nearby bands of a lower energy give a positive contribution.

4.2.3 Semi-Classical Equation of Motion

Suppose now that we subject the electron to an external potential force of the form $\mathbf{F} = -\nabla U(\mathbf{x})$. The correct way to proceed is to add $U(\mathbf{x})$ to the Hamiltonian and solve again for the eigenstates. However, in many circumstances, we can work semi-classically. For this, we need that $U(\mathbf{x})$ is small enough that it does not distort the band structure and, moreover, does not vary greatly over distances comparable to the lattice spacing.

We continue to restrict attention to the electron lying in a single band. To proceed, we should think in terms of wavepackets, rather than plane waves. This means that the electron has some localised momentum $\mathbf{k}$ and some localised position $\mathbf{x}$, within the bounds allowed by the Heisenberg uncertainty relation. We then treat this wavepacket as if it was a classical particle, where the position $\mathbf{x}$ and momentum $\hbar\mathbf{k}$ depend on time. This is sometimes referred to as a *semi-classical* approach.

The total energy of this semi-classical particle is $E(\mathbf{k})+U(\mathbf{x})$ where $E(\mathbf{k})$ is the band energy. The position and momentum evolve such that the total energy is conserved. This gives

$$\frac{d}{dt}\Big(E(\mathbf{k}(t)) + U(\mathbf{x}(t))\Big) = \frac{\partial E}{\partial \mathbf{k}} \cdot \frac{d\mathbf{k}}{dt} + \nabla U \cdot \frac{d\mathbf{x}}{dt} = \mathbf{v} \cdot \left(\hbar \frac{d\mathbf{k}}{dt} + \nabla U\right) = 0$$

which is satisfied when

$$\hbar \frac{d\mathbf{k}}{dt} = -\nabla U = \mathbf{F} \tag{4.10}$$

This should be viewed as a variant of Newton's equation, now adapted to the lattice environment. In fact, we can make it look even more similar to Newton's equation. For an isotropic system, the effective "mass times acceleration" is

$$m^\star \frac{d\mathbf{v}}{dt} = \frac{m^\star}{\hbar}\frac{d}{dt}\left(\frac{\partial E}{\partial \mathbf{k}}\right) = \frac{m^\star}{\hbar}\left(\frac{d\mathbf{k}}{dt}\cdot\frac{\partial}{\partial \mathbf{k}}\right)\frac{\partial E}{\partial \mathbf{k}} = \hbar\frac{d\mathbf{k}}{dt} = \mathbf{F} \tag{4.11}$$

where you might want to use index notation to convince yourself of the step in the middle where we lost the effective mass $m^\star$. It's rather nice that, despite the complications of the lattice, we still get to use some old equations that we know and love. Of course, the key to this was really the definition (4.9) of what we mean by effective mass $m^\star$.

An Example: Bloch Oscillations

Consider a Bloch electron, exposed to a constant electric field $\boldsymbol{\mathcal{E}}$. The semi-classical equation of motion is

$$\hbar\dot{\mathbf{k}} = -e\boldsymbol{\mathcal{E}} \quad \Rightarrow \quad k(t) = k(0) - \frac{e\boldsymbol{\mathcal{E}}}{\hbar}t$$

So the crystal momentum $\mathbf{k}$ increases linearly. At first glance, this is unsurprising. But it leads to a rather surprising effect. This is because $\mathbf{k}$ is really periodic, valued in the Brillouin zone. Like a character in a 1980s video game, when the electron leaves one edge of the Brillouin zone, it reappears on the other side.

We can see what this means in terms of velocity. For a typical one-dimensional band structure shown on the right, the velocity $\mathbf{v} \sim \mathbf{k}$ in the middle of the band, but $\mathbf{v} \sim -\mathbf{k}$ as the particle approaches the edge of the Brillouin zone. In other words, a constant electric field gives rise to an oscillating velocity, and hence an oscillating current! This surprising effect is called *Bloch oscillations*.

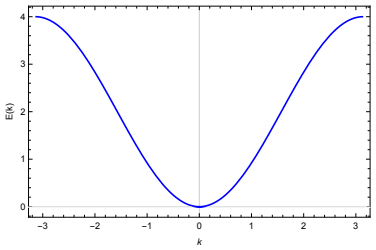


Figure 63:

As an example, consider a one-dimensional system with a tight-binding form of band structure

$$E = -C\cos(ka)$$

Then the velocity in a constant electric field oscillates as

$$v(k) = \frac{Ca}{\hbar}\sin(ka) = -\frac{Ca}{\hbar}\sin\left(\frac{e\mathcal{E}a}{\hbar}t\right)$$

The Bloch frequency is $\omega = e\mathcal{E}a/\hbar$. If we construct a wavepacket from several different energy eigenstates, then the position of the particle will similarly oscillate back and forth. This effect was first predicted by Leo Esaki in 1970.

Bloch oscillations are somewhat counterintuitive. They mean that a DC electric field applied to a pure crystal does *not* lead to a DC current! Yet we've all done experiments in school where we measure the DC current in a metal! This only arises because a metal is not a perfect crystal and the electrons are scattered by impurities or thermal lattice vibrations (phonons) which destroy the coherency of Bloch oscillations and lead to a current.

Bloch oscillations are delicate. The system must be extremely clean so that the particle does not collide with anything else over the time necessary to see the oscillations. This is too much to ask in solid state crystals. However, Bloch oscillations have been observed in other contexts, such as cold atoms in an artificial lattice. The time variation of the velocity of Caesium atoms in an optical lattice is shown in the figure[8].

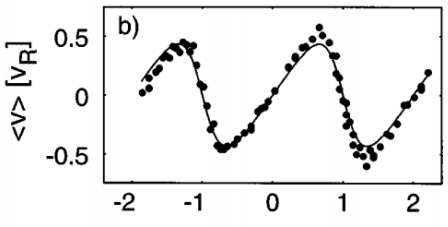


Figure 64:

4.2.4 Holes

Consider a totally filled band, and remove one electron. We're left with a vacancy in the otherwise filled band. In a zen-like manoeuvre, we ascribe properties to the absence of the particle. Indeed, as we will now see, this vacancy moves as if it were itself an independent particle. We call this particle a *hole*.

Recall that our definition (4.9) means that the effective mass of electrons is negative near the top of the band. Indeed, expanding around the maximum, the dispersion relation for electrons reads

$$E(\mathbf{k}) = E_{\rm max} + \frac{\hbar^2}{2m^\star}|\mathbf{k} - \mathbf{k}_{\rm max}|^2 + \ldots$$

and the negative effective mass $m^\star < 0$ ensures that electrons have less energy as the move away from the maximum.

Now consider filling all states except one. As the hole moves away from the maximum, it costs more energy (because we're subtracting less energy!). This suggests that we should write the energy of the hole as

$$E_{\rm hole}(\mathbf{k}) = -E(\mathbf{k}) = -E_{\rm max} + \frac{\hbar^2}{2m^\star_{\rm hole}}|\mathbf{k} - \mathbf{k}_{\rm max}|^2 + \ldots$$

where

$$m^\star_{\rm hole} = -m^\star$$

so that the effective mass of the hole is positive near the top of the band, but becomes negative if the hole makes it all the way down to the bottom.

[8]This data is taken from "*Bloch Oscillations of Atoms in an Optical Potential*" by Dahan et. al., Phys. Rev. Lett. vol 76 (1996), which reported the first observation of this effect.

The hole has other properties. Suppose that we take away an electron with momentum $\mathbf{k}$. Then the resulting hole can be thought of as having momentum $-\mathbf{k}$. This suggests that we define

$$\mathbf{k}_{\rm hole} = -\mathbf{k} \tag{4.12}$$

However, the velocity of the hole is the same as that of the missing electron

$$\mathbf{v}_{\rm hole} = \frac{1}{\hbar}\frac{\partial E_{\rm hole}}{\partial \mathbf{k}_{\rm hole}} = \frac{1}{\hbar}\frac{\partial E}{\partial \mathbf{k}} = \mathbf{v}$$

This too is intuitive, since the hole is moving in the same direction as the electron that we took away.

The definitions above mean that the hole obeys the Newtonian force law with

$$m^\star_{\rm hole}\frac{d\mathbf{v}_{\rm hole}}{dt} = -\mathbf{F} = \mathbf{F}_{\rm hole} \tag{4.13}$$

At first sight, this is surprising: the hole experiences an opposite force to the electron. But there's a very simple interpretation. The force that we typically wish to apply to our system is an electric field $\boldsymbol{\mathcal{E}}$ which, for an electron, gives rise to

$$\mathbf{F} = -e\boldsymbol{\mathcal{E}}$$

The minus sign in (4.13) is simply telling us that the hole should be thought of as carrying charge $+e$, the opposite of the electron,

$$\mathbf{F}_{\rm hole} = +e\boldsymbol{\mathcal{E}}$$

We can also reach this same conclusion by computing the current. We saw in (4.8) that a fully filled band carries no current. This means that the current carried by a partially filled band is

$$\mathbf{j} = -2e\int_{\rm filled}\frac{d^3k}{(2\pi)^3}\,\mathbf{v}(\mathbf{k}) = +2e\int_{\rm unfilled}\frac{d^3k}{(2\pi)^3}\,\mathbf{v}(\mathbf{k})$$

The filled states are electrons carrying charge $-e$; the unfilled states are holes, carrying charge $+e$.

Finally, it's worth mentioning that the idea of holes in band structure provides a fairly decent analogy for anti-matter in high-energy physics. There too the electron has a positively charged cousin, now called the positron. In both cases, the two particles can come together and annihilate. In solids, this releases a few eV of energy, given by the gap between bands. In high-energy physics, this releases a million times more energy, given by the rest mass of the electron.

4.2.5 Drude Model Again

The essence of Bloch's theorem is that electrons can travel through perfect crystals unimpeded. And yet, in the real world, this does not happen. Even the best metals have a resistance, in which any current degrades and ultimately relaxes to zero. This happens because metals are not perfect crystals, and the electrons collide with impurities and vacancies, as well as thermally vibrations called phonons.

We can model these effects in our semi-classical description by working with the electron equation of motion called the *Drude model*

$$m^\star \dot{\mathbf{v}} = -e\boldsymbol{\mathcal{E}} - \frac{m^\star}{\tau}\mathbf{v} \tag{4.14}$$

Here $\boldsymbol{\mathcal{E}}$ is the applied electric field and τ is the *scattering time*, which should be thought of as the average time between collisions.

We have already met the Drude model in the lectures on Electromagnetism when we tried to describe the conductivity in metals classically. We have now included the quantum effects of lattices and the Fermi surface yet, rather remarkably, the equation remains essentially unchanged. The only difference is that the effective mass $m^\star$ will depend on $\mathbf{k}$, and hence on $\mathbf{v}$, if the electron is not close to the minimum of the band.

In equilibrium, the velocity of the electron is

$$\mathbf{v} = -\frac{e\tau}{m^\star}\boldsymbol{\mathcal{E}} \tag{4.15}$$

The proportionality constant is called the *mobility*, $\mu = |e\tau/m^\star|$. The total current density $\mathbf{j} = -en\mathbf{v}$ where n is the density of charge carriers. The equation (4.15) then becomes $\mathbf{j} = \sigma\boldsymbol{\mathcal{E}}$ where σ is the conductivity,

$$\sigma = \frac{e^2\tau n}{m^\star} \tag{4.16}$$

We also define the resistivity $\rho = 1/\sigma$. This is the same result that we found in our earlier classical analysis, except the mass m is replaced by the effective mass $m^\star$.

There is, however, one crucial difference that the existence of the Fermi surface has introduced. When bands are mostly unfilled, it is best to think of the charge carriers in terms of negatively charged electrons, with positive effective mass $m^\star$. But when bands are mostly filled, it is best to think of the charge carriers in terms of positively charged holes, also with positive mass $m^\star_{\rm hole}$. In this case, we should replace the Drude model (4.14) with the equivalent version for holes,

$$m^\star_{\rm hole}\dot{\mathbf{v}} = +e\boldsymbol{\mathcal{E}} - \frac{m^\star_{\rm hole}}{\tau}\mathbf{v} \tag{4.17}$$

This means that certain materials can appear to have positive charge carriers, even though the only things actually moving are electrons. The different sign in the charge carrier doesn't show up in the conductivity (4.16), which depends on e^2. To see it, we need to throw in an extra ingredient.

Hall Resistivity

The standard technique to measure the charge of a material is to apply a magnetic field $\mathbf{B}$. Classically, particles of opposite charges will bend in a opposite directions, perpendicular to $\mathbf{B}$. In a material, this results in the classical *Hall effect*.

We will discuss the motion of Bloch electrons in a magnetic field in much more detail in Section 4.3. (And we will discuss the Hall effect in much much more detail in other lectures.) Here, we simply want to show how this effect reveals the difference between electrons and holes. For electrons, we adapt the Drude model (4.14) by adding a Lorentz force,

$$m^\star \dot{\mathbf{v}} = -e(\boldsymbol{\mathcal{E}} + \mathbf{v} \times \mathbf{B}) - \frac{m^\star}{\tau}\mathbf{v}$$

We once again look for equilibrium solutions with $\dot{\mathbf{v}} = 0$. Writing $\mathbf{j} = -ne\mathbf{v}$, we now must solve the vector equation

$$\frac{1}{ne}\mathbf{j} \times \mathbf{B} + \frac{m^\star}{ne^2\tau}\mathbf{j} = \boldsymbol{\mathcal{E}}$$

The solution to this is

$$\boldsymbol{\mathcal{E}} = \rho \mathbf{j}$$

where the resistivity ρ is now a 3×3 matrix. If we take $\mathbf{B} = (0, 0, B)$, then we have

$$\rho = \begin{pmatrix} \rho_{xx} & \rho_{xy} & 0 \\ -\rho_{xy} & \rho_{xx} & 0 \\ 0 & 0 & \rho_{xx} \end{pmatrix}$$

where the diagonal, *longitudinal resistivity* is $\rho_{xx} = 1/\sigma$ where σ is given in (4.16). The novelty is the off-diagonal, Hall resistivity

$$\rho_{xy} = \frac{B}{ne}$$

We often define the *Hall coefficient* R_H as

$$R_H = \frac{\rho_{xy}}{B} = \frac{1}{ne}$$

This, as promised, depends on the charge e. This means that if we were to repeat the above analysis for holes (4.17) rather than electrons, we would find a Hall coefficient which differs by a minus sign.

There are metals – such as beryllium and magnesium – whose Hall coefficient has the "wrong sign". We drew the Fermi surface for beryllium in Section 4.1.1; it contains both electrons and holes. In this case, we should add to two contributions with opposite signs. It turns out that the holes are the dominant charge carrier.

4.3 Bloch Electrons in a Magnetic Field

In this section, we continue our study of Bloch electrons, but now subjected to an external magnetic field $\mathbf{B}$. (Note that what we call $\mathbf{B}$ should really be called $\mathbf{H}$; it is the magnetising field, after taking into account any bound currents.) Magnetic fields play a particularly important role in solids because, as we shall see, they allow us to map out the Fermi surface.

4.3.1 Semi-Classical Motion

We again use our semi-classical equation of motion (4.10) for the electron, now with the Lorentz force law

$$\hbar \frac{d\mathbf{k}}{dt} = -e\mathbf{v} \times \mathbf{B} \tag{4.18}$$

where the velocity and momentum are once again related by

$$\mathbf{v} = \frac{1}{\hbar}\frac{\partial E}{\partial \mathbf{k}} \tag{4.19}$$

From these two equations, we learn two facts. First, the component of $\mathbf{k}$ parallel to $\mathbf{B}$ is constant: $\frac{d}{dt}(\mathbf{k}\cdot\mathbf{B}) = 0$. Second, the electron traces out a path of constant energy in $\mathbf{k}$-space. This is because

$$\frac{dE}{dt} = \frac{\partial E}{\partial \mathbf{k}} \cdot \frac{\partial \mathbf{k}}{\partial t} = -e\mathbf{v}\cdot(\mathbf{v}\times\mathbf{B}) = 0$$

These two facts are sufficient for us to draw the orbit in k-space. The Fermi surface is, by definition, a surface of constant energy. The electrons orbit the surface, perpendicular to $\mathbf{B}$. It's pictured on the right for a spherical Fermi surface, corresponding to free electrons.

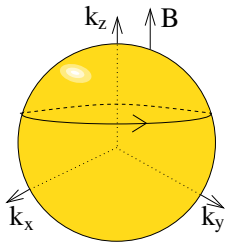


Figure 65:

Holes have an opposite electric charge, and so traverse the Fermi surface in the opposite direction. However, we have to also remember that we call $\mathbf{k}_{\text{hole}}$ also has a relative minus sign (4.12). As an example, consider a metal with $Z = 2$, which has both electron and hole Fermi surfaces. In Figure 66, we have drawn the Fermi surfaces of holes (in purple) and electrons (in yellow) in the extended zone scheme, and shown their direction of propagation in a magnetic field.

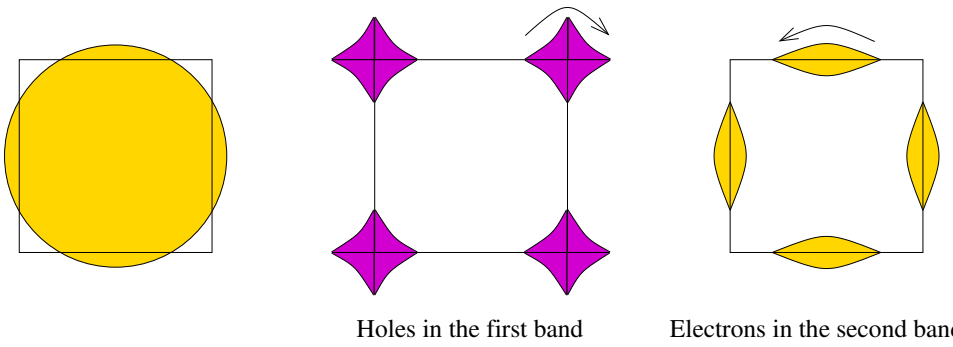


Figure 66: Pockets of electrons and holes for free electrons with $Z = 2$.

Orbits in Real Space

We can also look at the path $\mathbf{r}(t)$ that these orbits trace out in real space. Consider

$$\hat{\mathbf{B}} \times \hbar\dot{\mathbf{k}} = -e\hat{\mathbf{B}} \times (\dot{\mathbf{r}} \times \mathbf{B}) = -eB\,\dot{\mathbf{r}}_\perp \tag{4.20}$$

where $\mathbf{r}_\perp$ is the position of the electron, projected onto a plane perpendicular to $\mathbf{B}$,

$$\mathbf{r}_\perp = \mathbf{r} - (\hat{\mathbf{B}} \cdot \mathbf{r})\,\hat{\mathbf{B}}$$

Integrating (4.20), we find

$$\mathbf{r}_\perp(t) = \mathbf{r}_\perp(0) - \frac{\hbar}{eB}\hat{\mathbf{B}} \times \big(\mathbf{k}(t) - \mathbf{k}(0)\big) \tag{4.21}$$

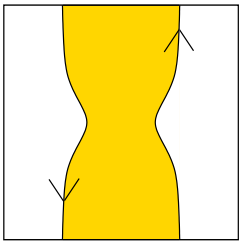


Figure 67:

In other words, the the particle follows the same shape trajectory as in $\mathbf{k}$-space, but rotated about $\mathbf{B}$ and scaled by the magnetic length $l_B^2 = \hbar/eB$. For free electrons, with a spherical Fermi surface, this reproduces the classical result that electrons move in circles. However, as the Fermi surface becomes distorted by band effects this need no longer be the case, and the orbits in real space are no longer circles. For example, the electrons trace out the rosette-like shape in the $Z = 3$ Fermi surface that we saw in Figure 58. In extreme cases its possible for the real space orbits to not be closed curves at all. This happens, for example, if the Fermi surface is distorted more in one direction than another, so it looks like the picture on the right, with electrons performing a loop in the Brillouin zone. These are called *open Fermi surfaces*.

4.3.2 Cyclotron Frequency

Let's compute the time taken for the electron to complete a closed orbit in $\mathbf{k}$-space. The time taken to travel between two points on the orbit $\mathbf{k}_1 = \mathbf{k}(t_1)$ and $\mathbf{k}_2 = \mathbf{k}(t_2)$ is given by the line integral

$$t_2 - t_1 = \int_{\mathbf{k}_1}^{\mathbf{k}_2} \frac{d\mathbf{k}}{|\dot{\mathbf{k}}|}$$

We can use (4.20) to relate $|\dot{\mathbf{k}}|$ to the perpendicular velocity,

$$|\dot{\mathbf{k}}| = \frac{eB}{\hbar}|\dot{\mathbf{r}}_\perp| = \frac{eB}{\hbar^2}\left|\left(\frac{\partial E}{\partial \mathbf{k}}\right)_\perp\right|$$

so we have

$$t_2 - t_1 = \frac{\hbar^2}{eB}\int_{\mathbf{k}_1}^{\mathbf{k}_2} \frac{d\mathbf{k}}{\left|\left(\frac{\partial E}{\partial \mathbf{k}}\right)_\perp\right|}$$

This has a rather nice geometric interpretation. Consider two orbits, both lying in the same plane perpendicular to $\mathbf{B}$, but with the second having a slightly higher Fermi energy $E + \Delta E$. To achieve this, the orbit must sit slightly outside the first, with momentum

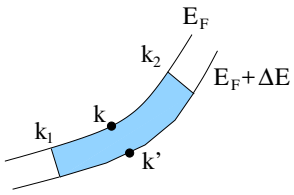


Figure 68:

$$\mathbf{k}' = \mathbf{k} + \left(\frac{\partial E}{\partial \mathbf{k}}\right)_\perp \Delta(\mathbf{k})$$

where, as the notation suggests, $\Delta(\mathbf{k})$, can change as we move around the orbit. We require that $\Delta(\mathbf{k})$ is such that the second orbit also has constant energy,

$$\Delta E = \left|\left(\frac{\partial E}{\partial \mathbf{k}}\right)_\perp\right| \Delta(\mathbf{k})$$

The time taken to traverse the orbit can then be written as

$$t_2 - t_1 = \frac{\hbar^2}{eB}\frac{1}{\Delta E}\int_{\mathbf{k}_1}^{\mathbf{k}_2} \Delta(\mathbf{k})\; d\mathbf{k}$$

But this is simply the area of the strip that separates the two orbits; this area, which we call A_{12}, is coloured in the figure. In the limit $\Delta E \to 0$, we have

$$t_2 - t_1 = \frac{\hbar^2}{eB}\frac{\partial A_{12}}{\partial E}$$

We can now apply this formula to compute the time taken to complete a closed orbit. Let $A(E)$ denote the area enclosed by the orbit. (Note that this will depend not only on E but also on the component of the momentum $\mathbf{k}\cdot\mathbf{B}$ parallel to the magnetic field.) The time taken to complete an orbit is

$$T = \frac{\hbar^2}{eB}\frac{\partial A(E)}{\partial E}$$

The *cyclotron frequency* is defined as

$$\omega_c = \frac{2\pi}{T} \tag{4.22}$$

One can check that the cyclotron frequency agrees with the usual result, $\omega_B = eB/m$ for free electrons.

The fact that the cyclotron frequency ω_c depends on some property of the Fermi surface – namely $\partial A/\partial E$ – is important because the cyclotron frequency is something that can be measured in experiments, since the electrons sit at resonance to absorb microwaves tuned to the same frequency. This gives us our first hint as to how we might measure properties of the Fermi surface.

4.3.3 Onsager-Bohr-Sommerfeld Quantisation

The combination of magnetic fields and Fermi surfaces gives rise to a host of further physics but to see this we will have to work a little harder.

The heart of the problem is that, in classical physics, the Lorentz force does no work. In the Hamiltonian formalism, this translates into the statement that the energy does not depend on $\mathbf{B}$ when written in terms of the canonical momenta. Whenever the energetics of a system depend on the magnetic field, there must be some quantum mechanics going on underneath. In the present case, this means that we need to go slightly beyond the simple semi-classical description that we've met above, to find some of the discreteness that quantum mechanics introduces into the problem.

(As an aside: this problem is embodied in the Bohr-van-Leeuwen theorem, which states that there can be no classical magnetism. We describe how quantum mechanics can circumvent this in the discussion of Landau diamagnetism in the lectures on Statistical Physics.)

To proceed, we would ideally like to quantise electrons in the presence of both a lattice and a magnetic field. This is hard. We've learned how to quantise in the presence of a magnetic field in Section 6 and in the presence of lattice in Section 3, but including both turns out to be a much more difficult problem. Nonetheless, as we now show, there's a way to cobble together an approximation solution.

This cobbled-together quantisation was first proposed by Onsager, but follows an earlier pre-quantum quantisation of Bohr and Sommerfield which suggests that, in any system, an approximation to the quantisation of energy levels can be found by setting

$$\frac{1}{2\pi}\oint \mathbf{p}\cdot d\mathbf{r} = \hbar(n+\gamma) \tag{4.23}$$

with $n \in \mathbf{Z}$ and γ an arbitrary constant. This Bohr-Sommerfeld quantisation does not, in general, agree with the exact result from solving the Schrödinger equation. However, it tends to capture the correct physics for large n, where the system goes over to its semi-classical description.

In the present context, we apply Bohr-Sommerfeld quantisation to our semi-classical model (4.18) and (4.19). We have

$$\frac{1}{2\pi}\oint \mathbf{p}\cdot d\mathbf{r} = \frac{\hbar}{2\pi}\oint \mathbf{k}\cdot d\mathbf{r} = \frac{\hbar^2}{2\pi eB}\oint \mathbf{k}\cdot(d\mathbf{k}\times\hat{\mathbf{B}})$$

where, in the last equality, we have used our result (4.20). But this integral simply captures the cross-sectional area of the orbit in k-space. This is the area $A(E)$ that we met above. We learn that the Bohr-Sommerfeld quantisation condition (4.23) leads to a quantisation of the cross-sectional areas of the Fermi surface in the presence of a magnetic field,

$$A_n = \frac{2\pi eB}{\hbar}(n+\gamma) \tag{4.24}$$

This quantisation of area is actually a variant of the Landau level quantisation that we met in Section 6.2. There are different ways of seeing this. First, note that, for fixed k_z, we can write the cyclotron frequency (4.22) as the difference between consecutive energy levels

$$\omega_c = \frac{2\pi eB}{\hbar^2}\frac{E_{n+1}-E_n}{A_{n+1}-A_n} = \frac{E_{n+1}-E_n}{\hbar}$$

Rearranging, this gives

$$E_n = \hbar\omega_c(n+\text{constant})$$

which coincides with our Landau level spectrum (6.14), except that the old cyclotron frequency $\omega_B = eB/m$ has been replaced by ω_c.

Alternatively, we could look at the quantisation of area in real space, rather than in **k**-space. We saw in (4.21), that the orbit in real space has the same shape as that in **k**-space, but is scaled by a factor of $l_B^2 = \hbar/eB$. This means that the flux through any such orbit is given by

$$\Phi_n = \left(\frac{\hbar}{eB}\right)^2 BA_n = (n+\gamma)\Phi_0 \tag{4.25}$$

where $\Phi_0 = 2\pi\hbar/e$ is the so-called *quantum of flux*. But this ties in nicely with our discussion in Section 6.2 of Landau levels in the absence of a lattice, where we saw that the degeneracy of states in each level is (6.17)

$$\mathcal{N} = \frac{\Phi}{\Phi_0}$$

which should clearly be an integer.

The quantisation (4.24) due to a background magnetic field results in a re-arrangement of the Fermi surface, which now sit in *Landau tubes* whose areas are quantised. A typical example is shown on the right.

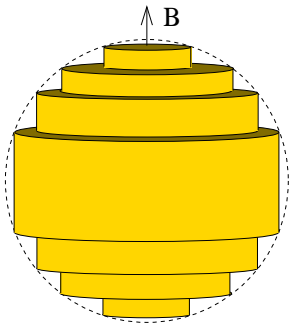


Figure 69:

4.3.4 Quantum Oscillations

The formation of Landau tubes gives rise to a number of fairly striking experimental signatures.

Consider a Fermi surface with energy E_F and a second surface slightly inside with energy $E_F - dE$. The region between these contains the accessible states if we probe the system with a small amount of energy dE. Now consider a Landau tube of cross-sectional area A_n, intersecting our Fermi surface. Typically, the Landau tube will intersect the Fermi surface only in some small region, as shown in left-hand picture of Figure 70. This means that the number of states that can contribute to physical processes will be fairly small. In the language that we introduced in the Statistical Physics lectures, the density of states $g(E_F)dE$ within this Landau tube will be small.

However, something special happens if the area A_n happens to coincide with an extremal area of the Fermi surface. Because the Fermi surface curves much more slowly at such points, the density of states $g(E_F)dE$ is greatly enhanced at this point. This is shown in the right-hand picture of Figure 70. In fact, one can show that the density of states actually diverges at this point as $g(E) \sim (E - E_\star)^{-1/2}$.

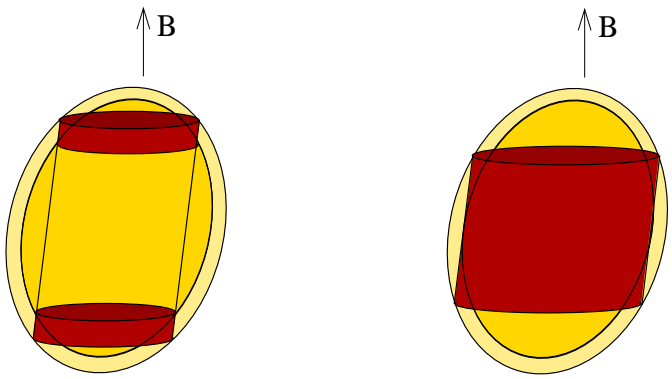


Figure 70: Landau tubes intersecting the Fermi surface: when the area of the tube coincides with an extremal cross-section of the Fermi surface, there is a large enhancement in the available states.

We learn that when the area quantisation takes special values, there are many more electrons that can contribute to any physical process. However, the area quantisation condition (4.24) changes with the magnetic field. This means that as we increase the magnetic field, the areas of Landau tubes will increase and will, occasionally, overlap with an extremal area in the Fermi surface. Indeed, if we denote the extremal cross-sectional area of the Fermi surface as $A_{\rm ext}$, we must get an enhancement in the density of available states whenever

$$A_n = \frac{2\pi eB}{\hbar}(n+\gamma) = A_{\rm ext}$$

for some n. We don't know what γ is, but this doesn't matter: the density of states should occur over and over again, at intervals given by

$$\Delta\left(\frac{1}{B}\right) = \frac{2\pi e}{\hbar}\frac{1}{A_{\rm ext}}$$

Such oscillations are seen in a wide variety of physical measurements and go by the collective name of *quantum oscillations*.

The first, and most prominent example of quantum oscillation is the *de Haas-van Alphen effect*, in which the magnetisation $M = -\partial F/\partial B$ varies with magnetic field. The experimental data for gold is shown in the Figure[9] 71. Note that there are two oscillation frequencies visible in the data. The Fermi surface of gold is shown on the

[9]The data is taken from I.M.Templeton, Proceedings of the Royal Society A, vol 292 (1965). Note the old school graph paper.

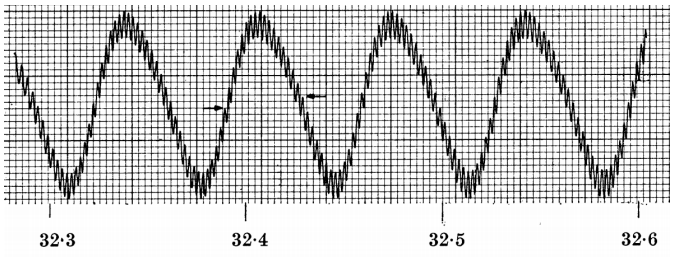


Figure 71: dHvA oscillations for gold. The horizontal axis is B, plotted in kG.

right. For the oscillations above, the magnetic field is parallel to the neck of the Fermi surface, as shown in the figure. The two frequencies then arise because there are two extremal cross-sections – the neck and the belly. As the direction of the magnetic field is changed, different extremal cross-sections become relevant. In this way, we can map out the entire Fermi surface.

The magnetisation is not the only quantity to exhibit oscillations. In fact, the large enhancement in the density of states affects nearly all observables. For example, oscillations in the conductivity are known as the *Shubikov-de Haas effect*.

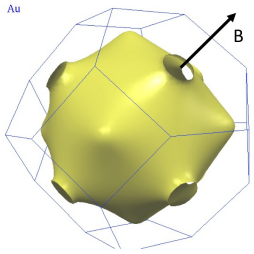


Figure 72: Gold

The experimental technique for measuring Fermi surfaces was pioneered by Brian Pippard, Cavendish professor and the first president of Clare Hall. Today, the techniques of quantum oscillations play an important role in attempts to better understand some of the more mysterious materials, such as unconventional superconductors.

5. Phonons

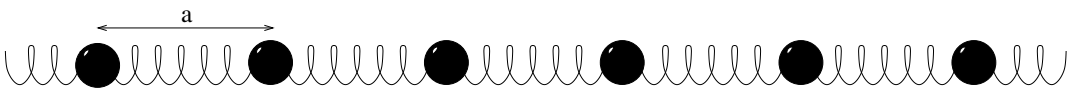
Until now, we've discussed lattices in which the atoms are fixed in place. This is, of course, somewhat unrealistic. In materials, atoms can jiggle, oscillating back and forth about their equilibrium position. The result of their collective effort is what we call sound waves or, at the quantum level, *phonons*. In this section we explore the physics of this jiggling.

5.1 Lattices in One Dimension

Much of the interesting physics can be illustrated by sticking to one-dimensional examples.

5.1.1 A Monotonic Chain

We start with a simple one-dimensional lattice consisting of N equally spaced, identical atoms, each of mass m. This is shown below.



We denote the position of each atom as x_n, with $n = 1, \ldots, N$. In equilibrium, the atoms sit at

$$x_n = na$$

with a the lattice spacing.

The potential that holds the atoms in place takes the form $\sum_n V(x_n - x_{n-1})$. For small deviations from equilibrium, a generic potential always looks like a harmonic oscillator. The deviation from equilibrium for the n^{th} atom is given by

$$u_n(t) = x_n(t) - na$$

The Hamiltonian governing the dynamics is then a bunch of coupled harmonic oscillators

$$H = \sum_n \frac{p_n^2}{2m} + \frac{\lambda}{2}\sum_n (u_n - u_{n-1})^2 \tag{5.1}$$

where $p_n = m\dot{u}_n$ and λ is the spring constant. (It is not to be confused with the wavelength.) The resulting equations of motion are

$$m\ddot{u}_n = -\lambda(2u_n - u_{n-1} - u_{n+1}) \tag{5.2}$$

To solve this equation, we need to stipulate some boundary conditions. It's simplest to impose periodic boundary conditions, extending $n \in \mathbf{Z}$ and requiring $u_{n+N} = u_n$. For $N \gg 1$, which is our interest, other boundary conditions do not qualitatively change the physics. We can then write the solution to (5.2) as

$$u_n = A\, e^{-i\omega t - ikna} \tag{5.3}$$

Because the equation is linear, we can always take real and imaginary parts of this solution. Moreover, the linearity ensures that the overall amplitude A will remain arbitrary.

The properties of the lattice put restrictions on the allowed values of k. First note that the solution is invariant under $k \to k + 2\pi/a$. This means that we can restrict k to lie in the first Brillouin zone,

$$k \in \left[-\frac{\pi}{a}, \frac{\pi}{a}\right)$$

Next, the periodic boundary conditions $u_{N+1} = u_1$ require that k takes values

$$k = \frac{2\pi}{Na} l \quad \text{with} \quad l = -\frac{N}{2}, \ldots, \frac{N}{2}$$

where, to make life somewhat easier, we will assume that N is even so l is an integer. We see that, as in previous sections, the short distance structure of the lattice determines the range of k. Meanwhile, the macroscopic size of the lattice determines the short distance structure of k. This, of course, is the essence of the Fourier transform. Before we proceed, it's worth mentioning that the minimum wavenumber $k = 2\pi/Na$ was something that we required when discussing the Debye model of phonons in the Statistical Physics lectures.

Our final task is to determine the frequency ω in terms of k. Substituting the ansatz into the formula (5.2), we have

$$m\omega^2 = \lambda\left(2 - e^{ika} - e^{-ika}\right) = 4\lambda \sin^2\left(\frac{ka}{2}\right)$$

We find the dispersion relation

$$\omega = 2\sqrt{\frac{\lambda}{m}}\left|\sin\left(\frac{ka}{2}\right)\right|$$

This dispersion relation is sketched Figure 73, with k ranging over the first Brillouin zone.

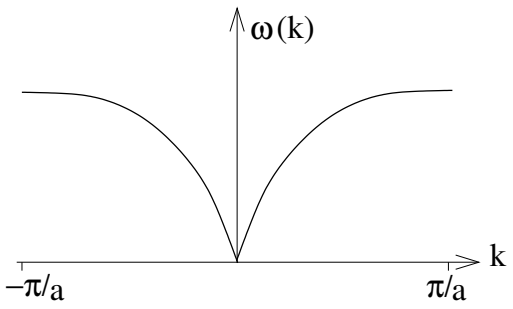


Figure 73: Phonon dispersion relation for a monatomic chain.

Many aspects of the above discussion are familiar from the discussion of electrons in the tight-binding model. In both cases, we end up with a dispersion relation over the Brillouin zone. But there are some important differences. In particular, at small values of k, the dispersion relation for phonons is linear

$$\omega \approx \sqrt{\frac{\lambda}{m}}\, ak$$

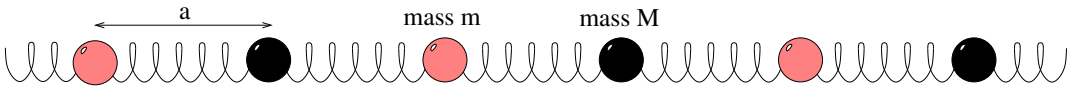
This is in contrast to the electron propagation where we get the dispersion relation for a non-relativistic, massive particle (3.6). Instead, the dispersion relation for phonons is more reminiscent of the massless, relativistic dispersion relation for light. For phonons, the ripples travel with speed

$$c_s = \sqrt{\frac{\lambda}{m}}a \tag{5.4}$$

This is the *speed of sound* in the material.

5.1.2 A Diatomic Chain

Consider now a linear chain of atoms, consisting of alternating atoms of different types.



The atoms on even sites have mass m; those on odd sites have mass M. For simplicity, we'll take the restoring forces between these atoms to be the same. The equations of motion are

$$\begin{aligned} m\ddot{u}_{2n} &= -\lambda(2u_{2n} - u_{2n-1} - u_{2n+1}) \\ M\ddot{u}_{2n+1} &= -\lambda(2u_{2n+1} - u_{2n} - u_{2n+2}) \end{aligned}$$

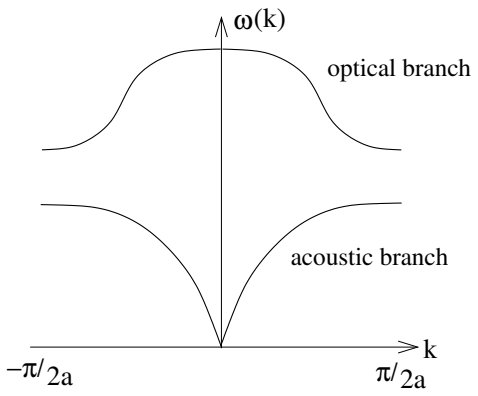


Figure 74: Phonon dispersion relation for a diatomic chain.

We make the ansatz

$$u_{2n} = A\, e^{-i\omega t - 2ikna} \quad \text{and} \quad u_{2n+1} = B\, e^{-i\omega t - 2ikna}$$

Note that these solutions are now invariant under $k \to k + \pi/a$. This reflects the fact that, if we take the identity of the atoms into account, the periodicity of the lattice is doubled. Correspondingly, the Brillouin zone is halved and k now lies in the range

$$k \in \left[-\frac{\pi}{2a}, \frac{\pi}{2a}\right) \tag{5.5}$$

Plugging our ansatz into the two equations of motion, we find a relation between the two amplitudes A and B,

$$\omega^2 \begin{pmatrix} m & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \lambda \begin{pmatrix} 2 & -(1+e^{-2ika}) \\ -(1+e^{2ika}) & 2 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} \tag{5.6}$$

This is viewed as an eigenvalue equation. The frequency ω is determined in terms of the wavenumber k by requiring that the appropriate determinant vanishes. This time we find that there are two frequencies for each wavevector, given by

$$\omega_\pm^2 = \frac{\lambda}{mM}\left[m + M \pm \sqrt{(m-M)^2 + 4mM\cos^2(ka)}\right]$$

The resulting dispersion relation is sketched in Figure 74 in the first Brillouin zone (5.5). Note that there is a gap in the spectrum on the boundary of the Brillouin zone, $k = \pm\pi/2a$, given by

$$\Delta E = \hbar(\omega_+ - \omega_-) = \hbar\sqrt{2\lambda}\left|\frac{1}{\sqrt{m}} - \frac{1}{\sqrt{M}}\right|$$

For $m = M$, the gap closes, and we reproduce the previous dispersion relation, now plotted on half the original Brillouin zone.

The lower ω_- part of the dispersion relation is called the *acoustic branch*. The upper ω_+ part is called the *optical branch*. To understand where these names come from, we need to look a little more closely at the the physical origin of these two branches. This comes from studying the eigenvectors of (5.6) which tells us the relative amplitudes of the two types of atoms.

This is simplest to do in the limit $k \to 0$. In this limit the acoustic branch has $\omega_- = 0$ and is associated to the eigenvector

$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The atoms move in phase in the acoustic branch. Meanwhile, in the optical branch we have $\omega_+^2 = 2\lambda(M^{-1} + m^{-1})$ with eigenvector

$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} M \\ -m \end{pmatrix}$$

In the optical branch, the atoms move out of phase.

Now we can explain the name. Often in a lattice, different sites contain ions of alternating charges: say, + on even sites and − on odd sites. But alternating charges oscillating out of phase create an electric dipole of frequency $\omega_+(k)$. This means that these vibrations of the lattice can emit or absorb light. This is the reason they are called "optical" phonons.

Although our discussion has been restricted to one-dimensional lattices, the same basic characterisation of phonon branches occurs for higher dimensional lattices. Acoustic branches have linear dispersion $\omega \sim k$ for low momenta, while optical branches have non-vanishing frequency, typically higher than the acoustic branch. The data for the phonon spectrum of $NaCl$ is shown on the right[10] and clearly exhibits these features.

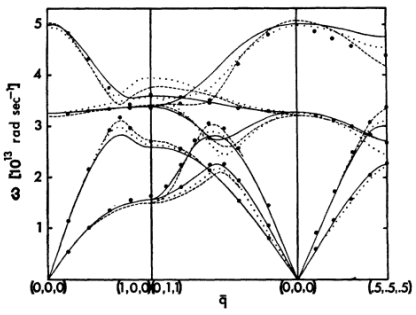


Figure 75:

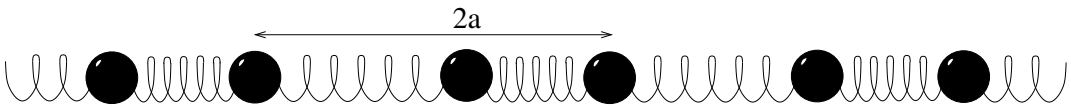
5.1.3 Peierls Transition

We now throw in two separate ingredients: we will consider the band structure of electrons, but also allow the underlying atoms to move. There is something rather special and surprising that happens for one-dimensional lattices.

[10]This was taken from "*Phonon Dispersion Relations in NaCl*", by G. Raumo, L. Almqvist and R. Stedman, Phys Rev. 178 (1969).

We consider the simple situation described in Section 5.1.1 where we have a one-dimensional lattice with spacing a. Suppose, further, that there is a single electron per lattice site. Because of the spin degree of freedom, it results in a half-filled band, as explained in Section 3.1. In other words, we have a conductor.

Consider a distortion of the lattice, in which successive pairs of atoms move closer to each other, as shown below.



Clearly this costs some energy since the atoms move away from their equilibrium positions. If each atom moves by an amount δx, we expect that the total energy cost is of order

$$U_{\rm lattice} \sim N\lambda(\delta x)^2 \tag{5.7}$$

What effect does this have on the electrons? The distortion has changed the lattice periodicity from a to $2a$. This, in turn, will halve the Brillouin zone so the electron states are now labeled by

$$k \in \Big[-\frac{\pi}{2a}, \frac{\pi}{2a}\Big)$$

More importantly, from the analysis of Section 3.1, we expect that a gap will open up in the electron spectrum at the edges of the Brillouin zone, $k = \pm\pi/2a$. In particular, the energies of the filled electron states will be pushed down; those of the empty electron states will be pushed up, as shown in the Figure 76. The question that we want to ask is: what is the energy reduction due to the electrons? In particular, is this more or less than the energy $U_{\rm lattice}$ that it cost to make the distortion in the first place?

Let's denote the dispersion relation before the distortion as $E_0(k)$, and the dispersion relation after the distortion as $E_-(k)$ for $|k| \in [0, \pi/2a)$ and $E_+(k)$ for $|k| \in [\pi/2a, \pi/a)$. The energy cost of the distortion due to the electrons is

$$U_{\rm electron} = -2\,\frac{Na}{2\pi}\int_{-\pi/2a}^{\pi/2a} dk\,\Big(E_0(k) - E_-(k)\Big) \tag{5.8}$$

Here the overall minus sign is because the electrons lose energy, the factor of 2 is to account for the spin degree of freedom, while the factor of $Na/2\pi$ is the density of states of the electrons.

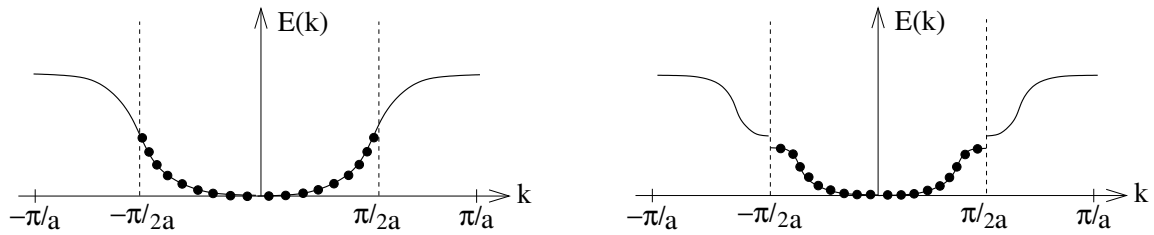


Figure 76: The distortion of the lattice reduces the energy of the Fermi sea of electrons.

To proceed, we need to get a better handle on $E_0(k)$ and $E_-(k)$. Neither are particularly nice functions. However, for a small distortion, we expect that the band structure is changed only in the immediate vicinity of $k = \pi/2a$. Whatever the form of $E_0(k)$, we can always approximate it by a linear function in this region,

$$E_0(k) \approx \mu + \nu q \quad \text{with } q = k - \frac{\pi}{2a} \tag{5.9}$$

where $\mu = E_0(\pi/2a)$ and $\nu = \partial E_0/\partial k$, again evaluated at $k = \pi/2a$. Note that $q < 0$ for the filled states, and $q > 0$ for the unfilled states.

We can compute $E_-(k)$ in this region by the same kind of analysis that we did in Section 3.1. Suppose that the distortion opens up a gap Δ at $k = \pi/2a$. Since there is no gap unless there is a distortion of the lattice, we expect that

$$\Delta \sim \delta x \tag{5.10}$$

(or perhaps δx to some power). To compute $E_-(k)$ in the vicinity of the gap, we can use our earlier result (3.16). Adapted to the present context, the energy E close to $k = \pi/2a$ is given by

$$\Big(E_0(\pi/2a + q) - E\Big)\Big(E_0(\pi/2a - q) - E\Big) - \frac{\Delta^2}{4} = 0$$

Using our linearisation (5.9) of E_0, we can solve this quadratic to find the dispersion relation

$$E_\pm(q) = \mu \pm \sqrt{\nu^2 q^2 + \frac{\Delta^2}{4}}$$

Note that when evaluated at $q = 0$, we find the gap $E_+ - E_- = \Delta$, as expected. The filled states sit in the lower branch E_-. The energy gained by the electrons (5.8) is

dominated by the regions $k = \pm\pi/2a$. By symmetry, it is the same in both and given by

$$U_{\text{electron}} \approx -\frac{Na}{\pi}\int_{-\Lambda}^{0} dq\ \left(\nu q + \sqrt{\nu^2 q^2 + \frac{\Delta^2}{4}}\right)$$

Here we have introduced a lower cut-off $-\Lambda$ on the integral; it will not ultimately be important where we take this cut-off, although we will require $\nu\Lambda \gg \Delta$. The integral is straightforward to evaluate exactly. However, our interest lies in what happens when Δ is small. In this limit, we have

$$U_{\text{electron}} \approx -\frac{Na}{\pi}\left[\frac{\Delta^2}{16\nu} - \frac{\Delta^2}{8\nu}\log\left(\frac{\Delta}{4\nu\Lambda}\right)\right]$$

Both terms contribute to the decrease in energy of the electrons. The first term is of order Δ^2 and hence, through (5.10), of order δx^2. This competes with the energy cost from the lattice distortion (5.7), but there is no guarantee that it is either bigger or smaller. The second term with the log is more interesting. For small Δ, this always beats the quadratic cost of the lattice distortion (5.7).

We reach a surprising conclusion: a half-filled band in one-dimension is unstable. The lattice rearranges itself to turn the metal into an insulator. This is known as the Peierls transition; it is an example of a metal-insulator transition. This striking behaviour can be seen in one-dimensional polymer chains, such as the catchily named TTF-TCNQ shown in the figure[11]. The resistivity – plotted on the vertical axis – rises sharply when the temperature drops to the scale Δ. (The figure also reveals another feature: as the pressure is increased, the resistivity no longer rises quite as sharply, and by the time you get to 8 GPa there is no rise at all. This is because of the interactions between electrons become important.)

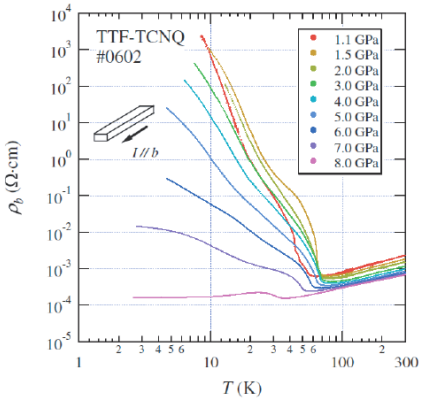


Figure 77:

5.1.4 Quantum Vibrations

Our discussion so far has treated the phonons purely classically. Now we turn to their quantisation. At heart this is not difficult – after all, we just have a bunch of harmonic oscillators. However, they are coupled in an interesting way and the trick is to disentangle them. It turns out that we've already achieved this disentangling by writing down the classical solutions.

[11]This data is taken from "*Recent progress in high-pressure studies on organic conductors*", by S. Yasuzuka and K. Murata (2009)

We have a classical solution (5.3) for each $k_l = 2\pi l/Na$ with $l = -N/2, \ldots, N/2$. We will call the corresponding frequency $\omega_l = 2\sqrt{\lambda/m}|\sin(k_l a/2)|$. We can introduce a different amplitude for each l. The most general classical solution then takes the form

$$u_n(t) = X_0(t) + \sum_{l\neq 0}\Big[\alpha_l\, e^{-i(\omega_l t - k_l na)} + \alpha_l^\dagger\, e^{i(\omega_l t - k_l na)}\Big] \tag{5.11}$$

This requires some explanation. First, we sum over all modes $l = -N/2, \ldots, +N/2$ with the exception of $l = 0$. This has been singled out and written as $X_0(t)$. It is the centre of mass, reflecting the fact that the entire lattice can move as one. The amplitudes for each $l \neq 0$ mode are denoted α_l. Finally, we have taken the real part of the solution because, ultimately, $u_n(t)$ should be real. Note that we've denoted the complex conjugation by $\alpha_l^\dagger$ rather than $\alpha_l^\star$ in anticipation of the quantisation that we will turn to shortly.

The momentum $p_n(t) = m\dot{u}_n$ is given by

$$p_n(t) = P_0(t) + \sum_{l\neq 0}\Big[-im\omega_l\alpha_l\, e^{-i(\omega_l t - k_l na)} + im\omega_l\alpha_l^\dagger\, e^{i(\omega_l t - k_l na)}\Big]$$

Now we turn to the quantum theory. We promote u_n and p_n to operators acting on a Hilbert space. We should think of $u_n(t)$ and $p_n(t)$ as operators in the Heisenberg representation; we can get the corresponding operators in the Schrödinger representation simply by setting $t = 0$.

Since u_n and p_n are operators, the amplitudes α_l and $\alpha_l^\dagger$ must also be operators if we want these equations to continue to make sense. We can invert the equations above by setting $t = 0$ and looking at

$$\sum_{n=1}^{N} u_n\, e^{-ik_l na} = \sum_n\sum_{l'}\Big[\alpha_{l'}\, e^{-i(k_l - k_{l'})na} + \alpha_{l'}^\dagger\, e^{-i(k_l + k_{l'})na}\Big] = N(\alpha_l + \alpha_{-l}^\dagger)$$

Similarly,

$$\sum_{n=1}^{N} p_n\, e^{ik_l na} = \sum_n\sum_{l'}\Big[-im\omega_{l'}\alpha_{l'}\, e^{-i(k_l - k_{l'})na} + im\omega_{l'}\alpha_{l'}^\dagger\, e^{-i(k_l + k_{l'})na}\Big] = -iNm\omega_l(\alpha_l - \alpha_{-l}^\dagger)$$

where we've used the fact that $\omega_l = \omega_{-l}$. We can invert these equations to find

$$\begin{aligned}\alpha_l &= \frac{1}{2m\omega_l N}\sum_n e^{-ik_l na}\big(m\omega_l u_n + ip_n\big)\\ \alpha_l^\dagger &= \frac{1}{2m\omega_l N}\sum_n e^{ik_l na}\big(m\omega_l u_n - ip_n\big)\end{aligned} \tag{5.12}$$

Similarly, we can write the centre of mass coordinates — which are also now operators — as

$$X_0 = \frac{1}{N}\sum_n u_n \quad \text{and} \quad P_0 = \frac{1}{N}\sum_n p_n \tag{5.13}$$

At this point, we're ready to turn to the commutation relations. The position and momentum of each atom satisfy

$$[u_n, p_{n'}] = i\hbar\delta_{n,n'}$$

A short calculation using the expressions above reveals that X_0 and P_0 obey the relations

$$[X_0, P_0] = \frac{i\hbar}{N}$$

Meanwhile, the amplitudes obey the commutation relations

$$[\alpha_l, \alpha_{l'}^\dagger] = \frac{\hbar}{2m\omega_l N}\delta_{l,l'} \quad \text{and} \quad [\alpha_l, \alpha_{l'}] = [\alpha_l^\dagger, \alpha_{l'}^\dagger] = 0$$

This is something that we've seen before: they are simply the creation and annihilation operators of a simple harmonic oscillator. We rescale

$$\alpha_l = \sqrt{\frac{\hbar}{2m\omega_l N}}\, a_l \tag{5.14}$$

then our new operators a_l obey

$$[a_l, a_{l'}^\dagger] = \delta_{l,l'} \quad \text{and} \quad [a_l, a_{l'}] = [a_l^\dagger, a_{l'}^\dagger] = 0$$

Phonons

We now turn to the Hamiltonian (5.1). Substituting in our expressions (5.12) and (5.13), and after a bit of tedious algebra, we find the Hamiltonian

$$H = \frac{P_0^2}{2M} + \sum_{l\neq 0}\left(a_l^\dagger a_l + \frac{1}{2}\right)\hbar\omega_l$$

Here $M = Nm$ is the mass of the entire lattice. Since this is a macroscopically large object, we set $P_0 = 0$ and focus on the Hilbert space arising from the creation operators $a_l^\dagger$. After our manipulations, these are simply N, decoupled harmonic oscillators.

The ground state of the system is a state $|0\rangle$ obeying

$$a_l|0\rangle = 0 \quad \forall\ l$$

Each harmonic oscillator gives a contribution of $\hbar\omega_l/2$ to the zero-point energy E_0 of the ground state. However, this is of no interest. All we care about is the energy difference between excited states and the ground state. For this reason, it's common practice to redefine the Hamiltonian to be simply

$$H = \sum_{l\neq 0} \hbar\omega_l a_l^\dagger a_l$$

so that $H|0\rangle = 0$.

The excited states of the lattice are identical to the excited states of the harmonic oscillators. For each l, the first excited state is given by $a_l^\dagger|0\rangle$ and has energy $E = \hbar\omega_l$. However, although the mathematics is identical to that of the harmonic oscillator, the physical interpretation of this state is rather different. That's because it has a further quantum number associated to it: this state carries crystal momentum $\hbar k_l$. But an object which carries both energy and momentum is what we call a particle! In this case, it's a particle which, like all momentum eigenstates, is not localised in space. This particle is a quantum of the lattice vibration. It is called the *phonon*.

Note that the coupling between the atoms has lead to a quantitative change in the physics. If there was no coupling between atoms, each would oscillate with frequency $m\lambda$ and the minimum energy required to excite the system would be $\sim \hbar m\lambda$. However, when the atoms are coupled together, the normal modes now vibrate with frequencies ω_l. For small k, these are $\omega_l \approx \sqrt{\frac{\lambda\pi^2}{m}}\frac{l}{N}$. The key thing to notice here is the factor of $1/N$. In the limit of an infinite lattice, $N \to \infty$, there are excited states with infinitesimally small energies. We say that the system is *gapless*, meaning that there is no gap betwen the ground state and first excited state. In general, the question of whether a bunch interacting particles is gapped or gapless is one of the most basic (and, sometimes, most subtle) questions that you can ask about a system.

Any state in the Hilbert space can be written in the form

$$|\psi\rangle = \prod_l \frac{(a_l^\dagger)^{n_l}}{\sqrt{n_l!}}\,|0\rangle$$

and has energy

$$H|\psi\rangle = \sum_l \hbar n_l\omega_l$$

This state should be thought of as described $\sum_l n_l$ phonons and decomposes into n_l phonons with momentum $\hbar k_l$ for each l. The full Hilbert space constructed in this way contains states consisting of an arbitrary number of particles. It is referred to as a *Fock space*.

Because the creation operators $a_l^\dagger$ commute with each other, there is no difference between the state $|\psi\rangle \sim a_l^\dagger a_{l'}^\dagger|0\rangle$ and $|\psi\rangle \sim a_{l'}^\dagger a_l^\dagger|0\rangle$. This is the statement that phonons are bosons.

The idea that harmonic oscillator creation operators actually create particles sometimes goes by the terrible name of *second quantisation*. It is misleading — nothing has been quantised twice.

Quantisation of Acoustic and Optical Phonons

It is not difficult to adapt the discussion above to vibrations of a diatomic lattice that we met in Section 5.1.2. We introduce two *polarization vectors*, $\mathbf{e}_\pm(k)$. These are eigenvectors obeying the matrix equation (5.6),

$$\begin{pmatrix} 2 & -(1+e^{-2ika}) \\ -(1+e^{2ika}) & 2 \end{pmatrix} \mathbf{e}_\pm(k) = \frac{\omega_\pm^2}{\lambda} \begin{pmatrix} m & 0 \\ 0 & M \end{pmatrix} \mathbf{e}_\pm(k)$$

We then write the general solution as

$$\begin{pmatrix} u_{2n}(t) \\ u_{2n+1}(t) \end{pmatrix} = \sum_{k\in BZ} \sum_{s=\pm} \sqrt{\frac{\hbar}{2N\omega_s(k)}} \left[a_s(k)\mathbf{e}_s(k)e^{i(\omega_s t+2kna)} + a_s^\dagger(k)\mathbf{e}_s^\star(k)e^{-i(\omega_s t+2kna)} \right]$$

where the creation operators obey

$$[a_s(k), a_{s'}(k')^\dagger] = \delta_{s,s'}\delta_{k,k'} \quad \text{and} \quad [a_s(k), a_{s'}(k')] = [a_s^\dagger(k), a_{s'}(k')^\dagger] = 0$$

Now the operators $a_-^\dagger(k)$ create acoustic phonons while $a_+^\dagger(k)$ create optical phonons, each with momentum $\hbar k$.

5.1.5 The Mössbauer Effect

There's a rather nice application of phonons that goes by the name of the *Mössbauer effect*. This is to do with how nuclei in solids absorb gamma rays.

To understand this, we first need to think about atoms absorb light, and then contrast this with how nuclei absorb light. To this end, consider a gas of atoms, all sitting in the ground state. If we shine light on the atoms at very specific frequencies, then the atoms will absorb the light by jumping to excited states. The frequency should be

$$E_\gamma = \hbar\nu = E_{\rm excite}$$

where $E_{\rm excite}$ is the energy difference between the excited state and the ground state. Once the atom absorbs a photon, it will sit in the excited state for some time and then decay. If it drops back down to the ground state, the emitted photon will again have energy E_γ and can be absorbed by another atom. This then repeats, a process known as *resonant absoption*.

However, a little thought shows that the situation is slightly more complicated than we've made out. Suppose, for simplicity, that the original atom was at rest. In the collision with the atom, both energy and momentum must be conserved. The momentum of the incoming photon is $p_\gamma = E_\gamma/c$ and, after the collision, this is transferred to the atom, so $p_{\rm atom} = E_\gamma/c$. This means that the atom has kinetic energy from the recoil,

$$E_{\rm recoil} = \frac{p_{\rm atom}^2}{2M} = \frac{E_\gamma^2}{2Mc^2} \tag{5.15}$$

where M is the mass of the atom. (The speed of the atom is small enough that we can use the non-relativistic form of kinetic energy.) So we see that it's not quite right to say that the energy of the photon should be tuned to the energy difference $E_{\rm excite}$ because this ignores the energy that goes into the recoil. Instead, the incoming photon should have slightly higher energy, $E_\gamma = E_{\rm excite} + E_{\rm recoil}$, or

$$E_\gamma = E_{\rm excite} + \frac{E_\gamma^2}{2Mc^2} \quad \Rightarrow \quad E_\gamma \approx E_{\rm excite} + \frac{(E_{\rm excite})^2}{2Mc^2} + \ldots \tag{5.16}$$

Meanwhile, when the atom now decays back to the ground state, it will emit the photon in a random direction. This means that the atom typically remains in motion; indeed, it's quite possible that the kinetic energy of atom increases yet again if it emits the photon back in the direction it came. All of this means that the energy of the emitted photon that the atom emits is smaller than the energy of the photon that it absorbed.

The question is: what happens next? In particular, is it possible for this emitted photon to be re-absorbed by a different atom so that we get resonant absorption? This is now a quantitative question, rather than a qualitative question. The key point is that you don't need to tune the frequency of light exactly to $E_{\rm excite}$ in order to excite

an atom. Instead, there is a range of energies – a so-called *line width* – that will do the job. This line width is related to the lifetime τ of the excited state by $\Delta E \sim \hbar/\tau$. (See the chapter on scattering in the lectures on Topics in Quantum Mechanics for more details.)

Let's put in some numbers. The energy needed to excite an electron from one level to another is measured in $E_{\rm excite} \approx$ eV. Meanwhile the mass of, say, an iron atom is around $Mc^2 \sim 5 \times 10^4$ MeV. This means that the correction term (5.16) in the photon energy is of order $\Delta E_\gamma \approx 10^{-11}$ eV. This is significantly smaller than the line width of atomic excitations, and the discussion above has no relevance to absorption of light due to transitions of electrons from one energy level to another.

However, things are very different when it comes to nuclear transitions. Now the relevant excitation energy is of order $E_{\rm excite} \approx 10^4$ eV, corresponding to soft gamma rays, and the correction term (5.16) in the photon energy due to recoil effects is $\Delta E \approx 10^{-3}$ eV. This time the energy is significantly larger than the line width: a typical nuclear excitation has lifetime $\tau \sim 10^{-7}$ seconds and a width $\Gamma \sim 10^{-8}$ eV. The upshot of this argument is that, while X-ray absorption lines are seen corresponding to atomic excitations, we should not expect to see a repeat in the gamma-ray spectrum associated to nuclear excitations.

And yet.... while it's true that gamma ray resonant absorption lines are not seen in gasses, they are seen solids. This is the *Mössbauer effect*. The important point is that a nucleus in an atom is coupled to all the other atoms through the bonds in a solid. A nucleus will recoil when hit by a photon, as in the discussion above, but now the atom will bounce back into position and the energy $E_{\rm recoil}$ will typically be distributed into phonon degrees of freedom. When there are a large number of phonons excited, the story is not different from that told above, and the emitted photon has a sufficiently different frequency to kill resonant absorption. However, there is some probability that no phonons are created, but instead the entire solid moves absorbs the momentum of the photon. In this case, the recoil energy is still given by (5.15) but with M is the mass of the solid, rather than the mass of a single atom. This gives an extra factor of around 10^{23} in the denominator, and the recoil energy becomes negligible. For this to happen, the entire solid must react coherently as a single quantum object! The resulting gamma ray resonant absorption spectrum is indeed observed.

5.2 From Atoms to Fields

If we look at a solid at suitably macroscopic distances, we don't notice the underlying atomic structure. Nonetheless, it's still straightforward to detect sound waves. This suggests that we should be able to formulate a continuum description of the solid that is ignorant of the underlying atomic make-up.

With this in mind, we define the *displacement field* for a one-dimensional lattice. This is a function $u(x,t)$. It is initially defined only at the lattice points

$$u(x = na) = u_n$$

However, we then extend this field to all $x \in \mathbf{R}$, with the proviso that our theory will cease to make sense if $u(x)$ varies appreciably on scales smaller than a.

The equation governing the atomic displacements is (5.2)

$$m\ddot{u}_n = -\lambda(2u_n - u_{n-1} - u_{n+1})$$

In the continuum limit, this difference equation becomes the wave equation

$$\rho\frac{\partial^2 u}{\partial t^2} = -\lambda'\frac{\partial^2 u}{\partial x^2} \tag{5.17}$$

where $\rho = m/a$ is the density of our one-dimensional solid, and $\lambda' = \lambda a$. These are the macroscopic parameters. Note, in particular, that the speed of sound (5.4) can be written purely in terms of these macroscopic parameters, $c_s^2 = \lambda'/\rho$.

The equation of motion (5.17) can be derived from the action

$$S = \int dtdx\ \left[\frac{\rho}{2}\left(\frac{\partial u}{\partial t}\right)^2 - \frac{\lambda'}{2}\left(\frac{\partial u}{\partial x}\right)^2\right]$$

This is the field theory for the phonons of a one-dimensional solid.

5.2.1 Phonons in Three Dimensions

For three-dimensional solids, there are three displacement fields, $u_i(\mathbf{x})$, one for each direction in which the lattice can deform. In general, the resulting action can depend on various quantities $\partial u_i/\partial x^j$. However, if the underlying lattice is such that the long-wavelength dynamics is rotationally invariant, then the action can only be a function of the symmetric combination

$$u_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x^j} + \frac{\partial u_j}{\partial x^i}\right)$$

If we want an equation of motion linear in the displacement, then the most general action is a function of $u_{ij}u_{ij}$ or u_{kk}^2. (The term u_{kk} is a total derivative and does not affect the equation of motion). We have

$$S = \int dtd^3x \; \frac{1}{2}\left[\rho\left(\frac{\partial u_i}{\partial t}\right)^2 - 2\mu\, u_{ij}u_{ij} - \lambda\, u_{ii}u_{jj}\right] \tag{5.18}$$

The coefficients μ and λ are called Lamé coeffcients; they characterise the underlying solid.

This action gives rise to the equations of motion

$$\rho\frac{\partial^2 u_i}{\partial t^2} = (\mu+\lambda)\frac{\partial^2 u_j}{\partial x^i \partial x^j} + \mu\frac{\partial^2 u_i}{\partial x^j \partial x^j} \tag{5.19}$$

We can look for solutions of the form

$$u_i(\mathbf{x},t) = \epsilon_i\, e^{i(\mathbf{k}\cdot\mathbf{x}+\omega t)}$$

where ϵ_i determines the polarisation of the wave. Plugging this ansatz into the equation of motion gives us the relation

$$\rho\omega^2\epsilon_i = \mu k^2\epsilon_i + (\mu+\lambda)(\boldsymbol{\epsilon}\cdot\mathbf{k})k_i$$

The frequency of the wave depends on the polarisation. There are two different options. Longitudinal waves have $\mathbf{k}\sim\boldsymbol{\epsilon}$. These have dispersion

$$\omega^2 = \frac{2\mu+\lambda}{\rho}\,k^2 \tag{5.20}$$

Meanwhile, transverse waves have $\boldsymbol{\epsilon}\cdot\mathbf{k}=0$ and dispersion

$$\omega^2 = \frac{\mu}{\rho}k^2 \tag{5.21}$$

Note that both of these dispersion relations are linear. The continuum approximation only captures the low-k limit of the full lattice system and does not see the bending of the dispersion relation close to the edge of the Brillouin zone. This is because it is valid only at long wavelengths, $ka \ll 1$.

The general solution to (5.19) is then

$$u_i(\mathbf{x},t) = \sum_s \int \frac{d^3k}{(2\pi)^3}\;\frac{1}{2\rho\omega_s(k)}\,\epsilon_i^s\left(a_s(\mathbf{k})\,e^{i(\mathbf{k}\cdot\mathbf{x}-\omega_s t)} + a_s^\dagger(\mathbf{k})\,e^{-i(\mathbf{k}\cdot\mathbf{x}-\omega_s t)}\right) \tag{5.22}$$

where the s sum is over the three polarisation vectors, two transverse and one longitudinal. The frequencies $\omega_s(k)$ correspond to either (5.20) or (5.21) depending on the choice of s.

5.2.2 From Fields to Phonons

Although we have discarded the underlying atoms, this does not mean that we have lost the discrete nature of phonons. To recover them, we must quantise the field theory defined by the action (5.18). This is the subject of *Quantum Field Theory*. You will learn much (much) more about this in next year's lectures. What follows is merely a brief taster for things to come.

To quantise the field, we need only follow the same path that we took in Section 5.1.4. At every step, we simply replace the discrete index n with the continuous index $\mathbf{x}$. Note, in particular, that $\mathbf{x}$ is not a dynamical variable in field theory; it is simply a label.

First, we turn the field $u(\mathbf{x})$ into an operator. This means that the amplitudes $a_s(\mathbf{k})$ and $a_s^\dagger(\mathbf{k})$ in (5.22) also become operators. To proceed, we need the momentum conjugate to $u_i(\mathbf{x}, t)$. This too is now a field, and is determined by the usual rules of classical dynamics,

$$\pi_i(\mathbf{x}) = \frac{\partial L}{\partial \dot{u}_i} = \rho \dot{u}_i$$

Written in terms of the solution (5.22), we have

$$\pi_i(\mathbf{x}, t) = \rho \sum_s \int \frac{d^3k}{(2\pi)^3} \, \frac{1}{2\rho\omega_s(k)} \, \epsilon_i^s \left(-i\omega_s a_s(\mathbf{k}) \, e^{i(\mathbf{k}\cdot\mathbf{x} - \omega_s t)} + i\omega_s a_s^\dagger(\mathbf{k}) \, e^{-i(\mathbf{k}\cdot\mathbf{x} - \omega_s t)} \right)$$

The canonical commutation relations are the field-theoretical analog of the usual position-momentum commutation relations,

$$[u_i(\mathbf{x}), \boldsymbol{\pi}_j(\mathbf{x}')] = i\hbar \, \delta_{ij} \, \delta^3(\mathbf{x} - \mathbf{x}')$$

At this point we have some straightforward but tedious calculations ahead of us. We will skip these on the grounds that you will see them in glorious detail in later courses. The first is an inverse Fourier transform, which expresses $a_s(\mathbf{k})$ and $a_s^\dagger(\mathbf{k})$ in terms of $u_i(\mathbf{x})$ and $\pi_i(\mathbf{x})$. The result is analogous to (5.12). We then use this to determine the commutation relations,

$$[a_s(\mathbf{k}), a_{s'}^\dagger(\mathbf{k}')] = \delta_{s,s'} \, \delta^3(\mathbf{k} - \mathbf{k}') \quad \text{and} \quad [a_s(\mathbf{k}), a_{s'}(\mathbf{k}')] = [a_s^\dagger(\mathbf{k}), a_{s'}^\dagger(\mathbf{k}')] = 0$$

This is the statement that these are creation and annihilation operators for harmonic oscillators, now labelled by both a discrete polarisation index $s = 1, 2, 3$ as well as the continuous momentum index $\mathbf{k}$.

The next fairly tedious calculation is the Hamiltonian. This too follows from standard rules of classical dynamics, together with a bunch of Fourier transforms. When the dust settles, we find that, up to an irrelevant overall constant,

$$H = \sum_s \int \frac{d^3k}{(2\pi)^3}\ \hbar\omega_s(k) a_s^\dagger(\mathbf{k}) a_s(\mathbf{k})$$

This is simply the Hamiltonian for an infinite number of harmonic oscillators.

The interpretation is the same as we saw in Section 5.1.4. We define the ground state of the field theory to obey $a_s(\mathbf{k})|0\rangle = 0$ for all s and for all $\mathbf{k}$. The Fourier modes of the field $a_s^\dagger(\mathbf{k})$ are then to be viewed as creating and destroying phonons which carry momentum $\hbar\mathbf{k}$, polarisation $\boldsymbol{\epsilon}_s$ and energy $\hbar\omega_s(k)$. In this way, we see particles emerging from an underlying field.

Lessons for the Future

This has been a very quick pass through some basic quantum field theory, applied to the vibrations of the lattice. Buried within the mathematics of this section are two, key physical ideas. The first is that a coarse grained description of atomic vibrations can be described in terms of a continuous field. The second is that quantisation of the field results in particles that, in the present context, we call phonons.

There is a very important lesson to take from the second of these ideas, a lesson which extends well beyond the study of solids. All of the fundamental particles that we know of in Nature – whether electrons, quarks, photons, or anything else — arise from the quantisation of an underlying field. This is entirely analogous to the way that phonons arose in the discussion above.

Is there also a lesson to take away from the first idea above? Could it be that the fundamental fields of Nature themselves arise from coarse-graining something smaller? The honest answer is that we don't know. However, perhaps surprisingly, all signs point towards this *not* being the case. First, and most importantly, there is no experimental evidence that the fundamental fields in our Universe have a discrete underpinning. But at the theoretical level, there are some deep mathematical reasons — to do with chiral fermions and topology — which suggest that it is not possible to find a discrete system from which the known laws of physics emerge. It would appear that our Universe does not have something akin to the atomic lattice which underlies the phonon field. Understanding these issues remains a vibrant topic of research, both in condensed matter physics and in high energy physics.

6. Particles in a Magnetic Field

The purpose of this chapter is to understand how quantum particles react to magnetic fields. In contrast to later sections, we will not yet place these particles inside solids, for the simple reason that there is plenty of interesting behaviour to discover before we do this. Later, in Section 4.1, we will understand how these magnetic fields affect the electrons in solids.

Before we get to describe quantum effects, we first need to highlight a few of the more subtle aspects that arise when discussing classical physics in the presence of a magnetic field.

6.1 Gauge Fields

Recall from our lectures on Electromagnetism that the electric field $\mathbf{E}(\mathbf{x}, t)$ and magnetic field $\mathbf{B}(\mathbf{x}, t)$ can be written in terms a scalar potential $\phi(\mathbf{x}, t)$ and a vector potential $\mathbf{A}(\mathbf{x}, t)$,

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t} \quad \text{and} \quad \mathbf{B} = \nabla \times \mathbf{A} \tag{6.1}$$

Both ϕ and $\mathbf{A}$ are referred to as *gauge fields*. When we first learn electromagnetism, they are introduced merely as handy tricks to help solve the Maxwell equations. However, as we proceed through theoretical physics, we learn that they play a more fundamental role. In particular, they are necessary if we want to discuss a Lagrangian or Hamiltonian approach to electromagnetism. We will soon see that these gauge fields are quite indispensable in quantum mechanics.

The Lagrangian for a particle of charge q and mass m moving in a background electromagnetic fields is

$$L = \frac{1}{2}m\dot{\mathbf{x}}^2 + q\dot{\mathbf{x}} \cdot \mathbf{A} - q\phi \tag{6.2}$$

The classical equation of motion arising from this Lagrangian is

$$m\ddot{\mathbf{x}} = q\left(\mathbf{E} + \dot{\mathbf{x}} \times \mathbf{B}\right)$$

This is the *Lorentz force law*.

Before we proceed I should warn you of a minus sign issue. We will work with a general charge q. However, many textbooks work with the charge of the electron, written as $q = -e$. If this minus sign leans to confusion, you should blame Benjamin Franklin.

An Example: Motion in a Constant Magnetic Field

We'll take a constant magnetic field, pointing in the z-direction: $\mathbf{B} = (0, 0, B)$. We'll take $\mathbf{E} = 0$. The particle is free in the z-direction, with the equation of motion $m\ddot{z} = 0$. The more interesting dynamics takes place in the (x, y)-plane where the equations of motion are

$$m\ddot{x} = qB\dot{y} \quad \text{and} \quad m\ddot{y} = -qB\dot{x} \tag{6.3}$$

which has general solution is

$$x(t) = X + R\sin(\omega_B(t - t_0)) \quad \text{and} \quad y(t) = Y + R\cos(\omega_B(t - t_0))$$

We see that the particle moves in a circle which, for $B > 0$ and $q > 0$, is in a clockwise direction. The *cyclotron frequency* is defined by

$$\omega_B = \frac{qB}{m} \tag{6.4}$$

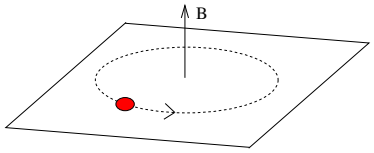


Figure 78:

The centre of the circle (X, Y), the radius of the circle R and the phase t_0 are all arbitrary. These are the four integration constants expected in the solution of two, second order differential equations.

6.1.1 The Hamiltonian

The *canonical momentum* in the presence of gauge fields is

$$\mathbf{p} = \frac{\partial L}{\partial \dot{\mathbf{x}}} = m\dot{\mathbf{x}} + q\mathbf{A} \tag{6.5}$$

This clearly is not the same as what we naively call momentum, namely $m\dot{\mathbf{x}}$.

The Hamiltonian is given by

$$H = \dot{\mathbf{x}} \cdot \mathbf{p} - L = \frac{1}{2m}(\mathbf{p} - q\mathbf{A})^2 + q\phi$$

Written in terms of the velocity of the particle, the Hamiltonian looks the same as it would in the absence of a magnetic field: $H = \frac{1}{2}m\dot{\mathbf{x}}^2 + q\phi$. This is the statement that a magnetic field does no work and so doesn't change the energy of the system. However, there's more to the Hamiltonian framework than just the value of H. We need to remember which variables are canonical. This information is encoded in the Poisson bracket structure of the theory (or, in fancy language, the symplectic structure on phase space). The fact that $\mathbf{x}$ and $\mathbf{p}$ are canonical means that

$$\{x_i, p_j\} = \delta_{ij} \quad \text{with} \quad \{x_i, x_j\} = \{p_i, p_j\} = 0$$

In the quantum theory, this structure transferred onto commutation relations between operators, which become

$$[x_i, p_j] = i\hbar\delta_{ij} \quad \text{with} \quad [x_i, x_j] = [p_i, p_j] = 0$$

6.1.2 Gauge Transformations

The gauge fields $\mathbf{A}$ and ϕ are not unique. We can change them as

$$\phi \to \phi - \frac{\partial \alpha}{\partial t} \quad \text{and} \quad \mathbf{A} \to \mathbf{A} + \nabla\alpha \tag{6.6}$$

for any function $\alpha(\mathbf{x}, t)$. Under these transformations, the electric and magnetic fields (6.1) remain unchanged. The Lagrangian (6.2) changes by a total derivative, but this is sufficient to ensure that the resulting equations of motion (6.3) are unchanged. Different choices of α are said to be different choices of *gauge*. We'll see some examples below.

The existence of gauge transformations is a redundancy in our description of the system: fields which differ by the transformation (6.6) describe physically identical configurations. Nothing that we can physically measure can depend on our choice of gauge. This, it turns out, is a beautifully subtle and powerful restriction. We will start to explore some of these subtleties in Sections 6.3 and 6.4

The canonical momentum $\mathbf{p}$ defined in (6.5) is not gauge invariant: it transforms as $\mathbf{p} \to \mathbf{p} + q\nabla\alpha$. This means that the numerical value of $\mathbf{p}$ can't have any physical meaning since it depends on our choice of gauge. In contrast, the velocity of the particle $\dot{\mathbf{x}}$ is gauge invariant, and therefore physical.

The Schrödinger Equation

Finally, we can turn to the quantum theory. We'll look at the spectrum in the next section, but first we wish to understand how gauge transformations work. Following the usual quantisation procedure, we replace the canonical momentum with

$$\mathbf{p} \mapsto -i\hbar\nabla$$

The time-dependent Schrödinger equation for a particle in an electric and magnetic field then takes the form

$$i\hbar\frac{\partial\psi}{\partial t} = H\psi = \frac{1}{2m}\Big(-i\hbar\nabla - q\mathbf{A}\Big)^2\psi + q\phi\psi \tag{6.7}$$

The shift of the kinetic term to incorporate the vector potential $\mathbf{A}$ is sometimes referred to as *minimal coupling*.

Before we solve for the spectrum, there are two lessons to take away. The first is that it is not possible to formulate the quantum mechanics of particles moving in electric and magnetic fields in terms of $\mathbf{E}$ and $\mathbf{B}$ alone. We're obliged to introduce the gauge fields $\mathbf{A}$ and ϕ. This might make you wonder if, perhaps, there is more to $\mathbf{A}$ and ϕ than we first thought. We'll see the answer to this question in Section 6.3. (Spoiler: the answer is yes.)

The second lesson follows from looking at how (6.7) fares under gauge transformations. It is simple to check that the Schrödinger equation transforms covariantly (i.e. in a nice way) only if the wavefunction itself also transforms with a position-dependent phase

$$\psi(\mathbf{x}, t) \to e^{iq\alpha(\mathbf{x},t)/\hbar}\psi(\mathbf{x}, t) \tag{6.8}$$

This is closely related to the fact that $\mathbf{p}$ is not gauge invariant in the presence of a magnetic field. Importantly, this gauge transformation does not affect physical probabilities which are given by $|\psi|^2$.

The simplest way to see that the Schrödinger equation transforms nicely under the gauge transformation (6.8) is to define the *covariant derivatives*

$$\mathcal{D}_t = \frac{\partial}{\partial t} + \frac{iq}{\hbar}\phi \quad \text{and} \quad \mathcal{D}_i = \frac{\partial}{\partial x^i} - \frac{iq}{\hbar}A_i$$

In terms of these covariant derivatives, the Schrödinger equation becomes

$$i\hbar\mathcal{D}_t\psi = -\frac{\hbar}{2m}\mathcal{D}^2\psi \tag{6.9}$$

But these covariant derivatives are designed to transform nicely under a gauge transformation (6.6) and (6.8). You can check that they pick up only a phase

$$\mathcal{D}_t\psi \to e^{iq\alpha/\hbar}\,\mathcal{D}_t\psi \quad \text{and} \quad \mathcal{D}_i\psi \to e^{iq\alpha/\hbar}\,\mathcal{D}_i\psi$$

This ensures that the Schrödinger equation (6.9) transforms covariantly.

6.2 Landau Levels

Our task now is to solve for the spectrum and wavefunctions of the Schrödinger equation. We are interested in the situation with vanishing electric field, $\mathbf{E} = 0$, and constant magnetic field. The quantum Hamiltonian is

$$H = \frac{1}{2m}(\mathbf{p} - q\mathbf{A})^2 \tag{6.10}$$

We take the magnetic field to lie in the z-direction, so that $\mathbf{B} = (0, 0, B)$. To proceed, we need to find a gauge potential $\mathbf{A}$ which obeys $\nabla \times \mathbf{A} = \mathbf{B}$. There is, of course, no unique choice. Here we pick

$$\mathbf{A} = (0, xB, 0) \tag{6.11}$$

This is called *Landau gauge*. Note that the magnetic field $\mathbf{B} = (0, 0, B)$ is invariant under both translational symmetry and rotational symmetry in the (x, y)-plane. However, the choice of $\mathbf{A}$ is not; it breaks translational symmetry in the x direction (but not in the y direction) and rotational symmetry. This means that, while the physics will be invariant under all symmetries, the intermediate calculations will not be manifestly invariant. This kind of compromise is typical when dealing with magnetic field.

The Hamiltonian (6.10) becomes

$$H = \frac{1}{2m}\left(p_x^2 + (p_y - qBx)^2 + p_z^2\right)$$

Because we have manifest translational invariance in the y and z directions, we have $[p_y, H] = [p_z, H] = 0$ and can look for energy eigenstates which are also eigenstates of p_y and p_z. This motivates the ansatz

$$\psi(\mathbf{x}) = e^{ik_y y + ik_z z}\,\chi(x) \tag{6.12}$$

Acting on this wavefunction with the momentum operators $p_y = -i\hbar\partial_y$ and $p_z = -i\hbar\partial_z$, we have

$$p_y\psi = \hbar k_y \psi \quad \text{and} \quad p_z\psi = \hbar k_z \psi$$

The time-independent Schrödinger equation is $H\psi = E\psi$. Substituting our ansatz (6.12) simply replaces p_y and p_z with their eigenvalues, and we have

$$H\psi(\mathbf{x}) = \frac{1}{2m}\Big[p_x^2 + (\hbar k_y - qBx)^2 + \hbar^2 k_z^2\Big]\psi(\mathbf{x}) = E\psi(\mathbf{x})$$

We can write this as an eigenvalue equation for the equation $\chi(x)$. We have

$$\tilde{H}\chi(x) = \left(E - \frac{\hbar^2 k_z^2}{2m}\right)\chi(x)$$

where $\tilde{H}$ is something very familiar: it's the Hamiltonian for a harmonic oscillator in the x direction, with the centre displaced from the origin,

$$\tilde{H} = \frac{1}{2m}p_x^2 + \frac{m\omega_B^2}{2}(x - k_y l_B^2)^2 \tag{6.13}$$

The frequency of the harmonic oscillator is again the cyloctron frequency $\omega_B = qB/m$, and we've also introduced a length scale l_B. This is a characteristic length scale which governs any quantum phenomena in a magnetic field. It is called the *magnetic length*.

$$l_B = \sqrt{\frac{\hbar}{qB}}$$

To give you some sense for this, in a magnetic field of $B = 1$ *Tesla*, the magnetic length for an electron is $l_B \approx 2.5 \times 10^{-8}$ m.

Something rather strange has happened in the Hamiltonian (6.13): the momentum in the y direction, $\hbar k_y$, has turned into the position of the harmonic oscillator in the x direction, which is now centred at $x = k_y l_B^2$.

We can immediately write down the energy eigenvalues of (6.13); they are simply those of the harmonic oscillator

$$E = \hbar\omega_B \left(n + \frac{1}{2}\right) + \frac{\hbar^2 k_z^2}{2m} \qquad n = 0, 1, 2, \ldots \tag{6.14}$$

The wavefunctions depend on three quantum numbers, $n \in \mathbf{N}$ and $k_y, k_z \in \mathbf{R}$. They are

$$\psi_{n,k}(x, y) \sim e^{ik_y y + ik_z z}\, H_n(x - k_y l_B^2) e^{-(x - k_y l_B^2)^2/2l_B^2} \tag{6.15}$$

with H_n the usual Hermite polynomial wavefunctions of the harmonic oscillator. The $\sim$ reflects the fact that we have made no attempt to normalise these these wavefunctions.

The wavefunctions look like strips, extended in the y direction but exponentially localised around $x = k_y l_B^2$ in the x direction. However, you shouldn't read too much into this. As we will see shortly, there is large degeneracy of wavefunctions and by taking linear combinations of these states we can cook up wavefunctions that have pretty much any shape you like.

6.2.1 Degeneracy

The dynamics of the particle in the z-direction is unaffected by the magnetic field $\mathbf{B} = (0, 0, B)$. To focus on the novel physics, let's restrict to particles with $k_z = 0$. The energy spectrum then coincides with that of a harmonic oscillator,

$$E_n = \hbar\omega_B \left(n + \frac{1}{2}\right) \tag{6.16}$$

In the present context, these are called *Landau levels*. We see that, in the presence of a magnetic field, the energy levels of a particle become equally spaced, with the gap between each level proportional to the magnetic field B. Note that the energy spectrum looks very different from a free particle moving in the (x, y)-plane.

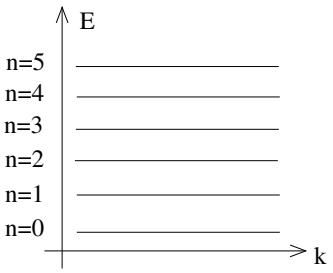


Figure 79: Landau Levels

The states in a given Landau level are not unique. Instead, there is a huge degeneracy, with many states having the same energy. We can see this in the form of the wavefunctions (6.15) which, when $k_z = 0$, depend on two quantum numbers, n and k_y. Yet the energy (6.16) is independent of k_y.

Let's determine how large this degeneracy of states is. To do so, we need to restrict ourselves to a finite region of the (x, y)-plane. We pick a rectangle with sides of lengths L_x and L_y. We want to know how many states fit inside this rectangle.

Having a finite size L_y is like putting the system in a box in the y-direction. The wavefunctions must obey

$$\psi(x, y + L_y, z) = \psi(x, y, z) \quad \Rightarrow \quad e^{ik_y L_y} = 1$$

This means that the momentum k_y is quantised in units of $2\pi/L_y$.

Having a finite size L_x is somewhat more subtle. The reason is that, as we mentioned above, the gauge choice (6.11) does not have manifest translational invariance in the x-direction. This means that our argument will be a little heuristic. Because the wavefunctions (6.15) are exponentially localised around $x = k_y l_B^2$, for a finite sample restricted to $0 \leq x \leq L_x$ we would expect the allowed k_y values to range between $0 \leq k_y \leq L_x/l_B^2$. The end result is that the number of states in each Landau level is given by

$$\mathcal{N} = \frac{L_y}{2\pi} \int_0^{L_x/l_B^2} dk = \frac{L_x L_y}{2\pi l_B^2} = \frac{qBA}{2\pi\hbar} \tag{6.17}$$

where $A = L_x L_y$ is the area of the sample. Strictly speaking, we should take the integer part of the answer above.

The degeneracy (6.17) is very very large. Throwing in some numbers, there are around 10^{10} degenerate states per Landau level for electrons in a region of area $A = 1\ cm^2$ in a magnetic field $B \sim 0.1\ T$. This large degeneracy ultimately, this leads to an array of dramatic and surprising physics.

6.2.2 Symmetric Gauge

It is worthwhile to repeat the calculations above using a different gauge choice. This will give us a slightly different perspective on the physics. A natural choice is *symmetric gauge*

$$\mathbf{A} = -\frac{1}{2}\mathbf{x} \times \mathbf{B} = \frac{B}{2}(-y, x, 0) \tag{6.18}$$

This choice of gauge breaks translational symmetry in both the x and the y directions. However, it does preserve rotational symmetry about the origin. This means that angular momentum is now a good quantum number to label states.

In this gauge, the Hamiltonian is given by

$$\begin{aligned} H &= \frac{1}{2m}\left[\left(p_x + \frac{qBy}{2}\right)^2 + \left(p_y - \frac{qBx}{2}\right)^2 + p_z^2\right] \\ &= -\frac{\hbar^2}{2m}\nabla^2 + \frac{qB}{2m}L_z + \frac{q^2B^2}{8m}(x^2 + y^2) \end{aligned} \tag{6.19}$$

where we've introduced the angular momentum operator

$$L_z = xp_y - yp_x$$

We'll again restrict to motion in the (x, y)-plane, so we focus on states with $k_z = 0$. It turns out that complex variables are particularly well suited to describing states in symmetric gauge, in particular in the lowest Landau level with $n = 0$. We define

$$w = x + iy \quad \text{and} \quad \bar{w} = x - iy$$

Correspondingly, the complex derivatives are

$$\partial = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right) \quad \text{and} \quad \bar{\partial} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)$$

which obey $\partial w = \bar{\partial}\bar{w} = 1$ and $\partial\bar{w} = \bar{\partial}w = 0$. The Hamiltonian, restricted to states with $k_z = 0$, is then given by

$$H = -\frac{2\hbar^2}{m}\partial\bar{\partial} - \frac{\omega_B}{2}L_z + \frac{m\omega_B^2}{8}w\bar{w}$$

where now

$$L_z = \hbar(w\partial - \bar{w}\bar{\partial})$$

It is simple to check that the states in the lowest Landau level take the form

$$\psi_0(w,\bar{w}) = f(w)e^{-|w|^2/4l_B^2}$$

for *any* holomorphic function $f(w)$. These all obey

$$H\psi_0(w,\bar{w}) = \frac{\hbar\omega_B}{2}\psi_0(w,\bar{w})$$

which is the statement that they lie in the lowest Landau level with $n = 0$. We can further distinguish these states by requiring that they are also eigenvalues of L_z. These are satisfied by the monomials,

$$\psi_0 = w^M e^{-|w|^2/4l_B^2} \quad \Rightarrow \quad L_z\psi_0 = \hbar M\psi_0 \tag{6.20}$$

for some positive integer M.

Degeneracy Revisited

In symmetric gauge, the profiles of the wavefunctions (6.20) form concentric rings around the origin. The higher the angular momentum M, the further out the ring. This, of course, is very different from the strip-like wavefunctions that we saw in Landau gauge (6.15). You shouldn't read too much into this other than the fact that the profile of the wavefunctions is not telling us anything physical as it is not gauge invariant.

However, it's worth revisiting the degeneracy of states in symmetric gauge. The wavefunction with angular momentum M is peaked on a ring of radius $r = \sqrt{2M}l_B$. This means that in a disc shaped region of area $A = \pi R^2$, the number of states is roughly (the integer part of)

$$\mathcal{N} = R^2/2l_B^2 = A/2\pi l_B^2 = \frac{qBA}{2\pi\hbar}$$

which agrees with our earlier result (6.17).

6.2.3 An Invitation to the Quantum Hall Effect

Take a system with some fixed number of electrons, which are restricted to move in the (x, y)-plane. The charge of the electron is $q = -e$. In the presence of a magnetic field, these will first fill up the $\mathcal{N} = eBA/2\pi\hbar$ states in the $n = 0$ lowest Landau level. If any are left over they will then start to fill up the $n = 1$ Landau level, and so on.

Now suppose that we increase the magnetic field B. The number of states $\mathcal{N}$ housed in each Landau level will increase, leading to a depletion of the higher Landau levels. At certain, very special values of B, we will find some number of Landau levels that are exactly filled. However, generically there will be a highest Landau level which is only partially filled.

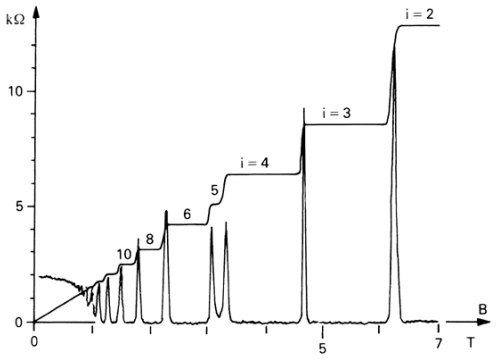


Figure 80: The integer quantum Hall effect.

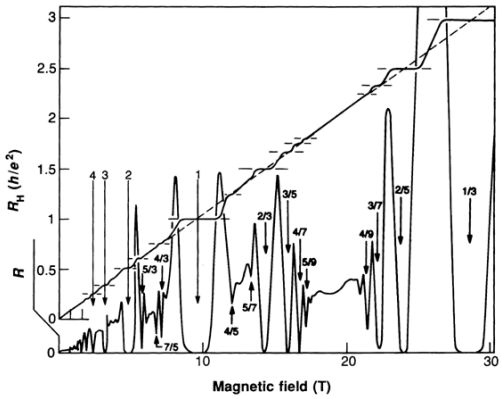


Figure 81: The fractional quantum Hall effect.

This successive depletion of Landau levels gives rise to a number of striking signatures in different physical quantities. Often these quantities oscillate, or jump discontinuously as the number of occupied Landau levels varies. One particular example is the de Haas van Alphen oscillations seen in the magnetic susceptibility which we describe in Section 4.3.4. Another example is the behaviour of the resistivity ρ. This relates the current density $\mathbf{J} = (J_x, J_y)$ to the applied electric field $\mathbf{E} = (E_x, E_y)$,

$$\mathbf{E} = \rho \mathbf{J}$$

In the presence of an applied magnetic field $\mathbf{B} = (0, 0, B)$, the electrons move in circles. This results in components of the current which are both parallel and perpendicular to the electric field. This is modelled straightforwardly by taking ρ to be a matrix

$$\rho = \begin{pmatrix} \rho_{xx} & \rho_{xy} \\ -\rho_{xy} & \rho_{xx} \end{pmatrix}$$

where the form of the matrix follows from rotational invariance. Here ρ_{xx} is called the *longitudinal resistivity* while ρ_{xy} is called the *Hall resistivity*.

In very clean samples, in strong magnetic fields, both components of the resistivity exhibit very surprising behaviour. This is shown in the left-hand figure above. The Hall resistivity ρ_{xy} increases with B by forming a series of plateaux, on which it takes values

$$\rho_{xy} = \frac{2\pi\hbar}{e^2}\frac{1}{\nu} \qquad \nu \in \mathbf{N}$$

The value of ν (which is labelled $i = 2, 3, \ldots$ in the data shown above) is measured to be an integer to extraordinary accuracy — around one part in 10^9. Meanwhile,

the longitudinal resistivity vanishes when ρ_{xy} lies on a plateaux, but spikes whenever there is a transition between different plateaux. This phenomenon, called the *integer Quantum Hall Effect*, was discovered by Klaus von Klitzing in 1980. For this, he was awarded the Nobel prize in 1985.

It turns out that the integer quantum Hall effect is a direct consequence of the existence of discrete Landau levels. The plateaux occur when precisely $\nu \in \mathbf{Z}^+$ Landau levels are filled. Of course, we're very used to seeing integers arising in quantum mechanics — this, after all, is what the "quantum" in quantum mechanics means. However, the quantisation of the resistivity ρ_{xy} is something of a surprise because this is a macroscopic quantity, involving the collective behaviour of many trillions of electrons, swarming through a hot and dirty system. A full understanding of the integer quantum Hall effect requires an appreciation of how the mathematics of topology fits in with quantum mechanics. David Thouless (and, to some extent, Duncan Haldane) were awarded the 2016 Nobel prize for understanding the underlying role of topology in this system.

Subsequently it was realised that similar behaviour also happens when Landau levels are partially filled. However, it doesn't occur for any filling, but only very special values. This is referred to as the *fractional quantum Hall effect*. The data is shown in the right-hand figure. You can see clear plateaux when the lowest Landau level has $\nu = \frac{1}{3}$ of its states filled. There is another plateaux when $\nu = \frac{2}{5}$ of the states are filled, followed by a bewildering pattern of further plateaux, all of which occur when ν is some rational number. This was discovered by Tsui and Störmer in 1982. It called the *Fractional Quantum Hall Effect*. The 1998 Nobel prize was awarded to Tsui and Stormer, together with Laughlin who pioneered the first theoretical ideas to explain this behaviour.

The fractional quantum Hall effect cannot be explained by treating the electrons as free. Instead, it requires us to take interactions into account. We have seen that each Landau level has a macroscopically large degeneracy. This degeneracy is lifted by interactions, resulting in a new form of quantum liquid which exhibits some magical properties. For example, in this state of matter the electron — which, of course, is an indivisible particle — can split into constituent parts! The $\nu = \frac{1}{3}$ state has excitations which carry 1/3 of the charge of an electron. In other quantum Hall states, the excitations have charge 1/5 or 1/4 of the electron. These particles also have a number of other, even stranger properties to do with their quantum statistics and there is hope that these may underly the construction of a quantum computer.

We will not delve into any further details of the quantum Hall effect. Suffice to say that it is one of the richest and most beautiful subjects in theoretical physics. You can find a fuller exploration of these ideas in the lecture notes devoted to the Quantum Hall Effect.

6.3 The Aharonov-Bohm Effect

In our course on Electromagnetism, we learned that the gauge potential A_μ is unphysical: the physical quantities that affect the motion of a particle are the electric and magnetic fields. Yet we've seen above that we cannot formulate quantum mechanics without introducing the gauge fields $\mathbf{A}$ and ϕ. This might lead us to wonder whether there is more to life than $\mathbf{E}$ and $\mathbf{B}$ alone. In this section we will see that things are, indeed, somewhat more subtle.

6.3.1 Particles Moving around a Flux Tube

Consider the set-up shown in the figure. We have a solenoid of area A, carrying magnetic field $\mathbf{B} = (0, 0, B)$ and therefore magnetic flux $\Phi = BA$. Outside the solenoid the magnetic field is zero. However, the vector potential is not. This follows from Stokes' theorem which tells us that the line integral outside the solenoid is given by

$$\oint \mathbf{A} \cdot d\mathbf{x} = \int \mathbf{B} \cdot d\mathbf{S} = \Phi$$

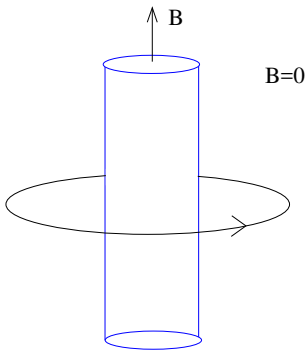


Figure 82:

This is simply solved in cylindrical polar coordinates by

$$A_\phi = \frac{\Phi}{2\pi r}$$

Now consider a charged quantum particle restricted to lie in a ring of radius r outside the solenoid. The only dynamical degree of freedom is the angular coordinate $\phi \in [0, 2\pi)$. The Hamiltonian is

$$H = \frac{1}{2m}(p_\phi - qA_\phi)^2 = \frac{1}{2mr^2}\left(-i\hbar\frac{\partial}{\partial\phi} - \frac{q\Phi}{2\pi}\right)^2$$

We'd like to see how the presence of this solenoid affects the particle. The energy eigenstates are simply

$$\psi = \frac{1}{\sqrt{2\pi r}}\, e^{in\phi} \qquad n \in \mathbf{Z} \tag{6.21}$$

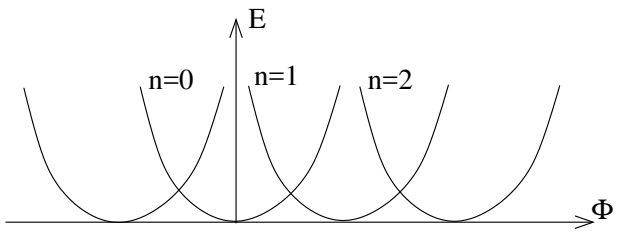


Figure 83: The energy spectrum for a particle moving around a solenoid.

where the requirement that ψ is single valued around the circle means that we must take $n \in \mathbf{Z}$. Plugging this into the time independent Schrödinger equation $H\psi = E\psi$, we find the spectrum

$$E = \frac{1}{2mr^2}\left(\hbar n - \frac{q\Phi}{2\pi}\right)^2 = \frac{\hbar^2}{2mr^2}\left(n - \frac{\Phi}{\Phi_0}\right)^2 \qquad n \in \mathbf{Z}$$

where we've defined the *quantum of flux* $\Phi_0 = 2\pi\hbar/q$. (Usually this quantum of flux is defined using the electron charge $q = -e$, with the minus signs massaged so that $\Phi_0 \equiv 2\pi\hbar/e > 0$.)

Note that if Φ is an integer multiple of Φ_0, then the spectrum is unaffected by the solenoid. But if the flux in the solenoid is not an integral multiple of Φ_0 — and there is no reason that it should be — then the spectrum gets shifted. We see that the energy of the particle knows about the flux Φ even though the particle never goes near the region with magnetic field. The resulting energy spectrum is shown in Figure 83.

There is a slightly different way of looking at this result. Away from the solenoid, the gauge field is a total divergence

$$\mathbf{A} = \nabla\alpha \quad \text{with} \quad \alpha = \frac{\Phi\phi}{2\pi}$$

This means that we can try to remove it by redefining the wavefunction to be

$$\psi \to \tilde{\psi} = \exp\left(\frac{-iq\alpha}{\hbar}\right)\psi = \exp\left(\frac{-iq\Phi}{2\pi\hbar}\phi\right)\psi$$

However, there is an issue: the wavefunction should be single-valued. This, after all, is how we got the quantisation condition $n \in \mathbf{Z}$ in (6.21). This means that the gauge transformation above is allowed only if Φ is an integer multiple of $\Phi_0 = 2\pi\hbar/q$. Only in this case is the particle unaffected by the solenoid. The obstacle arises from the fact that the wavefunction of the particle winds around the solenoid. We see here the first glimpses of how topology starts to feed into quantum mechanics.

There are a number of further lessons lurking in this simple quantum mechanical set-up. You can read about them in the lectures on the Quantum Hall Effect (see Section 1.5.3) and the lectures on Gauge Theory (see Section 3.6.1).

6.3.2 Aharonov-Bohm Scattering

The fact that a quantum particle can be affected by $\mathbf{A}$ even when restricted to regions where $\mathbf{B} = 0$ was first pointed out by Aharonov and Bohm in a context which is closely related to the story above. They revisited the famous double-slit experiment, but now with a twist: a solenoid carrying flux Φ is hidden behind the wall. This set-up is shown in the figure below. Once again, the particle is forbidden from going near the solenoid. Nonetheless, the presence of the magnetic flux affects the resulting interference pattern, shown as the dotted line in the figure.

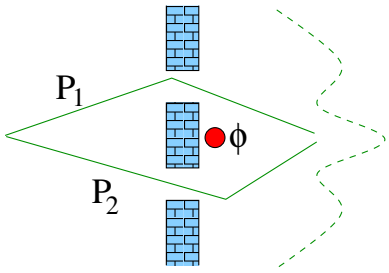


Figure 84:

Consider a particle that obeys the free Schrödinger equation,

$$\frac{1}{2m}\Big(-i\hbar\nabla - q\mathbf{A}\Big)^2\psi = E\psi$$

We can formally remove the gauge field by writing

$$\psi(\mathbf{x}) = \exp\left(\frac{iq}{\hbar}\int^{\mathbf{x}} \mathbf{A}(\mathbf{x}')\cdot d\mathbf{x}'\right)\phi(\mathbf{x})$$

where the integral is over any path. Crucially, however, in the double-slit experiment there are two paths, P_1 and P_2. The phase picked up by the particle due to the gauge field differs depending on which path is taken. The phase difference is given by

$$\Delta\theta = \frac{q}{\hbar}\int_{P_1}\mathbf{A}\cdot d\mathbf{x} - \frac{q}{\hbar}\int_{P_2}\mathbf{A}\cdot d\mathbf{x} = \frac{q}{\hbar}\oint\mathbf{A}\cdot d\mathbf{x} = \frac{q}{\hbar}\int\mathbf{B}\cdot d\mathbf{S}$$

Note that neither the phase arising from path P_1, nor the phase arising from path P_2, is gauge invariant. However, the difference between the two phases is gauge invariant. As we see above, it is given by the flux through the solenoid. This is the Aharonov-Bohm phase, $e^{iq\Phi/\hbar}$, an extra contribution that arises when charged particles move around magnetic fields.

The Aharonov-Bohm phase manifests in the interference pattern seen on the screen. As Φ is changed, the interference pattern shifts, an effect which has been experimentally observed. Only when Φ is an integer multiple of Φ_0 is the particle unaware of the presence of the solenoid.

6.4 Magnetic Monopoles

A *magnetic monopole* is a hypothetical object which emits a radial magnetic field of the form

$$\mathbf{B} = \frac{g\hat{\mathbf{r}}}{4\pi r^2} \quad \Rightarrow \quad \int d\mathbf{S}\cdot\mathbf{B} = g \tag{6.22}$$

Here g is called the *magnetic charge*.

We learned in our first course on Electromagnetism that magnetic monopoles don't exist. First, and most importantly, they have never been observed. Second there's a law of physics which insists that they can't exist. This is the Maxwell equation

$$\nabla\cdot\mathbf{B} = 0$$

Third, this particular Maxwell equation would appear to be non-negotiable. This is because it follows from the definition of the magnetic field in terms of the gauge field

$$\mathbf{B} = \nabla\times\mathbf{A} \quad \Rightarrow \quad \nabla\cdot\mathbf{B} = 0$$

Moreover, as we've seen above, the gauge field $\mathbf{A}$ is necessary to describe the quantum physics of particles moving in magnetic fields. Indeed, the Aharonov-Bohm effect tells us that there is non-local information stored in $\mathbf{A}$ that can only be detected by particles undergoing closed loops. All of this points to the fact that we would be wasting our time discussing magnetic monopoles any further.

Happily, there is a glorious loophole in all of these arguments, first discovered by Dirac, and magnetic monopoles play a crucial role in our understanding of the more subtle effects in gauge theories. The essence of this loophole is that there is an ambiguity in how we define the gauge potentials. In this section, we will see how this arises.

6.4.1 Dirac Quantisation

It turns out that not any magnetic charge g is compatible with quantum mechanics. Here we present several different arguments for the allowed values of g.

We start with the simplest and most physical of these arguments. Suppose that a particle with charge q moves along some closed path C in the background of some gauge potential $\mathbf{A}(\mathbf{x})$. Then, upon returning to its initial starting position, the wavefunction of the particle picks up a phase

$$\psi \to e^{iq\alpha/\hbar}\psi \quad \text{with} \quad \alpha = \oint_C \mathbf{A}\cdot d\mathbf{x} \tag{6.23}$$

This is the Aharonov-Bohm phase described above.

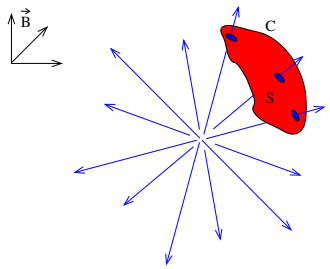


Figure 85: Integrating over S...

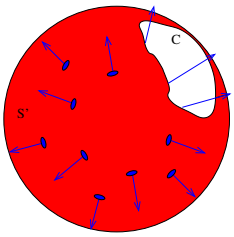


Figure 86: ...or over S'.

The phase of the wavefunction is not an observable quantity in quantum mechanics. However, as we described above, the phase in (6.23) is really a *phase difference*. We could, for example, place a particle in a superposition of two states, one of which stays still while the other travels around the loop C. The subsequent interference will depend on the phase $e^{iq\alpha/\hbar}$, just like in the Aharonov-Bohm effect.

Let's now see what this has to do with magnetic monopoles. We place our particle, with electric charge q, in the background of a magnetic monopole with magnetic charge g. We keep the magnetic monopole fixed, and let the electric particle undergo some journey along a path C. We will ask only that the path C avoids the origin where the magnetic monopole is sitting. This is shown in the left-hand panel of the figure. Upon returning, the particle picks up a phase $e^{iq\alpha/\hbar}$ with

$$\alpha = \oint_C \mathbf{A} \cdot d\mathbf{x} = \int_S \mathbf{B} \cdot d\mathbf{S}$$

where, as shown in the figure, S is the area enclosed by C. Using the fact that $\int_{\mathbf{S}^2} \mathbf{B} \cdot d\mathbf{S} = g$, if the surface S makes a solid angle Ω, this phase can be written as

$$\alpha = \frac{\Omega g}{4\pi}$$

However, there's an ambiguity in this computation. Instead of integrating over S, it is equally valid to calculate the phase by integrating over S', shown in the right-hand panel of the figure. The solid angle formed by S' is $\Omega' = 4\pi - \Omega$. The phase is then given by

$$\alpha' = -\frac{(4\pi - \Omega)g}{4\pi}$$

where the overall minus sign comes because the surface S' has the opposite orientation to S. As we mentioned above, the phase shift that we get in these calculations is

observable: we can't tolerate different answers from different calculations. This means that we must have $e^{iq\alpha/\hbar} = e^{iq\alpha'/\hbar}$. This gives the condition

$$qg = 2\pi\hbar n \quad \text{with } n \in \mathbf{Z} \tag{6.24}$$

This is the famous Dirac quantisation condition. The smallest such magnetic charge has $n = 1$. It coincides with the quantum of flux, $g = \Phi_0 = 2\pi\hbar/q$.

Above we worked with a single particle of charge q. Obviously, the same argument must hold for any other particle of charge q'. There are two possibilities. The first is that all particles carry charge that is an integer multiple of some smallest unit. In this case, it's sufficient to impose the Dirac quantisation condition (6.24) where q is the smallest unit of charge. For example, in our world we should take $q = \pm e$ to be the electron or proton charge (or, if we look more closely in the Standard Model, we might choose to take $q = -e/3$, the charge of the down quark).

The second possibility is that the particles carry electric charges which are irrational multiples of each other. For example, there may be a particle with charge q and another particle with charge $\sqrt{2}q$. In this case, no magnetic monopoles are allowed.

It's sometimes said that the existence of a magnetic monopole would imply the quantisation of electric charges. This, however, has it backwards. (It also misses the point that we have a wonderful explanation of the quantisation of charges from the story of anomaly cancellation in the Standard Model.) There are two possible groups that could underly gauge transformations in electromagnetism. The first is $U(1)$; this has integer valued charges and admits magnetic monopoles. The second possibility is $\mathbf{R}$; this has irrational electric charges and forbids monopoles. All the evidence in our world points to the fact that electromagnetism is governed by $U(1)$ and that magnetic monopoles should exist.

Above we looked at an electrically charged particle moving in the background of a magnetically charged particle. It is simple to generalise the discussion to particles that carry both electric and magnetic charges. These are called *dyons*. For two dyons, with charges (q_1, g_1) and (q_2, g_2), the generalisation of the Dirac quantisation condition requires

$$q_1 g_2 - q_2 g_1 \in 2\pi\hbar\mathbf{Z}$$

This is sometimes called the *Dirac-Zwanziger* condition.

6.4.2 A Patchwork of Gauge Fields

The discussion above shows how quantum mechanics constrains the allowed values of magnetic charge. It did not, however, address the main obstacle to constructing a magnetic monopole out of gauge fields $\mathbf{A}$ when the condition $\mathbf{B} = \nabla \times \mathbf{A}$ would seem to explicitly forbid such objects.

Let's see how to do this. Our goal is to write down a configuration of gauge fields which give rise to the magnetic field (6.22) of a monopole which we will place at the origin. However, we will need to be careful about what we want such a gauge field to look like.

The first point is that we won't insist that the gauge field is well defined at the origin. After all, the gauge fields arising from an electron are not well defined at the position of an electron and it would be churlish to require more from a monopole. This fact gives us our first bit of leeway, because now we need to write down gauge fields on $\mathbf{R}^3/\{0\}$, as opposed to $\mathbf{R}^3$ and the space with a point cut out enjoys some non-trivial topology that we will make use of.

Consider the following gauge connection, written in spherical polar coordinates

$$A^N_\phi = \frac{g}{4\pi r}\frac{1-\cos\theta}{\sin\theta} \tag{6.25}$$

The resulting magnetic field is

$$\mathbf{B} = \nabla \times \mathbf{A} = \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(A^N_\phi \sin\theta)\,\hat{\mathbf{r}} - \frac{1}{r}\frac{\partial}{\partial r}(rA^N_\phi)\hat{\boldsymbol{\theta}}$$

Substituting in (6.25) gives

$$\mathbf{B} = \frac{g\hat{\mathbf{r}}}{4\pi r^2} \tag{6.26}$$

In other words, this gauge field results in the magnetic monopole. But how is this possible? Didn't we learn in kindergarten that if we can write $\mathbf{B} = \nabla \times \mathbf{A}$ then $\int d\mathbf{S}\cdot\mathbf{B} = 0$? How does the gauge potential (6.25) manage to avoid this conclusion?

The answer is that $\mathbf{A}^N$ in (6.25) is actually a singular gauge connection. It's not just singular at the origin, where we've agreed this is allowed, but it is singular along an entire half-line that extends from the origin to infinity. This is due to the $1/\sin\theta$ term which diverges at $\theta = 0$ and $\theta = \pi$. However, the numerator $1-\cos\theta$ has a zero when $\theta = 0$ and the gauge connection is fine there. But the singularity along the half-line $\theta = \pi$ remains. The upshot is that this gauge connection is not acceptable along the line of the south pole, but is fine elsewhere. This is what the superscript N is there to remind us: we can work with this gauge connection s long as we keep north.

Now consider a different gauge connection

$$A_\phi^S = -\frac{g}{4\pi r}\frac{1+\cos\theta}{\sin\theta} \tag{6.27}$$

This again gives rise to the magnetic field (6.26). This time it is well behaved at $\theta = \pi$, but singular at the north pole $\theta = 0$. The superscript S is there to remind us that this connection is fine as long as we keep south.

At this point, we make use of the ambiguity in the gauge connection. We are going to take $\mathbf{A}^N$ in the northern hemisphere and $\mathbf{A}^S$ in the southern hemisphere. This is allowed because the two gauge potentials are the same up to a gauge transformation, $\mathbf{A} \to \mathbf{A} + \nabla\alpha$. Recalling the expression for $\nabla\alpha$ in spherical polars, we find that for $\theta \neq 0, \pi$, we can indeed relate A_ϕ^N and A_ϕ^S by a gauge transformation,

$$A_\phi^N = A_\phi^S + \frac{1}{r\sin\theta}\,\partial_\phi\alpha \quad \text{where} \quad \alpha = \frac{g\phi}{2\pi} \tag{6.28}$$

However, there's still a question remaining: is this gauge transformation allowed? The problem is that the function α is not single valued: $\alpha(\phi = 2\pi) = \alpha(\phi = 0) + g$. And this should concern us because, as we've seen in (6.8), the gauge transformation also acts on the wavefunction of a quantum particle

$$\psi \to e^{iq\alpha/\hbar}\psi$$

There's no reason that we should require the gauge transformation α to be single-valued, but we do want the wavefunction ψ to be single-valued. This holds for the gauge transformation (6.28) provided that we have

$$qg = 2\pi\hbar n \quad \text{with } n \in \mathbf{Z}$$

This, of course, is the Dirac quantisation condition (6.24).

Mathematically, we have constructed of a topologically non-trivial $U(1)$ bundle over the $\mathbf{S}^2$ surrounding the origin. In this context, the integer n is called the first Chern number.

6.4.3 Monopoles and Angular Momentum

Here we provide yet another derivation of the Dirac quantisation condition, this time due to Saha. The key idea is that the quantisation of magnetic charge actually follows from the more familiar quantisation of angular momentum. The twist is that, in the presence of a magnetic monopole, angular momentum isn't quite what you thought.

To set the scene, let's go back to the Lorentz force law

$$\frac{d\mathbf{p}}{dt} = q\,\dot{\mathbf{x}} \times \mathbf{B}$$

with $\mathbf{p} = m\dot{\mathbf{x}}$. Recall from our discussion in Section 6.1.1 that $\mathbf{p}$ defined here is not the canonical momentum, a fact which is hiding in the background in the following derivation. Now let's consider this equation in the presence of a magnetic monopole, with

$$\mathbf{B} = \frac{g}{4\pi}\frac{\mathbf{r}}{r^3}$$

The monopole has rotational symmetry so we would expect that the angular momentum, $\mathbf{x} \times \mathbf{p}$, is conserved. Let's check:

$$\begin{aligned}\frac{d(\mathbf{x}\times\mathbf{p})}{dt} &= \dot{\mathbf{x}}\times\mathbf{p} + \mathbf{x}\times\dot{\mathbf{p}} = \mathbf{x}\times\dot{\mathbf{p}} = q\mathbf{x}\times(\dot{\mathbf{x}}\times\mathbf{B}) \\ &= \frac{qg}{4\pi r^3}\,\mathbf{x}\times(\dot{\mathbf{x}}\times\mathbf{x}) = \frac{qg}{4\pi}\left(\frac{\dot{\mathbf{x}}}{r} - \frac{\dot{r}\mathbf{x}}{r^2}\right) \\ &= \frac{d}{dt}\left(\frac{qg}{4\pi}\hat{\mathbf{r}}\right)\end{aligned}$$

We see that in the presence of a magnetic monopole, the naive angular momentum $\mathbf{x} \times \mathbf{p}$ is not conserved! However, as we also noticed in the lectures on Classical Dynamics (see Section 4.3.2), we can easily write down a modified angular momentum that is conserved, namely

$$\mathbf{L} = \mathbf{x}\times\mathbf{p} - \frac{qg}{4\pi}\hat{\mathbf{r}}$$

The extra term can be thought of as the angular momentum stored in $\mathbf{E} \times \mathbf{B}$. The surprise is that the system has angular momentum even when the particle doesn't move.

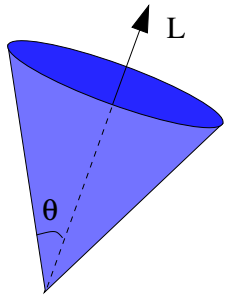


Figure 87:

Before we move on, there's a nice and quick corollary that we can draw from this. The angular momentum vector $\mathbf{L}$ does not change with time. But the angle that the particle makes with this vector is

$$\mathbf{L}\cdot\hat{\mathbf{r}} = -\frac{qg}{4\pi} = \text{constant}$$

This means that the particle moves on a cone, with axis $\mathbf{L}$ and angle $\cos\theta = -qg/4\pi L$.

So far, our discussion has been classical. Now we invoke some simple quantum mechanics: the angular momentum should be quantised. In particular, the angular momentum in the z-direction should be $L_z \in \frac{1}{2}\hbar\mathbf{Z}$. Using the result above, we have

$$\frac{qg}{4\pi} = \frac{1}{2}\hbar n \quad \Rightarrow \quad qg = 2\pi\hbar n \qquad \text{with } n \in \mathbf{Z}$$

Once again, we find the Dirac quantisation condition.

6.5 Spin in a Magnetic Field

As we've seen in previous courses, particles often carry an intrinsic angular momentum called *spin* $\mathbf{S}$. This spin is quantised in half-integer units. For examples, electrons have spin $\frac{1}{2}$ and their spin operator is written in terms of the Pauli matrices $\boldsymbol{\sigma}$,

$$\mathbf{S} = \frac{\hbar}{2}\boldsymbol{\sigma}$$

Importantly, the spin of any particle couples to a background magnetic field $\mathbf{B}$. The key idea here is that the intrinsic spin acts like a magnetic moment $\mathbf{m}$ which couples to the magnetic field through the Hamiltonian

$$H = -\mathbf{m}\cdot\mathbf{B}$$

The question we would like to answer is: what magnetic moment $\mathbf{m}$ should we associate with spin?

A full answer to this question would require an extended detour into the Dirac equation. Here we provide only some basic motivation. First consider a particle of charge q moving with velocity $\mathbf{v}$ around a circle of radius $\mathbf{r}$ as shown in the figure. From our lectures on Electromagnetism, we know that the associated magnetic moment is given by

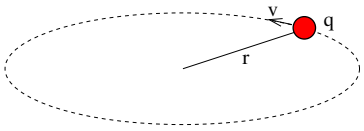


Figure 88:

$$\mathbf{m} = -\frac{q}{2}\mathbf{r}\times\mathbf{v} = \frac{q}{2m}\mathbf{L}$$

where $\mathbf{L} = m\mathbf{r}\times\mathbf{v}$ is the orbital angular momentum of the particle. Indeed, we already saw the resulting coupling $H = -(q/2m)\mathbf{L}\cdot\mathbf{B}$ in our derivation of the Hamiltonian in symmetric gauge (6.19).

Since the spin of a particle is another contribution to the angular momentum, we might anticipate that the associated magnetic moment takes the form

$$\mathbf{m} = g\frac{q}{2m}\mathbf{S}$$

where g is some dimensionless number. (Note: g is unrelated to the magnetic charge that we discussed in the previous section!) This, it turns out, is the right answer. However, the value of g depends on the particle under consideration. The upshot is that we should include a term in the Hamiltonian of the form

$$H = -g\frac{q}{2m}\mathbf{S}\cdot\mathbf{B} \tag{6.29}$$

The g-factor

For fundamental particles with spin $\frac{1}{2}$ — such as the electron — there is a long and interesting history associated to determining the value of g. For the electron, this was first measured experimentally to be

$$g_e = 2$$

Soon afterwards, Dirac wrote down his famous relativistic equation for the electron. One of its first successes was the theoretical prediction $g_e = 2$ for any spin $\frac{1}{2}$ particle. This means, for example, that the neutrinos and quarks also have $g = 2$.

This, however, was not the end of the story. With the development of quantum field theory, it was realised that there are corrections to the value $g_e = 2$. These can be calculated and take the form of a series expansion, starting with

$$g_e = 2\left(1 + \frac{\alpha}{2\pi} + \ldots\right) \approx 2.00232$$

where $\alpha = e^2/4\pi\epsilon_0\hbar c \approx 1/137$ is the dimensionless *fine structure constant* which characterises the strength of the Coulomb force. The most accurate experimental measurement of the electron magnetic moment now yields the result

$$g_e \approx 2.00231930436182 \pm 2.6 \times 10^{-13}$$

Theoretical calculations agree to the first ten significant figures or so. This is the most impressive agreement between theory and experiment in all of science! Beyond that, the value of α is not known accurately enough to make a comparison. Indeed, now the measurement of the electron magnetic moment is used to *define* the fine structure constant α.

While all fundamental spin $\frac{1}{2}$ particles have $g \approx 2$, this does not hold for more complicated objects. For example, the proton has

$$g_p \approx 5.588$$

while the neutron — which of course, is a neutral particle, but still carries a magnetic moment — has

$$g_n \approx -3.823$$

where, because the neutron is neutral, the charge $q = e$ is used in the formula (6.29). These measurements were one of the early hints that the proton and neutron are composite objects.

6.5.1 Spin Precession

Consider a constant magnetic field $\mathbf{B} = (0, 0, B)$. We would like to understand how this affects the spin of an electron. We'll take $g_e = 2$. We write the electric charge of the electron as $q = -e$ so the Hamiltonian is

$$H = \frac{e\hbar}{2m}\boldsymbol{\sigma} \cdot \mathbf{B}$$

The eigenstates are simply the spin-up $|\uparrow\rangle$ and spin-down $|\downarrow\rangle$ states in the z-direction. They have energies

$$H|\uparrow\rangle = \frac{\hbar\omega_B}{2}|\uparrow\rangle \quad \text{and} \quad H|\downarrow\rangle = -\frac{\hbar\omega_B}{2}|\downarrow\rangle$$

where $\omega_B = eB/m$ is the cyclotron frequency which appears throughout this chapter.

What happens if we do not sit in an energy eigenstate. A general spin state can be expressed in spherical polar coordinates as

$$|\psi(\theta, \phi)\rangle = \cos(\theta/2)|\uparrow\rangle + e^{i\phi}\sin(\theta/2)|\downarrow\rangle$$

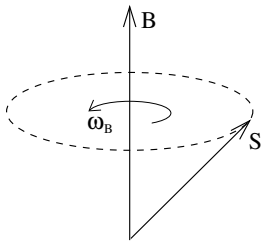


Figure 89:

As a check, note that $|\psi(\theta = \pi/2, \phi)\rangle$ is an eigenstate of σ^x when $\phi = 0, \pi$ and an eigenstate of σ^y when $\phi = \pi/2, 3\pi/2$ as it should be. The evolution of this state is determined by the time-dependent Schrödinger equation

$$i\hbar\frac{\partial|\psi\rangle}{\partial t} = H|\psi\rangle$$

which is easily solved to give

$$|\psi(\theta,\phi;t)\rangle = e^{i\omega_B t/2}\Big[\cos(\theta/2)|\uparrow\rangle + e^{i(\phi-\omega_B t)}\sin(\theta/2)|\downarrow\rangle\Big]$$

We see that the effect of the magnetic field is to cause the spin to precess about the **B** axis, as shown in the figure.

6.5.2 A First Look at the Zeeman Effect

The Zeeman effect describes the splitting of atomic energy levels in the presence of a magnetic field. Consider, for example, the hydrogen atom with Hamiltonian

$$H = -\frac{\hbar^2}{2m}\nabla^2 - \frac{1}{4\pi\epsilon_0}\frac{e^2}{r}$$

The energy levels are given by

$$E_n = -\frac{\alpha^2 mc^2}{2}\frac{1}{n^2} \qquad n \in \mathbf{Z}$$

where α is the fine structure constant. Each energy level has a degeneracy of states. These are labelled by the angular momentum $l = 0, 1, \ldots, n-1$ and the z-component of angular momentum $m_l = -l, \ldots, +l$. Furthermore, each electron carries one of two spin states labelled by $m_s = \pm\frac{1}{2}$. This results in a degeneracy given by

$$\text{Degeneracy} = 2\sum_{l=0}^{n-1}(2l+1) = 2n^2$$

Now we add a magnetic field $\mathbf{B} = (0, 0, B)$. As we have seen, this results in perturbation to the Hamiltonian which, to leading order in B, is given by

$$\Delta H = \frac{e}{2m}(\mathbf{L} + g_e\mathbf{S})\cdot\mathbf{B}$$

In the presence of such a magnetic field, the degeneracy of the states is split. The energy levels now depend on the quantum numbers n, m_l and m_s and are given by

$$E_{n,m,s} = E_n + \frac{e}{2m}(m_l + 2m_s)B$$

The Zeeman effect is developed further in the Lectures on Topics in Quantum Mechanics.