Lecture Notes for Solid State Physics
(3rd Year Course 6)
Hilary Term 2011

Professor Steven H. Simon
Oxford University

January 8, 2011

When I was an undergraduate, I thought solid state physics (a sub-genre of condensed matter physics) was perhaps the worst course that any undergraduate could be forced to learn – boring and tedious, “squalid state” as it was commonly called[1]. How much would I really learn about the universe by studying the properties of crystals? I managed to avoid taking this course altogether. My opinion at the time was not a reflection of the subject matter, but rather was a reflection of how solid state physics was taught.

Given my opinion as an undergraduate, it is a bit ironic that I have become a condensed matter physicist. But once I was introduced to the subject properly, I found that condensed matter was my favorite subject in all of physics – full of variety, excitement, and deep ideas. Many many physicists have come to this same conclusion. In fact, condensed matter physics is by far the largest single subfield of physics (the annual meeting of condensed matter physicists in the United States attracts over 6000 physicists each year!). Sadly a first introduction to the topic can barely scratch the surface of what constitutes the broad field of condensed matter.

Last year when I was told that a new course was being prepared to teach condensed matter physics to third year Oxford undergraduates, I jumped at the opportunity to teach it. I felt that it *must* be possible to teach a condensed matter physics course that is just as interesting and exciting as any other course that an undergraduate will ever take. It must be possible to convey the excitement of real condensed matter physics to the undergraduate audience. I hope I will succeed in this task. You can judge for yourself.

The topics I was asked to cover (being given little leeway in choosing the syllabus) are not atypical for a solid state physics course. In fact, the new condensed matter syllabus is extremely similar to the old Oxford B2 syllabus – the main changes being the removal of photonics and device physics. A few other small topics, such as superconductivity, are also nonexaminable now, or are removed altogether (such as point group symmetries). A few other topics (thermal expansion, chemical bonding) are now added by mandate of the IOP[2].

At any rate, the changes to the old B2 syllabus are generally minor, so I recommend that Oxford students use the old B2 exams as a starting point for figuring out what it is they need to study as the exams approach. In fact, I have used precisely these old exams to figure out what I need to teach. Being that the same group of people will be setting the exams this year as set them last year, this seems like a good idea. As with most exams at Oxford, one starts to see patterns in terms of what type of questions are asked year after year. The lecture notes contained here are designed to cover exactly this crucial material. I realize that these notes are a lot of material, and for this I apologize. However, this is the *minimum* set of notes that covers all of the topics that have shown up on old B2 exams. The actual lectures for this course will try to cover everything in these notes, but a few of the less crucial pieces will necessarily be glossed over in the interest of time.

Many of these topics are covered well in standard solid state physics references that one might find online, or in other books. The reason I am giving these lectures (and not just telling students to go read a standard book) is because condensed matter/solid-state is an enormous subject — worth many years of lectures — and one needs a guide to decide what subset of topics are most important (at least in the eyes of the examination committee). I believe that the lectures contained here give depth in some topics, and gloss over other topics, so as to reflect the particular topics that are deemed important at Oxford. These topics may differ a great deal from what is

---

[1] This jibe against solid state physics can be traced back to the Nobel Laureate Murray Gell-Mann, discoverer of the quark, who famously believed that there was nothing interesting in any endeavor but particle physics. Interestingly he now studies complexity — a field that mostly arose from condensed matter.

[2] We can discuss elsewhere whether or not we should pay attention to such mandates in general – although these particular mandates do not seem so odious.

deemed important elsewhere. In particular, Oxford is extremely heavy on scattering theory (x-ray and neutron diffraction) compared with most solid state courses or books that I have seen. But on the other hand, Oxford does not appear to believe in group representations (which resulted in my elimination of point group symmetries from the syllabus).

I cannot emphasize enough that there are many many extremely good books on solid-state and condensed matter physics already in existence. There are also many good resources online (including the rather infamous "Britney Spears' guide to semiconductor physics" — which is tongue-in-cheek about Britney Spears[3], but actually is a very good reference about semiconductors). I will list here some of the books that I think are excellent, and throughout these lecture notes, I will try to point you to references that I think are helpful.

- **States of Matter** by David L. Goodstein, Dover Publishing
  Chapter 3 of this book is a very brief but well written and easy to read description of much of what we will need to cover (but not all, certainly). The book is also published by Dover which means it is super-cheap in paperback. Warning: It uses cgs units rather than SI units, which is a bit annoying.
- **Solid State Physics, 2nd ed** by J. R. Hook and H. E. Hall
  This is frequently the book that students like the most. It is a first introduction to the subject and is much more introductory than Ashcroft and Mermin.
- **The Solid State**, H M Rosenberg (OUP)
  This slightly more advanced book was written a few decades ago to cover what was the solid state course at Oxford at that time. Some parts of the course have since changed, but other parts are well covered in this book.
- **Solid-State Physics, 4ed**, by H. Ibach and H. Luth, Springer-Verlag
  Another very popular book on the subject, with quite a bit of information in it. More advanced than Hook and Hall
- **Solid State Physics** by N. W. Ashcroft and D. N. Mermin, Holt-Sanders publishing
  This is the standard complete introduction to solid state physics. It has many many chapters on topics we won't be studying, and goes into great depth on almost everything. It may be a bit overwhelming to try to use this as a reference because of information-overload, but it has good explanations of almost everything. On the whole, this is my favorite reference. Warning: Also uses cgs units.
- **Introduction to Solid State Physics, 8ed**, by Charles Kittel[4]
  This is a classic text. It gets mixed reviews by some as being unclear on many matters. It is somewhat more complete than Hooke and Hall, less so than Ashcroft and Mermin. Its selection of topics and organization may seem a bit strange in the modern era.
- **The Basics of Crystallography and Diffraction, 3ed**, by C Hammond (OUP)
  This book has historically been part of the syllabus, particularly for the scattering theory part of the course. I don't like it as much.
- **Structure and Dynamics**, by M.T. Dove, Oxford University Press
  This is a more advanced book that covers scattering in particular. It is used in the Condensed Matter option 4-th year course.

---

[3] This guide was written when Ms. Spears was just a popular young performer and not the complete train wreck that she appears to be now.

[4] Kittel happens to be my dissertation-supervisor's dissertation-supervisor's dissertation-supervisor's dissertation-supervisor, for whatever that is worth.

- **Magnetism in Condensed Matter**, by Stephen Blundell, Oxford University Press
  Well written advanced material on the magnetism part of the course. It is used in the Condensed Matter option 4-th year course.
- **Band Theory and Electronic Properties of Solids**, by John Singleton
  More advanced material on electrons in solids. Also used in the Condensed Matter option 4-th year course.
- **Solid State Physics**, by G. Burns, Academic
  Another more advanced book. Some of its descriptions are short but very good.

I will remind my reader that these notes are a first draft. I apologize that they do not cover the material uniformly. In some places I have given more detail than in others – depending mainly on my enthusiasm-level at the particular time of writing. I hope to go back and improve the quality as much as possible. Updated drafts will hopefully be appearing.

Perhaps this pile of notes will end up as a book, perhaps they will not. This is not my point. My point is to write something that will be helpful for this course. If you can think of ways that these notes could be improved (correction of errors or whatnot) please let me know. The next generation of students will certainly appreciate it and that will improve your Karma. ☺

Oxford, United Kingdom
January, 2011.

Acknowledgements

Needless to say, I pilfered a fair fraction of the content of this course from parts of other books (mostly mentioned above). The authors of these books put great thought and effort into their writing. I am deeply indebted to these giants who have come before me. Additionally, I have stolen many ideas about how this course should be taught from the people who have taught the course (and similar courses) at Oxford in years past. Most recently this includes Mike Glazer, Andrew Boothroyd, and Robin Nicholas.

I am also very thankful for all the people who have helped me proofread, correct, and otherwise tweak these notes and the homework problems. These include in particular Mike Glazer, Alex Hearmon, and Simon Davenport.

Finally, I thank my father for helping proofread and improve these notes... and for a million other things.

Contents

Chapter 1

About Condensed Matter Physics

This chapter is just my personal take on why this topic is interesting. It seems unlikely to me that any exam would ask you why you study this topic, so you should probably consider this section to be not examinable. Nonetheless, you might want to read it to figure out why you should think this course is interesting if that isn't otherwise obvious.

1.1 What is Condensed Matter Physics

Quoting Wikipedia:

> Condensed matter physics is the field of physics that deals with the macroscopic and microscopic physical properties of matter. In particular, it is concerned with the "condensed" phases that appear whenever the number of constituents in a system is extremely large and the interactions between the constituents are strong. The most familiar examples of condensed phases are solids and liquids, which arise from the electromagnetic forces between atoms.

The use of the term "Condensed Matter" being more general than just solid state was coined and promoted by Nobel-Laureate Philip W. Anderson.

1.2 Why do we study Condensed Matter Physics?

There are several very good answers to this question

1. **Because it is the world around us**
   Almost all of the physical world that we see is in fact condensed matter. Questions such as
   - why are metals shiny and why do they feel cold?
   - why is glass transparent?

- why is water a fluid, and why does fluid feel wet?
- why is rubber soft and stretchy?

These questions are all in the domain of condensed matter physics. In fact almost every question you might ask about the world around you, short of asking about the sun or stars, is probably related to condensed matter physics in some way.

2. **Because it is useful**

   Over the last century our command of condensed matter physics has enabled us humans to do remarkable things. We have used our knowledge of physics to engineer new materials and exploit their properties to change our world and our society completely. Perhaps the most remarkable example is how our understanding of solid state physics enabled new inventions exploiting semiconductor technology, which enabled the electronics industry, which enabled computers, iPhones, and everything else we now take for granted.

3. **Because it is deep**

   The questions that arise in condensed matter physics are as deep as those you might find anywhere. In fact, many of the ideas that are now used in other fields of physics can trace their origins to condensed matter physics.

   A few examples for fun:

   - The famous Higgs boson, which the LHC is searching for, is no different from a phenomenon that occurs in superconductors (the domain of condensed matter physicists). The Higgs mechanism, which gives mass to elementary particles is frequently called the "Anderson-Higgs" mechanism, after the condensed matter physicist Phil Anderson (the same guy who coined the term "condensed matter") who described much of the same physics before Peter Higgs, the high energy theorist.
   - The ideas of the renormalization group (Nobel prize to Kenneth Wilson in 1982) was developed simultaneously in both high-energy and condensed matter physics.
   - The ideas of topological quantum field theories, while invented by string theorists as theories of quantum gravity, have been discovered in the laboratory by condensed matter physicists!

   That this type of physics is deep is not just my opinion. The Nobel committee agrees with me. During this course we will discuss the work of no fewer than 50 Nobel laureates! (See the index of scientists at the end of this set of notes).

4. **Because Reductionism Doesn't Work**

   `begin{rant}` People frequently have the feeling that if you continually ask "what is it made of" you learn more about something. This approach to knowledge is known as *reductionism.* For example, asking what water is made of, someone may tell you it is made from molecules, then molecules are made of atoms, atoms of electrons and protons, protons of quarks, and quarks are made of who-knows-what. But none of this information tells you anything about why water is wet, about why protons and neutrons bind to form nuclei, why the atoms bind to form water, and so forth. Understanding physics inevitably involves understanding how many objects all interact with each other. And this is where things get difficult very quickly. We understand the Schroedinger equation extremely well for one particle, but the Schroedinger equations for four or more particles, while in principle solvable, in practice are never solved because they are too difficult – even for the world's biggest computers. Physics involves figuring out what to do then. How are we to understand how many quarks form a nucleus, or how many electrons and protons form an atom if we cannot solve the many particle Schroedinger equation?

Even more interesting is the possibility that we understand very well the microscopic theory of a system, but then we discover that macroscopic properties *emerge* from the system that we did not expect. My personal favorite example is when one puts together many electrons (each with charge $-e$) one can sometimes find new particles emerging, each with charge $-e/3$![1] Reductionism would never uncover this – it misses the point completely. `end{rant}`

5. **Because it is a Laboratory**

   Condensed matter physics is perhaps the best laboratory we have for studying quantum physics and statistical physics. Those of us who are fascinated by what quantum mechanics and statistical mechanics can do often end up studying condensed matter physics which is deeply grounded in both of these topics. Condensed matter is an infinitely varied playground for physicists to test strange quantum and statistical effects.

   I view this entire course as an extension of what you have already learned in quantum and statistical physics. If you enjoyed those courses, you will likely enjoy this as well. If you did not do well in those courses, you might want to go back and study them again because many of the same ideas will arise here.

---

[1] Yes, this truly happens. The Nobel prize in 1998 was awarded to Dan Tsui, Horst Stormer and Bob Laughlin, for discovery this phenomenon known as the fractional quantum Hall effect.

Part I

Physics of Solids without Considering Microscopic Structure: The Early Days of Solid State

Chapter 2

Specific Heat of Solids: Boltzmann, Einstein, and Debye

Our story of condensed matter physics starts around the turn of the last century. It was well known (and you should remember from last year) that the specific heat of a monatomic (ideal) gas is $C_v = 3k_B/2$ per atom with k_B being Boltzmann's constant. The statistical theory of gases described why this is so.

As far back as 1819, however, it had also been known that for many solids the specific heat is given by[1]

$$\begin{aligned} C &= 3k_B \quad \text{per atom} \\ &= 3R \end{aligned}$$

which is known as the *Law of Dulong-Petit*[2]. While this law is not always correct, it frequently is close to true. For example, at room temperature we have

Material	C/R
Aluminum	2.91
Antimony	3.03
Copper	2.94
Gold	3.05
Silver	2.99
Diamond	0.735

With the exception of diamond, the law seems to hold extremely well at room temperature, although at lower temperatures all materials start to deviate from this law, and typically C drops rapidly below some temperature. (And for diamond when the temperature is raised, the specific heat increases towards $3R$ as well, see Fig. 2.2 below).

In 1896 Boltzmann constructed a model that accounted for this law fairly well. In his model, each atom in the solid is bound to neighboring atoms. Focusing on a single particular atom, we

---

[1] Here I do not distinguish between C_p and C_v because they are very close to the same. Recall that $C_p - C_v = VT\alpha^2/\beta_T$ where β_T is the isothermal compressibility and α is the coefficient of thermal expansion. For a solid α is relatively small.

[2] Both Pierre Dulong and Alexis Petit were French chemists. Neither is remembered for much else besides this law.

imagine the atom as being in a harmonic well formed by the interaction its neighbors. In such a classical statistical mechanical model, the specific heat of the vibration of the atom is $3k_B$ per atom, in agreement with Dulong-Petit. (Proving this is a good homework assignment that you should be able to answer with your knowledge of stat mech and/or the equipartition theorem).

Several years later in 1907, Einstein started wondering about why this law does not hold at low temperatures (for Diamond, "low" temperature appears to be room temperature!). What he realized is that quantum mechanics is important!

Einstein's assumption is similar to that of Boltzmann. He assumed that every atom is in a harmonic well created by the interaction with its neighbors. Further he assumed that every atom is in an identical harmonic well and has an oscillation frequency ω.

The quantum mechanical problem of a simple harmonic oscillator is one whose solution we know. We will now use that knowledge to determine the specific heat of a single 1d harmonic oscillator. This entire calculation should look familiar from your statistical physics course.

2.1 Einstein's Calculation

In 1d, the eigenstates of a single harmonic oscillator are

$$E_n = \hbar\omega(n + 1/2)$$

with ω the frequency of the harmonic oscillator. The partition function is then

$$\begin{aligned} Z_{1D} &= \sum_{n \geqslant 0} e^{-\beta\hbar\omega(n+1/2)} \\ &= \frac{e^{-\beta\hbar\omega/2}}{1 - e^{-\beta\hbar\omega}} = \frac{2}{\sinh(\beta\hbar\omega/2)} \end{aligned}$$

(We will very frequently use the notation $\beta = 1/(k_B T)$). The expectation of energy is then

$$\langle E \rangle = -\frac{1}{Z}\frac{\partial Z}{\partial \beta} = (\hbar\omega/2)\coth(\beta\hbar\omega/2) = \hbar\omega\left(n_B(\beta\hbar\omega) + \frac{1}{2}\right)$$

where n_B is the Bose[3] occupation factor

$$n_B(x) = \frac{1}{e^x - 1}$$

Note that this result is easy to interpret. The mode ω is an excitation that is excited on average n_B times, or equivalently there is a "boson" which is "occupied" n_B times.

Differentiating this result,

$$C = \frac{\partial \langle E \rangle}{\partial T} = k_B(\beta\hbar\omega)^2 \frac{e^{\beta\hbar\omega}}{(e^{\beta\hbar\omega} - 1)^2}$$

Note that the high temperature limit of this expression gives $C = k_B$ (check this if it is not obvious!).

---

[3] Satyendra Bose worked out the idea of Bose statistics in 1924, but could not get it published until Einstein lent his support to the idea.

Generalizing to the 3D case,

$$E_{n_1,n_2,n_3} = \hbar\omega[(n_1 + 1/2) + (n_2 + 1/2) + (n_3 + 1/2)]$$

and

$$Z_{3d} = \sum_{n1,n2,n3 \geqslant 0} e^{-\beta E_{n_1,n_2,n_3}} = [Z_{1D}]^3$$

resulting in $\langle E_{3D} \rangle = 3\langle E_{1D} \rangle$ so correspondingly we obtain

$$C = 3k_B(\beta\hbar\omega)^2 \frac{e^{\beta\hbar\omega}}{(e^{\beta\hbar\omega} - 1)^2}$$

Plotted this looks like Fig. 2.1.

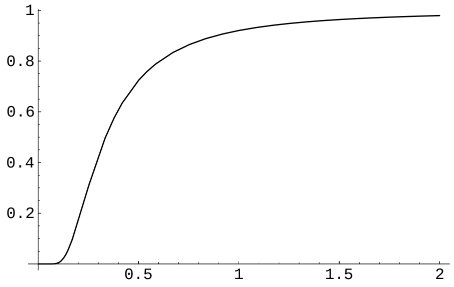


Figure 2.1: x-axis is $k_B T$ in units of $\hbar\omega$ and y axis is C in units of $3k_B$

Note that the high temperature limit $k_B T \gg \hbar\omega$ recover the law of Dulong-Petit — $3k_B$ heat capacity per atom. However, at low temperature (compared to $\hbar\omega/k_B$ the degrees of freedom "freeze out", the system gets stuck in only the ground state eigenstate, and the specific heat vanishes rapidly.

Einstein's theory reasonably accurately explained the behavior of the the heat capacity as a function of temperature with only a single fitting parameter, the so-called Einstein frequency ω. (Sometimes this frequency is quoted in terms of the Einstein temperature $\hbar\omega = k_B T_{Einstein}$). In Fig. 2.2 we show Einstein's original comparison to the heat capacity of diamond.

For most materials, the Einstein frequency ω is lower than room temperature, so the Dulong-Petit law hold fairly well (being relatively high temperature compared to the Einstein frequency). However, for diamond, ω is higher than room temperature, so the heat capacity is lower than $3R$ at room temperature. The reason diamond has such a high Einstein frequency is that the bonding between atoms in diamond is very strong (hence a high $\omega = \sqrt{k/m}$ oscillation frequency). These strong bonds also result in diamond being an exceptionally hard material.

This result was remarkable, not only in that it explained the temperature dependence of specific heat, but more importantly it told us something fundamental about quantum mechanics. Keep in mind that Einstein obtained this result 17 years before the Schroedinger equation was discovered![4]

---

[4] Einstein was a pretty smart guy.

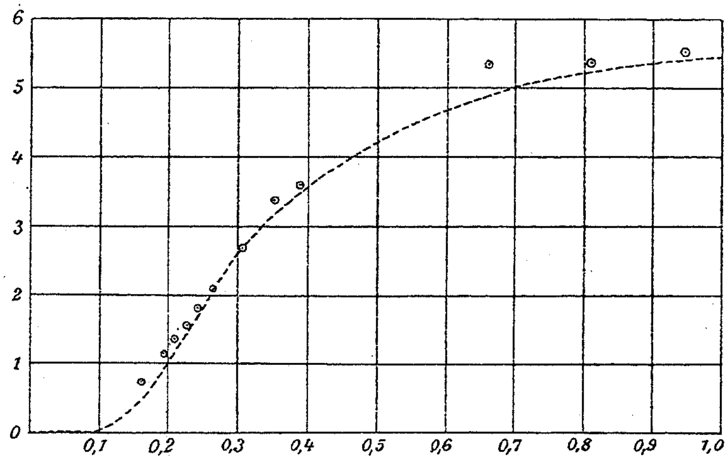


Figure 2.2: Plot of Specific Heat of Diamond from Einstein's original 1907 paper. The fit is to the Einstein theory of heat capacity. The x-axis is $k_B T$ in units of $\hbar\omega$ and y axis is C in units of cal/(K-mol). In these units, $3R \approx 5.96$).

2.2 Debye's Calculation

Einstein's theory of specific heat was remarkably successful, but still there were clear deviations from the predicted equation. Even in the plot in his first paper (Fig. 2.2 above) one can see that at low temperature the experimental data lies above the theoretical curve. This result turns out to be rather important! In fact, it was known that at low temperatures most materials have a heat capacity that is proportional to T^3 (Metals also have a very small additional term proportional to T which we will discuss later in section 3.2.2. Magnetic materials may have other additional terms as well. Nonmagnetic insulators have only the T^3 behavior). At any rate, Einstein's formula at low temperature is exponentially small in T, not agreeing at all with the actual experiments.

In 1912 Peter Debye[5] discovered how to better treat the quantum mechanics of oscillations of atoms, and managed to explain the T^3 specific heat. Debye realized that oscillation of atoms is the same thing as sound, and sound is a wave, so it should be quantized the same way as Planck quantized light waves. Besides the fact that the speed of light is much faster than that of sound, there is only one minor difference between light and sound: for light, there are two polarizations for each $\mathbf{k}$ whereas for sound, there are three modes for each $\mathbf{k}$ (a longitudinal mode, where the atomic motion is in the same direction as $\mathbf{k}$ and two transverse modes where the motion is perpendicular to $\mathbf{k}$. Light has only the transverse modes.). For simplicity of presentation here we will assume that the transverse and longitudinal modes have the same velocity, although in truth the longitudinal velocity is usually somewhat greater than the transverse velocity.

We now repeat essentially what was Planck's calculation for light. This calculation should also look familiar from your statistical physics course. First, however, we need some preliminary information about waves:

---

[5] Peter Debye later won a Nobel prize in Chemistry for something completely different.

2.2.1 About Periodic (Born-Von-Karman) Boundary Conditions

Many times in this course we will consider waves with periodic or "Born-Von-Karman" boundary conditions. It is easiest to describe this first in 1 dimension. Here, instead of having a 1d sample of length L with actual ends, we imagine that the two ends are connected together making the sample into a circle. The periodic boundary condition means that, any wave in this sample e^{ikr} is required to be the same for a position r as it is for $r + L$ (we have gone all the way around the circle). This then restricts the possible values of k to be

$$k = \frac{2\pi n}{L}$$

for n an integer. If we are ever required to sum over all possible values of k, for large enough L we can replace the sum with an integral obtaining

$$\sum_k = \sum_n \rightarrow \frac{L}{2\pi} \int_{-\infty}^{\infty} dk$$

where the prefactor is the Jacobian.[6]

In 3d, the story is extremely similar. For a sample of size L^3, we identify opposite ends of the sample (wrapping the sample up into a hypertorus!) so that if you go a distance L in any direction, you get back to where you started. As a result, our $\mathbf{k}$ values can only take values

$$\mathbf{k} = \frac{2\pi}{L}(n_1, n_2, n_3)$$

for integer values of n_i. Because of this discretization of values of $\mathbf{k}$, whenever we have a sum over all possible $\mathbf{k}$ values we obtain

$$\sum_{\mathbf{k}} \rightarrow \frac{L^3}{(2\pi)^3} \int d^3k$$

One might think that wrapping the sample up into a hypertorus is very unnatural compared to considering a system with real boundary conditions. However, these boundary conditions tend to simplify calculations quite a bit and most physical quantities you might measure far from the boundaries of the sample are independent of what you do with the boundary conditions.

2.2.2 Debye's Calculation Following Planck

Debye decided that the oscillation modes were waves with frequencies $\omega(\mathbf{k}) = v|\mathbf{k}|$ with v the sound velocity — and for each $\mathbf{k}$ there should be three possible oscillation modes. Thus he wrote an expression entirely analogous to Einstein's expression

$$\begin{aligned} \langle E \rangle &= 3 \sum_{\mathbf{k}} \hbar\omega(\mathbf{k}) \, [n_B(\beta\hbar\omega(\mathbf{k})) + \frac{1}{2}] \\ &= 3 \frac{L^3}{(2\pi)^3} \int d\mathbf{k} \, \hbar\omega(\mathbf{k}) \, [n_B(\beta\hbar\omega(\mathbf{k})) + \frac{1}{2}] \end{aligned}$$

---

[6] In your previous courses you may have used particle in a box boundary conditions where instead of plane waves $e^{i2\pi n r/L}$, you used particle in a box wavefunctions of the form $\sin(kn\pi/L)$. This gives you instead

$$\sum_k = \sum_n \rightarrow \frac{L}{\pi} \int_0^{\infty} dk$$

which will inevitably result in the same physical answers as for the periodic boundary condition case. All calculations can be done either way, but periodic Born-Von-Karmen boundary conditions are almost always simpler.

Each excitation mode is a boson of frequency $\omega(\mathbf{k})$ and it is occupied on average $n_B(\beta\hbar\omega(\mathbf{k}))$ times.

By spherical symmetry, we may convert the 3d integral to a 1d integral

$$\int d\mathbf{k} \to 4\pi \int_0^\infty k^2 dk$$

(recall that $4\pi k^2$ is the area of the surface of a sphere[7] of radius k) and we also use $k = \omega/v$ to obtain

$$\langle E \rangle = 3\frac{4\pi L^3}{(2\pi)^3} \int_0^\infty \omega^2 d\omega (1/v^3)(\hbar\omega)[n_B(\beta\hbar\omega) + \frac{1}{2}]$$

It is convenient to replace $nL^3 = N$ where n is the density of atoms. We then obtain

$$\langle E \rangle = \int_0^\infty d\omega g(\omega)(\hbar\omega)[n_B(\beta\hbar\omega) + \frac{1}{2}] \tag{2.1}$$

where the *density of states* is given by

$$g(\omega) = N\left[\frac{12\pi\omega^2}{(2\pi)^3 n v^3}\right] = N\frac{9\omega^2}{\omega_d^3}$$

where

$$\omega_d^3 = 6\pi^2 n v^3$$

This frequency will be known as the *Debye frequency*, and below we will see why we chose to define it this way with the factor of 9 removed.

This result, Eq. 2.1 for the quantum energy of the sound waves is strikingly similar to Planck's result for the quantum energy of light waves, only we have replaced $2/c^3$ by $3/v^3$ (replacing the 2 light modes by 3 sound modes). The other change from Planck's classic result is the $+1/2$ that we obtain as the zero point energy of each oscillator. This is due to the zero-point energy of each oscillator (Planck should have gotten this energy as well, but he didn't know about zero-point energy — in fact, since it was long before quantum mechanics was fully understood, Debye didn't actually have this term either). At any rate, this zero point energy gives us a contribution which is temperature independent. Since we are concerned with $C = \partial\langle E \rangle/\partial T$ this term will not contribute and we can throw it out. We thus obtain

$$\langle E \rangle = \frac{9N\hbar}{\omega_d^3} \int_0^\infty d\omega \frac{\omega^3}{e^{\beta\hbar\omega} - 1}$$

by defining a variable $x = \beta\hbar\omega$ this becomes

$$\langle E \rangle = \frac{9N\hbar}{\omega_d^3(\beta\hbar)^4} \int_0^\infty dx \frac{x^3}{e^x - 1}$$

The integral just gives some number – in fact the number is $\pi^4/15$ (we prove this below in an appendix to this chapter). Thus we obtain

$$\langle E \rangle = 9N\frac{(k_B T)^4}{(\hbar\omega_d)^3}\frac{\pi^4}{15}$$

Notice the similarity to Planck's derivation of the T^4 energy of photons. As a result, the specific heat is

$$C = \frac{\partial\langle E \rangle}{\partial T} = N k_B \frac{(k_B T)^3}{(\hbar\omega_d)^3}\frac{12\pi^4}{5} \sim T^3$$

---

[7] Or to be pedantic, $\int d\mathbf{k} \to \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \int k^2 dk$ and performing the angular integrals gives 4π.

This correctly obtains the desired T^3 specific heat. Furthermore, the prefactor of T^3 can be calculated in terms of things such as the sound velocity and the density of atoms. Note that the Debye frequency in this equation is replaced by a temperature

$$\hbar\omega_d = k_B T_{Debye}$$

known as the *Debye temperature*, so that this equation reads

$$C = \frac{\partial \langle E \rangle}{\partial T} = N k_B \frac{(T)^3}{(T_{Debye})^3} \frac{12\pi^4}{5}$$

2.2.3 Debye's "Interpolation"

Unfortunately, now Debye has a problem. In the expression derived above, the specific heat is proportional to T^3 up to arbitrarily high temperature. We know however, that the specific heat should level off to $3k_BN$ at high T. Debye understood that the problem with his approximation is that it allows an infinite number of sound wave modes — up to arbitrarily large $\mathbf{k}$. This would imply more sound wave modes than there are atoms in the entire system. Debye guessed that really there should be only as many modes as there are degrees of freedom in the system. We will see later in the course that this is an important principle. To fix this problem, Debye decided to not consider sound waves above some maximum frequency ω_{cutoff}, with this frequency chosen such that there are exactly $3N$ sound wave modes in the system (3 dimensions of motion times N particles). We thus define ω_{cutoff} via

$$3N = \int_0^{\omega_{cutoff}} d\omega \, g(\omega) \tag{2.2}$$

We correspondingly rewrite Eq. 2.1 for the energy (dropping the zero point contribution) as

$$\langle E \rangle = \int_0^{\omega_{cutoff}} d\omega \, g(\omega) \, \hbar\omega \, n_B(\beta\hbar\omega) \tag{2.3}$$

Let us now check that this gives us the correct high temperature limit. For high temperature

$$n_B(\beta\hbar\omega) = \frac{1}{e^{\beta\hbar\omega} - 1} \rightarrow \frac{k_B T}{\hbar\omega}$$

Thus in the high temperature limit, invoking Eq. 2.2 we obtain

$$\langle E \rangle = k_B T \int_0^{\omega_{cutoff}} d\omega g(\omega) \; = 3k_B T N$$

yielding the Dulong-Petit high temperature specific heat $C = 3k_BN$. For completeness, let us now evaluate our cutoff frequency,

$$3N = \int_0^{\omega_{cutoff}} d\omega g(\omega) = 9N \int_0^{\omega_{cutoff}} d\omega \frac{\omega^2}{\omega_d^3} = 3N \frac{\omega_{cutoff}^3}{\omega_d^3}$$

we thus see that the correct cutoff frequency is exactly the Debye frequency ω_d. Note that $k = \omega_d/v$ is on the order of the interatomic spacing of the solid.

More generally (in the neither high nor low temperature limit) one has to evaluate the integral 2.3, which cannot be done analytically. Nonetheless it can be done numerically and then can be compared to actual experimental data as shown in Fig. 2.3.

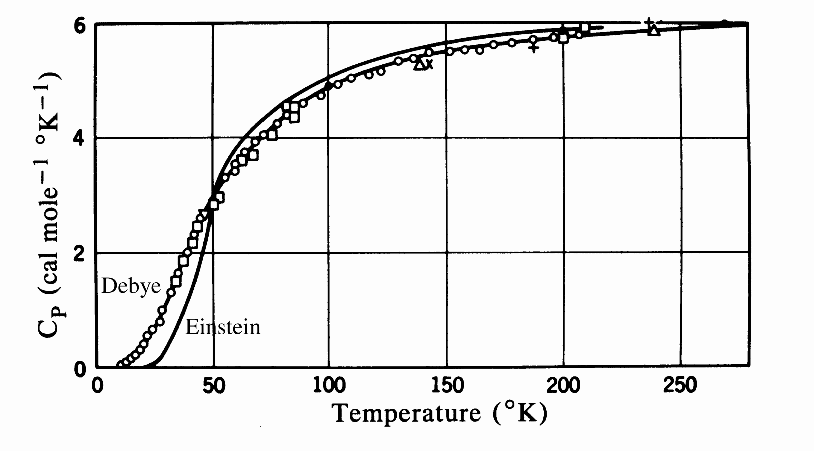


Figure 2.3: Plot of Specific Heat of Silver. The y axis is C in units of cal/(K-mol). In these units, $3R \approx 5.96$). Over the entire experimental range, the fit to the Debye theory is excellent. At low T it correctly recovers the T^3 dependence, and at high T it converges to the law of Dulong-Petit.

2.2.4 Some Shortcomings of the Debye Theory

While Debye's theory is remarkably successful, it does have a few shortcomings.

- The introduction of the cutoff seems very ad-hoc. This seems like a successful cheat rather than real physics

- We have assumed sound waves follow the law $\omega = vk$ even for very very large values of k (on the order of the inverse lattice spacing), whereas the entire idea of sound is a long wavelength idea, which doesn't seem to make sense for high enough frequency and short enough wavelength. At any rate, it is known that at high enough frequency the law $\omega = vk$ no longer holds.

- Experimentally, the Debye theory is accurate, but it is not exact

- In particular, at very very low temperatures, metals have a term in the heat capacity that is proportional to T, so the overall specific heat is $C = aT + bT^3$ and at low enough T the linear term will dominate. You can't see this contribution on the plot Fig. 2.3 but at very low T it becomes evident.

Of these shortcomings, the first three can be handled more properly by treating the details of the crystal structure of materials accurately (which we will do much later in this course). The final issue requires us to carefully study the behavior of electrons in metals to discover the origin of this linear T term (we shall do this next).

Nonetheless, despite these problems, Debye's theory was a substantial improvement over Einstein's[8]

---

[8] Debye was pretty smart too... even though he was a chemist.

2.3 Summary of Specific Heat of Solids

- (Much of the) Specific Heat of materials is due to atomic vibrations
- Boltzmann and Einstein models consider these vibrations as N simple harmonic oscillators.
- Boltzmann classical analysis obtains law of Dulong-Petit $C = 3Nk_B = 3R$
- Einstein quantum analysis shows that at temperatures below the oscillator frequency, degrees of freedom freeze out, and heat capacity drops exponentially
- Debye Model treats oscillations as sound waves.
  - $\omega = v|k|$, similar to light (but three polarizations not two)
  - quantization similar to Planck quantization of light
  - Maximum frequency cutoff ($\hbar\omega_{Debye} = k_B T_{Debye}$) necessary to obtain a total of only $3N$ degrees of freedom
  - obtains Dulong-Petit at high T and $C \sim T^3$ at low T
- Metals have an additional (albeit small) linear T term in the specific heat which we will discuss later.

References

Almost every book covers the material introduced in this chapter, but frequently it is done late in the book only after the idea of phonons is introduced. We will get to phonons in chapter 5. Before we get there the following references cover this material without discussion of phonons:

- Goodstein sections 3.1 and 3.2
- Rosenberg sections 5.1 through 5.13 (good problems included)
- Burns sections 11.3 through 11.5 (good problems included)

Once we get to phonons, we can look back at this material again. Discussions are then given also by

- Dove section 9.1 and 9.2
- Ashcroft and Mermin chapter 23
- Hook and Hall section 2.6
- Kittel beginning of chapter 5.

2.4 Appendix to this chapter

Here is how we evaluate the nasty integral:

$$\begin{aligned}\int_0^\infty dx \frac{x^3}{e^x - 1} &= \int_0^\infty dx \frac{x^3 e^{-x}}{1 - e^{-x}} = \int_0 x^3 e^{-x} \sum_{n=0}^{\infty} e^{-nx} \\ &= \sum_{n=1}^{\infty} \int_0 x^3 e^{-nx} = \sum_{n=1}^{\infty} \frac{3!}{n^4} = 3!\zeta(4) = \frac{\pi^4}{15}\end{aligned}$$

The last step is from looking up the Riemann zeta function on a table (which is much easier than actually doing the integral).

However, if we must do this calculation ourselves (in the unlikely event that we are stranded on a desert island) a sketch of the calculation goes as follows. First write a Fourier series for the function x^2 on the interval $[-\pi, \pi]$. The series is given by

$$x^2 = \sum_{n \geqslant 0} a_n \cos(nx)$$

with coefficients given by

$$a_n \quad = \quad \frac{1}{\pi} \int_{\pi}^{\pi} dx x^2 \cos(nx)$$

These can be calculated straightforwardly to give

$$x^2 = \frac{\pi^2}{3} + \sum_{n=1}^{\infty} \frac{4(-1)^n}{n^2} \cos(nx)$$

Now we invoke Parseval's theorem (here quoted for an even function f with Fourier modes a_n):

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 = \frac{1}{4} \sum_n |a_n|^2$$

(This theorem is not hard to prove). The left hand gives $\pi^4/5$ whereas the right hand gives

$$\frac{1}{4}\left(\frac{2\pi^2}{3}\right)^2 + \frac{1}{2} \sum_{n=1}^{\infty} \left(\frac{4}{n^2}\right)^2 = \frac{\pi^4}{9} + 8 \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{9} + 8\zeta(4).$$

which thus gives us the result $\zeta(4) = \pi^4/90$.

Chapter 3

Electrons in Metals – An Early Look

The fundamental characteristic of a metal is that it conducts electricity. At some level the reason for this conduction boils down to the fact that electrons are mobile in these materials. In later chapters we will be concerned with the question of why electrons are mobile in some materials but not in others, being that all materials have electrons in them! For now, we take as given that there are mobile electrons and would like to understand their properties.

3.1 Drude (Kinetic) Theory

J.J. Thomson's 1896 discovery of the electron ("corpuscles of charge" that could be pulled out of metal) raised the question of how these charge carriers might move within the metal. In 1900 Paul Drude[1] realized that he could apply Boltzmann's kinetic theory of gases to understanding electron motion within metals. This theory was remarkably successful, providing a first understanding of metallic conduction.[2]

Having studied kinetic theory of gas, Drude theory should be very easy to understand. We will make three assumptions about the motion of electrons

1. Electrons have a scattering time τ. The probability of scattering within a time interval dt is dt/τ.
2. Once a scattering event occurs, we assume the electron returns to momentum $\mathbf{p} = 0$.
3. In between scattering events, the electrons, which are charge $-e$ particles, respond to the externally applied electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$.

The first two of these assumptions are exactly those made in the kinetic theory of gases[3]. The

---

[1] pronounced roughly "Drood-a"

[2] Sadly, neither Boltzmann nor Drude lived to see how much influence this theory really had — in unrelated tragic events, both of them committed suicide in 1906. Boltzmann's famous student, Ehrenfest, also committed suicide some years later. Why so many highly successful statistical physicists took their own lives is a bit of a mystery.

[3] Ideally we would do a better job with our representation of the scattering of particles. Every collision should

third assumption is just a logical generalization to account for the fact that, unlike gases molecules, electrons are charged and must therefore respond to electromagnetic fields.

We consider an electron with momentum $\mathbf{p}$ at time t and we ask what momentum it will have at time $t + dt$. There are two terms in the answer, there is a probability dt/τ that it will scatter to momentum zero. If it does not scatter to momentum zero (with probability $1 - dt/\tau$) it simply accelerates as dictated by its usual equations of motion $d\mathbf{p}/dt = \mathbf{F}$. Putting the two terms together we have

$$\langle \mathbf{p}(t+dt) \rangle = \left(1 - \frac{dt}{\tau}\right)(\mathbf{p}(t) + \mathbf{F}dt) + \mathbf{0}dt/\tau$$

or[4]

$$\frac{d\mathbf{p}}{dt} = \mathbf{F} - \frac{\mathbf{p}}{\tau} \tag{3.1}$$

where here the force $\mathbf{F}$ on the electron is just the Lorentz force

$$\mathbf{F} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B})$$

One can think of the scattering term $-\mathbf{p}/\tau$ as just a drag force on the electron. Note that in the absence of any externally applied field the solution to this differential equation is just an exponentially decaying momentum

$$\mathbf{p}(t) = \mathbf{p}_{initial}\; e^{-t/\tau}$$

which is what we should expect for particles that lose momentum by scattering.

3.1.1 Electrons in an Electric Field

Let us start by considering the case where the electric field is nonzero but the magnetic field is zero. Our equation of motion is then

$$\frac{d\mathbf{p}}{dt} = -e\mathbf{E} - \frac{\mathbf{p}}{\tau}$$

In steady state, $d\mathbf{p}/dt = 0$ so we have

$$m\mathbf{v} = \mathbf{p} = -e\tau\mathbf{E}$$

with m the mass of the electron and $\mathbf{v}$ its velocity.

Now, if there is a density n of electrons in the metal, and they are all moving at velocity $\mathbf{v}$ then the electrical current is given by

$$\mathbf{j} = -en\mathbf{v} = \frac{e^2\tau n}{m}\mathbf{E}$$

---

have two particles having $\mathbf{p}_1^{initial}$ and $\mathbf{p}_2^{initial}$ and then scattering to $\mathbf{p}_1^{final}$ and $\mathbf{p}_2^{final}$ so as to conserve both energy and momentum. Unfortunately, keeping track of things so carefully makes the problem extremely difficult to solve. Assumption 1 is not so crazy as an approximation being that there really is a typical time between scattering events in a gas. Assumption 2 is a bit more questionable, but *on average* the final momentum after a scattering event is indeed zero (if you average momentum as a vector). However, obviously it is not correct that every particle has zero kinetic energy after a scattering event. This is a defect of the approach.

[4] Here we really mean $\langle \mathbf{p} \rangle$ when we write $\mathbf{p}$. Since our scattering is probabilistic, we should view all quantities (such as the momentum) as being an expectation over these random events. A more detailed theory would keep track of the entire distribution of momenta rather than just the average momentum. Keeping track of distributions in this way leads one to the Boltzmann Transport Equation, which we will not discuss in this course.

or in other words, the conductivity of the metal is[5]

$$\sigma = \frac{e^2 \tau n}{m} \tag{3.2}$$

By measuring the conductivity of the metal (assuming we know both the charge and mass of the electron) we can determine the product of the density and scattering time of the electron.

3.1.2 Electrons in Electric and Magnetic Fields

Let us continue on and see what other predictions come from Drude theory. We now consider the transport equation 3.1 for a system in both an electric and a magnetic field. We now have

$$\frac{d\mathbf{p}}{dt} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) - \mathbf{p}/\tau$$

Again setting this to zero in steady state, and using $\mathbf{p} = m\mathbf{v}$ and $\mathbf{j} = -ne\mathbf{v}$, we obtain an equation for the steady state current

$$0 = -e\mathbf{E} + \frac{\mathbf{j} \times \mathbf{B}}{n} + \frac{m}{ne\tau}\mathbf{j}$$

or

$$\mathbf{E} = \left(\frac{1}{ne}\mathbf{j} \times \mathbf{B} + \frac{m}{ne^2\tau}\mathbf{j} \right)$$

We now define the 3 by 3 resistivity matrix $\underset{\sim}{\rho}$ which relates the current vector to the electric field vector

$$\mathbf{E} = \underset{\sim}{\rho}\,\mathbf{j}$$

We then obtain components of this matrix

$$\rho_{xx} = \rho_{yy} = \rho_{zz} = \frac{m}{ne^2\tau}$$

and if we imagine $\mathbf{B}$ oriented in the $\hat{z}$ direction, then

$$\rho_{xy} = -\rho_{yx} = \frac{B}{ne}$$

(and all other components of $\underset{\sim}{\rho}$ are zero). This off-diagonal term in the resistivity is known as the "Hall"-resistivity, named after Edwin Hall who discovered in 1879 that there when a magnetic field is applied perpendicular to a current flow a voltage can be measured perpendicular to both current and magnetic field. (See Fig. 3.1). As a homework problem we will consider a further generalization to finite frequency conductivity, and we will see that Drude theory also gives some interesting (and frequently accurate) predictions.

The Hall coefficient R_H is defined as

$$R_H = \frac{\rho_{yx}}{|B|}$$

which in the Drude theory is given by

$$R_H = \frac{-1}{ne}$$

---

[5] A related quantity is the *mobility* defined as $\mu = e\tau/m$ which we will discuss in section 13.1.1 below.

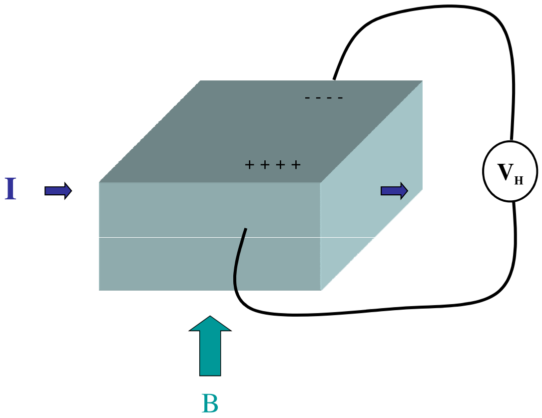


Figure 3.1: Edwin Hall's 1879 experiment. The voltage measured perpendicular to both the magnetic field and the current is known as the Hall voltage which is proportional to H and inversely proportional to the electron density (at least in Drude theory).

This then allows us to measure the density of electrons in a metal.

Aside: One can also consider turning this experiment on its head. If you know the density of electrons in your sample you can use a Hall measurement to determine the magnetic field. This is known as a Hall sensor. Since it is hard to measure small voltages, Hall sensor typically use materials, such as semiconductors, where the density of electrons is low so R_H and hence the resulting voltage is large.

Let us then calculate $n = -1/(eR_H)$ for various metals and divide it by the density of atoms. This should give us the number of free electrons per atom. Later on in this course we will see that it is frequently not so hard to estimate the number of electrons in a system. A short description is that electrons bound in the core shells of the atoms are never free to travel throughout the crystal, whereas the electrons in the outer shell may be free (we will discuss later in the course when these electrons are free and when they are not). The number of electrons in the outermost shell is known as the *valence* of the atom.

Material	$(-1/[eR_H])/$ [density of atoms] In Drude theory this should give the number of free electrons per atom which is the valence	Valence
Li	.8	1
Na	1.2	1
K	1.1	1
Cu	1.5	1 (usually)
Be	-0.2	2
Mg	-0.4	2

We see from the table that for many metals this Drude theory analysis seems to make sense — the "valence" of Li, Na, and K are all one . The effective valence of Cu is also one, so this is not surprising either. However, something has clearly gone seriously wrong for Be and Mg. In this case, the sign of the Hall coefficient has come out incorrect. From this result, one might conclude that the charge carrier for Be and Mg have the opposite charge from that of the electron! Indeed, we will see below in section 13.1.1 that this is indeed true and is a result of the so-called band structure of these materials. However, for many metals, simple Drude theory gives quite reasonable results.

If we believe the Hall effect measurement of the density of electrons in metals, using Eq. 3.2 we can then extract a scattering time from the expression for the conductivity. The Drude scattering time comes out to be in the range of $\tau \approx 10^{-14}$ seconds for most metals near room temperature.

3.1.3 Thermal Transport

Drude was brave enough to attempt to further calculate the thermal conductivity κ due to mobile electrons using Boltzmann's kinetic theory. Without rehashing the derivation, this result should look familiar to you from your previous encounters with kinetic theory of gas

$$\kappa = \frac{1}{3} n c_v \langle v \rangle \lambda$$

where c_v is the specific heat per particle, $\langle v \rangle$ is the average thermal velocity and $\lambda = \langle v \rangle \tau$ is the scattering length. For a conventional gas the specific heat per particle is

$$c_v = \frac{3}{2} k_B$$

and

$$\langle v \rangle = \sqrt{\frac{8 k_B T}{\pi m}}$$

Assuming this holds true for electrons, we obtain

$$\kappa = \frac{4}{\pi} \frac{n \tau k_B^2 T}{m}$$

While this quantity still has the unknown parameter τ in it, it is the same quantity that occurs in the electrical conductivity (Eq. 3.2). Thus we may look at the ratio of thermal conductivity to electrical conductivity, know as the Lorenz number[6,7]

$$L = \frac{\kappa}{T\sigma} = \frac{4}{\pi} \left(\frac{k_B}{e} \right)^2 \approx 0.94 \times 10^{-8} \text{ WattOhm/K}^2$$

A slightly different prediction is obtained by realizing that we have used $\langle v \rangle^2$ in our calculation, whereas perhaps we might have instead used $\langle v^2 \rangle$ which would have then given us instead

$$L = \frac{\kappa}{T\sigma} = \frac{3}{2} \left(\frac{k_B}{e} \right)^2 \approx 1.11 \times 10^{-8} \text{ WattOhm/K}^2$$

---

[6] This is named after Ludvig Lorenz, not Hendrik Lorentz who is famous for the Lorentz force and Lorentz contraction. However, just to confuse matters, the two of them worked on similar topics and there is even a Lorentz-Lorenz equation

[7] The dimensions here might look a bit funny, but κ, the thermal conductivity is measured in Watt/K and σ is measured in 1/Ohm. To see that WattOhm/K^2 is the same as $(k_b/e)^2$ note that k_b is J/K and e is Coulomb (C). So we need to show that (J/C)^2 is Watt-Ohm

$$(\text{J/C})^2 = (\text{J/sec})(\text{J/C})(1/(\text{C/sec}) = \text{WattVolt/Amp} = \text{WattOhm}$$

This result was viewed as a huge success, being that it was known for almost half a century that almost all metals have roughly the same value of this ratio, a fact known as the Wiedemann-Franz law.

In fact the value predicted for this ratio is only a bit lower than that measured experimentally

Material	$L \times 10^8$ (WattOhm/K^2)
Li	2.22
Na	2.12
Cu	2.20
Fe	2.61
Bi	3.53
Mg	2.14
Drude Prediction	0.98-1.11

So the result appears to be off by about a factor of 2, but still that is very good, considering that before Drude no one had any idea why this ratio should be a constant at all!

In retrospect we now realize that this calculation is completely incorrect (despite its successful result). The reason we know there is a problem is because we do not actually measure a specific heat of $c_v = \frac{3}{2}k_B$ per electron in metals (for certain systems where the density of electrons is very low, we do in fact measure this much specific heat, but not in metals). In fact, in most metals we measure only a vibrational (Debye) specific heat, plus a very small term linear in T at low temperatures. So why does this calculation give such a good result? It turns out (and we will see later below) that we have made two mistakes that roughly cancel each other. We have used a specific heat that is way too large, but we have also used a velocity that is way too small. We will see later that both of these mistakes are due to Fermi statistics of the electron (which we have so far ignored) and the Pauli exclusion principle.

We can see the problem much more clearly in some other quantities. The so-called *Peltier effect* is the fact that running electrical current through a material also transports heat. The so-called Peltier coefficient Π is defined by

$$\mathbf{j}^q = \Pi \mathbf{j}$$

where $\mathbf{j}^q$ is the heat current, and $\mathbf{j}$ is the electrical current.

Aside: The Peltier effect is used for thermoelectric refrigeration devices. By running electricity through a thermoelectric material, you can force heat to be transported through that material. You can thus transport heat away from one object and towards another. A good thermoelectric device has a high Peltier coefficient, but must also have a low resistivity, because running a current through an material with resistivity R will result in power I^2R being dissipated thus heating it up.

In kinetic theory the thermal current is

$$\mathbf{j}^q = \frac{1}{3}\mathbf{v}n(c_v T) \tag{3.3}$$

here $c_v T$ is the heat carried by one particle (with $c_v = 3k_B/2$ the specific heat per particle), n is the density of particles, and n is the density of particles (and $1/3$ is the geometric factor that is probably approximate anyway). Similarly the electrical current is

$$\mathbf{j} = -en\mathbf{v}$$

Thus the Peltier coefficient is

$$\Pi = \frac{-c_v T}{3e} = \frac{-k_B T}{2e} \tag{3.4}$$

Thus the ratio (known as thermopower, or Seebeck coefficient) $S = \Pi/T$ is given by

$$S = \frac{\Pi}{T} = \frac{-k_B}{2e} = -4.3 \times 10^{-4} V/K \tag{3.5}$$

in Drude theory. For most metals the actual value of this ratio is roughly 100 times smaller! This is a reflection of the fact that we have used $c_v = 3k_B/2$ whereas the actual specific heat per particle is much much lower (which we will understand in the next section when we consider Fermi statistics more carefully).

3.1.4 Summary of Drude Theory

- Based on kinetic theory of gases
- Assumes some scattering time τ, resulting in a conductivity $\sigma = ne^2\tau/m$
- Hall coefficient measures density of electrons
- Successes
  - Wiedemann-Franz ratio $\kappa/(\sigma T)$ comes out close to right for most materials
  - Many other transport properties predicted correctly (ex, conductivity at finite frequency)
  - Hall coefficient measurement of the density comes out reasonable for many metals.
- Failures
  - Hall coefficient frequently is measured to have the wrong sign, indicating a charge carrier with charge opposite to that of the electron
  - There is no $3k_B/2$ specific heat per particle measured for electrons in metals. This then makes the Peltier coefficient come out wrong by a factor of 100.

The latter of the two shortcomings will be addressed in the next section, whereas the former of the two will be addressed in chapter 13 below where we discuss band theory.

Despite the shortcomings of Drude theory, it nonetheless was the only theory of metallic conductivity for a quarter of a century (until the Sommerfeld theory improved it), and it remains quite useful today.

References for Drude Theory

- Ashcroft and Mermin, chapter 1
- Burns, chapter 9 part A
- Singleton, section 1.1-1.4
- Hook and Hall section 3.3 sort-of.

Actually, Hook and Hall are aimed mainly at Free electron (Sommerfeld) theory, but they end up doing Drude theory anyway (they don't use the word "Drude").

3.2 Sommerfeld (Free-Electron) Theory of Metals

In 1925 Pauli discovered the exclusion principle, that no two electrons may be in the exact same state. In 1926, Fermi and Dirac separately derived what we now call Fermi-Dirac statistics[8] Upon learning about Fermi Statistics, Sommerfeld[9] realized that Drude's theory of metals could easily be generalized to incorporate Fermi statistics, which is what we shall presently do.

3.2.1 Basic Fermi-Dirac Statistics

Given a system of electrons with chemical potential μ the probability of an eigenstate of energy E being occupied is given by the Fermi factor[10] (See Fig. 3.2)

$$n_F(\beta(E-\mu)) = \frac{1}{e^{\beta(E-\mu)}+1} \tag{3.6}$$

At low temperature this becomes a step function (states below the chemical potential are filled, those above the chemical potential are empty), whereas at higher temperatures the step function becomes more smeared out.

We will consider the electrons to be in a box of size $V = L^3$ and, as with our above discussion of sound waves, it is easiest to imagine that the box has periodic boundary conditions. The plane wavefunctions are of the form $e^{i\mathbf{k}\cdot\mathbf{r}}$ where $\mathbf{k}$ must take value $(2\pi/L)(n_1, n_2, n_3)$ with n_i integers due to the boundary conditions. These plane waves have corresponding energies

$$\epsilon(\mathbf{k}) = \frac{\hbar^2|k|^2}{2m} \tag{3.7}$$

with m the electron mass. Thus the total number of electrons in the system is given by

$$N = 2\sum_{\mathbf{k}} n_F(\beta(\epsilon(\mathbf{k})-\mu)) = 2\frac{V}{(2\pi)^3}\int d\mathbf{k}\; n_F(\beta(\epsilon(\mathbf{k})-\mu)) \tag{3.8}$$

where the prefactor of 2 accounts for the two possible spin states for each possible wavevector $\mathbf{k}$. In fact, in a metal, N will usually be given to us, and this equation will define the chemical potential as a function of temperature.

We now define a useful concept:

Definition 3.2.1. The *Fermi Energy*, E_F is the chemical potential at temperature $T = 0$.

This is also sometimes called the *Fermi Level*. The states that are filled at $T = 0$ are sometimes called the *Fermi Sea*. Frequently one also defines a *Fermi Temperature* $T_F = E_F/k_B$,

---

[8] All three, Pauli, Fermi, and Dirac, won Nobel prizes in the next few years — but you probably knew that already.

[9] Sommerfeld never won a Nobel prize, although he was nominated for it 81 times — more than any other physicist. He also was a research advisor for more Nobel Laureates than anyone else in history (6: Heisenberg, Pauli, Debye, Bethe, who were his PhD students and Pauling, Rabi who were postdoctoral researchers with him. He also was the first research advisor for Rudolf Peierls for whom the theory building at Oxford is named, Falthough Peierls eventually finished his PhD as a student of Pauli.)

[10] When we say that there are a particular set of N orbitals occupied by electrons, we really mean that the overall wavefunction of the system is an antisymmetric function $\Psi(1,\ldots,N)$ which can be expressed as a Slater determinant of N particle coordinates occupying the N orbitals. We will never need to actually write out such Slater determinant wavefunctions except in Appendix 19.4 which is nonexaminable.

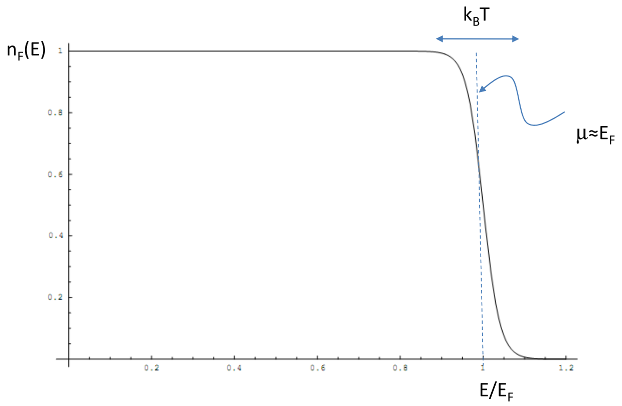


Figure 3.2: The Fermi Distribution for $k_B T \ll E_F$.

and also the *Fermi Wavevector* k_F defined via

$$E_F = \frac{\hbar^2 k_F^2}{2m} \tag{3.9}$$

and correspondingly a *Fermi momentum* $p_F = \hbar k_F$ and a *Fermi velocity*[11]

$$v_F = \hbar k_F / m \tag{3.10}$$

Aside: Frequently people think of the Fermi Energy as the energy of the most energetic occupied electron state in system. While this is correct in the case where you are filling a continuum of states, it can also lead you to errors in cases where the energy eigenstates are discrete, or more specifically when there is a gap between the most energetic occupied electron in the system, and the least energetic unoccupied electron state. More correctly the Fermi energy, i.e., the chemical potential at $T = 0$, will be half way between the most energetic occupied electron in the system, and the least energetic unoccupied electron state.

Let us now calculate the Fermi energy in a (3 dimensional) metal with N electrons in it. At $T = 0$ the Fermi function (Eq. 3.6) becomes a step function (which we write as Θ. I.e.,$\Theta(x) = 1$ for $x > 0$ and $= 0$ for $x < 0$) , so that Eq. 3.8 becomes

$$N = 2\frac{V}{(2\pi)^3} \int d\mathbf{k}\; \Theta(E_F - \epsilon(\mathbf{k})) = 2\frac{V}{(2\pi)^3} \int^{|k|<k_F} d\mathbf{k} \;\; = 2\frac{V}{(2\pi)^3}\left(\frac{4}{3}\pi k_F^3\right) \tag{3.11}$$

where in the last step we have use the fact that the volume of a ball is $4\pi/3$ times the cube of the radius, and at $T = 0$ the electrons fill a ball (a "Fermi sphere") up to radius k_F. The surface of this ball, the sphere of radius k_F is known as the *Fermi Surface* — the surface dividing filled from unfilled states at zero temperature.

Using the fact that the density $n = N/V$ we can rearrange Eq. 3.11 to give

$$k_F = (3\pi^2 n)^{1/3}$$

---

[11] Yes, Fermi got his name attached to many things. To help spread the credit around I've called this chapter "Basic Fermi-Dirac Statistics" instead of just "Basic Fermi Statistics".

and correspondingly

$$E_F = \frac{\hbar^2 (3\pi^2 n)^{2/3}}{2m} \tag{3.12}$$

Since we know roughly how many free electrons there are in a metal (say, one per atom for monovalent metals such as sodium or copper), we can estimate the Fermi energy, which, say for copper, turns out to be on the order of 7 eV, corresponding to a Fermi temperature of about 80,000 K. (!). This amazingly high energy scale is a result of Fermi statistics and the very high density of electrons in metals. It is crucial to remember that for all metals, $T_F \gg T$ for any temperature anywhere near room temperature. In fact metals melt (and even vaporize!) at temperatures far far below their Fermi temperatures.

Similarly, one can calculate the Fermi Velocity, which, for a typical metal such as copper, may be as large as 1% the speed of light! Again, this enormous velocity stems from the Pauli exclusion principle — all the lower momentum states are simply filled, so if the density of electrons is very high, the velocities will be very high as well.

With a Fermi energy that is so large, and therefore a Fermi sea that is very deep, any (not insanely large) temperature can only make excitations of electrons that are already very close to the Fermi surface. The electrons deep within the Fermi sea, near $\mathbf{k} = \mathbf{0}$, cannot be moved by any reasonably low energy perturbation simply because there are no available unfilled states for them to move to unless they absorb a very large amount of energy.

3.2.2 Electronic Specific Heat

We now turn to try to calculate the electronic specific heat of a metal. While doing the calculation properly is a bit tricky, it is reasonably easy to get the right order of magnitude.

Analogous to Eq. 3.8, the total energy of our system of electrons is given now by

$$E_{total} = \frac{2V}{(2\pi)^3} \int d\mathbf{k}\, \epsilon(\mathbf{k})\, n_F(\beta(\epsilon(\mathbf{k}) - \mu)) = \frac{2V}{(2\pi)^3} \int_0^\infty 4\pi k^2 dk\, \epsilon(\mathbf{k})\, n_F(\beta(\epsilon(\mathbf{k}) - \mu))$$

where the chemical potential is defined as above by

$$N = \frac{2V}{(2\pi)^3} \int d\mathbf{k}\, n_F(\beta(\epsilon(\mathbf{k}) - \mu)) = \frac{2V}{(2\pi)^3} \int_0^\infty 4\pi k^2 dk\; n_F(\beta(\epsilon(\mathbf{k}) - \mu))$$

(Here we have changed to spherical coordinates to obtain a one dimensional integral and a factor of $4\pi k^2$ out front).

It is convenient to replace k in this equation by the energy E by using Eq. 3.7 or equivalently

$$k = \sqrt{\frac{2\epsilon m}{\hbar^2}}$$

we then have

$$dk = \sqrt{\frac{m}{2\epsilon\hbar^2}}\; d\epsilon$$

We can then rewrite these expressions as

$$\begin{aligned} E_{total} &= V \int_0^\infty d\epsilon\; \epsilon\; g(\epsilon)\; n_F(\beta(\epsilon - \mu)) && (3.13) \\ N &= V \int_0^\infty d\epsilon\; g(\epsilon)\; n_F(\beta(\epsilon - \mu)) && (3.14) \end{aligned}$$

where

$$g(\epsilon)d\epsilon = \frac{2}{(2\pi)^3} 4\pi k^2 dk = \frac{2}{(2\pi)^3} 4\pi \left(\frac{2\epsilon m}{\hbar^2}\right) \sqrt{\frac{m}{2\epsilon\hbar^2}}\; d\epsilon = \frac{(2m)^{3/2}}{2\pi^2\hbar^3} \epsilon^{1/2} d\epsilon \tag{3.15}$$

is the *density of states per unit volume*, i.e., how many electron plane wave eigenstates there are in a energy range $d\epsilon$ per unit volume (and we have included both spin states here).

From Eq. 3.12 we can simply derive $(2m)^{3/2}/\hbar^3 = 3\pi^2 n/E_F^{3/2}$, thus we can simplify the density of states expression to

$$g(\epsilon) = \frac{3n}{2E_F}\left(\frac{\epsilon}{E_F}\right)^{1/2} \tag{3.16}$$

which is a fair bit simpler. Note that the density of states has dimensions of a density (an inverse volume) divided by an energy. It is clear that this is the dimensions it must have given Eq. 3.14 for example.

Note that the expression Eq. 3.14 should be thought of as defining the chemical potential given the number of electrons in the system and the temperature. Once the chemical potential is fixed, then Eq. 3.13 gives us the total kinetic energy of the system.

Recalling that for reasonable temperatures $T \ll T_F$ we should expand the Fermi factor n_F for low temperatures. However, this turns out to be a bit tricky (See Ashcroft and Mermin chapter 2 to see how it is done in detail). Nonetheless, we can roughly guess what is going to happen without too much work.

When $T = 0$ the Fermi function is a step function and the chemical potential is (by definition) the Fermi energy. For small T, the step function is smeared out as we see in Fig. 3.2. Note however, in this smearing the number of states that are removed from below the chemical potential is almost exactly the same as the number of states that are added above the chemical potential. Thus, for small T, one does not have to move the chemical potential from the Fermi energy in order to keep the number of particles fixed in Eq. 3.14. We conclude that $\mu \approx E_F$ for any low temperature. (In fact, in more detail we find that $\mu(T) = E_F + \mathcal{O}(T/T_F)^2$, see Ashcroft and Mermin chapter 2).

Thus we can focus on Eq. 3.13 with the assumption that $\mu = E_F$. At $T = 0$ let us call the kinetic energy[12] of the system $E(T = 0)$. At finite temperature, instead of a step function in Eq. 3.13 the step is smeared out as in Fig. 3.2. We see in the figure that only electrons within an energy range of roughly $k_B T$ of the Fermi surface can be excited – and in general they are excited above the Fermi surface by an energy of about $k_B T$. Thus we can approximately write

$$E(T) = E(T = 0) + (\gamma/2)[Vg(E_F)(k_B T)](k_B T) + \ldots$$

Here $Vg(E_F)$ is the density of states near the Fermi surface (Recall g is the density of states per unit volume), so the number of particles close enough to the fermi surface to be excited is $Vg(E_F)(k_B T)$, and the final factor of $(k_B T)$ is roughly the amount of energy that each one gets excited by. Here γ is an some constant which we cannot get right by such an approximate argument (but it can be derived more carefully, and it turns out that $\gamma = \pi^2/3$, see Ashcroft and Mermin).

We can then derive the heat capacity

$$C = \partial E/\partial T = \gamma k_B g(E_F) k_B T$$

which then using Eq. 3.16 we can rewrite as

$$C = \gamma \left(\frac{3Nk_B}{2}\right)\left(\frac{T}{T_F}\right)$$

---

[12] In fact $E(T = 0) = (3/5)NE_F$, which is not too hard to show. Try showing it!

The first term in brackets is just the classical result for the heat capacity of a gas, but the final factor T/T_F is tiny (0.01 or smaller!). This is the above promised linear T term in the specific heat of electrons, which is far smaller than one would get for a classical gas.

This Sommerfeld prediction for the electronic (linear T) contribution to the heat capacity of a metal is typically not far from being correct (The coefficient may be incorrect by factors of "order one"). A few metals, however, have specific heats that deviate from this prediction by as much as a factor of 10. Note that there are other measurements that indicate that these errors are associated with the electron mass being somehow changed in the metal. We will discover the reason for these deviations later in the course when we study band theory.

Realizing now that the specific heat of the electron gas is reduced from that of the classical gas by a factor of $T/T_F \lesssim 0.01$, we can return to the re-examine some of the above Drude calculations of thermal transport. We had above found (See Eq. 3.3-3.5) that Drude theory predicts a thermopower $S = \Pi/T = -c_v/(3e)$ that is too large by a factor of 100. Now it is clear that the reason for this error was that we used in this calculation (See Eq. 3.4) the specific heat per electron for a classical gas, which is too large by roughly $T_F/T \approx 100$. If we repeat the calculation using the proper specific heat, we will now get a prediction for thermopower which is reasonably close to what is actually measured in experiment for most metals.

We also used the specific heat per particle in the Drude calculation of the thermal conductivity $\kappa = \frac{1}{3} n c_v \langle v \rangle^2 \lambda$. In this case, the c_v that Drude used was too large by a factor of T_F/T, but on the other hand the value of $\langle v \rangle^2$ that he used was too small by roughly the same factor (Classically, one uses $mv^2/2 = k_B T$ whereas for the Sommerfeld model, one should use the Fermi velocity $mv_F^2/2 = k_B T_F$). Thus Drude's prediction for thermal conductivity came out roughly correct (and thus the Wiedemann-Franz law correctly holds).

3.2.3 Magnetic Spin Susceptibility (Pauli Paramagnetism)

Another property we can examine about the free electron gas is its response to an externally applied magnetic field. There are several ways that the electrons can respond to the magnetic field. First, the electrons motion can be curved due to the Lorentz force. We have discussed this previously, and we will return to discuss it again in section 15.2 below[13]. Secondly, the electron spins can flip over due to the applied magnetic field. Roughly, the Hamiltonian (neglecting the Lorentz force of the magnetic field) becomes[14].

$$\mathcal{H} = \frac{\mathbf{p}^2}{2m} + \mathfrak{g}\mu_B \mathbf{B} \cdot \boldsymbol{\sigma}$$

---

[13] For a free electron gas, the contribution to the magnetic susceptibility from the orbital motion of the electron is known as Landau diamagnetism and takes the value $\chi_{Landau} = -(1/3)\chi_{Pauli}$. This is a relatively tricky calculation. See Peierls' book for example. This effect is named after the famous Russian Nobel-Laureate Lev Landau, who kept a now famous ranking of how smart various physicist were — ranked on a logarithmic scale. Einstein was on top with a ranking of 0.5. Bose, Wigner, and Newton all received a ranking of 1. Schroedinger, Heisenberg, Bohr, and Dirac were ranked 2, and Landau modestly ranked himself a 2.5 but after winning the Nobel prize raised himself to 2. He said that anyone ranked below 4 was not worth talking to.

[14] Because the electron charge is negative, the electron dipole moment is actually opposite the direction of the electron spin (the current is rotating opposite the direction that the electron is spinning). Thus spins are lower energy when they are anti-aligned with the magnetic field! This is yet another annoyance caused by Benjamin Franklin who declared that the charge left on a glass rod when rubbed with silk is positive.

where $\mathfrak{g} = 2$ is the g-factor of the electron[15], $\mathbf{B}$ is the magnetic field[16] and $\boldsymbol{\sigma}$ is the spin of the electron which takes eigenvalues $\pm 1/2$. Here I have defined (and will use elsewhere) the useful version of the Bohr magneton

$$\mu_B = e\hbar/2m_e \approx .67(K/T)/k_B$$

the conventional Bohr magneton. Thus in the magnetic field the energy of an electron with spin up or down (with up meaning it points the same way as the applied field)

$$\begin{aligned} \epsilon(\mathbf{k},\uparrow) &= \frac{\hbar^2|\mathbf{k}|^2}{2m} + \mu_B|B| \\ \epsilon(\mathbf{k},\downarrow) &= \frac{\hbar^2|\mathbf{k}|^2}{2m} - \mu_B|B| \end{aligned}$$

The spin magnetization of the system (moment per unit volume) in the direction of the applied magnetic field will then be

$$M = -\frac{1}{V}\frac{dE}{dB} = -([\#\text{ up spins}] - [\#\text{ down spins}])\,\mu_B/V \qquad (3.17)$$

So when the magnetic field is applied, it is lower energy for the spins to be pointing down, so more of them will point up. Thus a magnetization develops in the same direction as the applied magnetic field. This is known as *Pauli Paramagnetism*. Here *Paramagnetism* means that the magnetization is in the direction of the applied magnetic field. *Pauli Paramagnetism* refers in particular to the spin magnetization of the free electron gas. (We will discuss paramagnetism in more detail in chapter 15).

Let us now calculate the Pauli paramagnetism of the free electron gas at $T = 0$. With zero magnetic field applied, both the spin up and spin down states are filled up to the Fermi energy (i.e, to the Fermi wavevector). Near the Fermi level the density of states per unit volume for spin up electrons is $g(E_F)/2$ and similarly the density of states per unit volume for spin down electrons is $g(E_F)/2$. When B is applied, the spin ups will be more costly by an energy $\mu_B B$. Thus, (assuming that the chemical potential does not change) we will have $(g(E_F)/2)\mu_B B$ fewer spin ups electrons per unit volume. Similarly, the spin downs will be less costly by the same amount, so we will have $(g(E_F)/2\mu_B B$ more spin downs per unit volume. Note that the total number of electrons in the system did not change, so our assumption that the chemical potential did not change is correct. (Recall that chemical potential is always adjusted so it gives the right total number of electrons in the system). Thus, using Eq. 3.17 the magnetization (magnetic moment per unit volume) is given by

$$M = g(E_F)\mu_B^2 B$$

and hence the magnetic susceptibility $\chi = \partial M/\partial H$ is given (at $T = 0$ by)[17]

$$\chi_{Pauli} = \frac{dM}{dH} = \mu_0\frac{dM}{dB} = \mu_0\mu_B^2 g(E_F)$$

with μ_0 the permeability of free space. In fact this result is not far from correct for simple metals such as Li,Cu, or Na.

---

[15] It is a yet another constant source of grief that the letter "g" is used both for density of states and for g-factor of the electron. On this page, I used $\mathfrak{g}$ for the g-factor, but I am probably not very consistent about this. I hope that the context will always make clear which is which. Similar grief is that we now have to write $\mathcal{H}$ for Hamiltonian because $H = B/\mu_0$ is frequently used for the magnetic field with μ_0 the permeability of free space.

[16] One should be careful to use the magnetic field seen by the actual electrons — this may be different from the magnetic field applied to the sample if the sample itself develops a magnetization.

[17] See also the very closely related derivation given in section 19.1.2 below.

3.2.4 Why Drude Theory Works so Well

In retrospect we can understand a bit more about why Drude theory was so successful. As mentioned above, we now realize that because of Fermi statistics, treating electrons as a classical gas is incorrect – resulting in a huge overestimation of the specific heat per particle,and in a huge underestimation of the typical velocity of particles. As described above, these two errors can sometimes cancel giving reasonable results nonetheless.

However, we can also ask why it is that Drude was successful in calculation of transport properties such as the conductivity and the Hall coefficient make sense. In these calculations neither the velocity of the particle nor the specific heat enter. But still, the idea that a single particle will accelerate freely for some amount of time, then will scatter back to zero momentum seems like it must be wrong, since the state at zero momentum is always fully occupied. The transport equation (Eq. 3.1) that we solve

$$\frac{d\mathbf{p}}{dt} = \mathbf{F} - \frac{\mathbf{p}}{\tau} \tag{3.18}$$

in the Drude theory describes the motion of each particle. However, we can just as well use the same equation to describe the motion of the center of mass of the entire Fermi sea! On the left of Fig. 3.3 we have a picture of a Fermi sphere of radius $\mathbf{k}_F$. The typical electron has a very large velocity on the order of the Fermi velocity v_F, but the average of all of the velocities is zero. When an electric field is applied (in the $\hat{y}$ direction as shown on the right of Fig. 3.3, so that the force is in the $\hat{y}$ direction since the charge of the electron is $-e$) every electron in the system accelerates together in the $-\hat{y}$ direction, and the center of the Fermi sea shifts. The shifted Fermi sea has some nonzero average velocity, known as the *drift velocity* $\mathbf{v}_{drift}$. Since the kinetic energy of the shifted Fermi sea is higher than the energy of the Fermi sea with zero average velocity, the electrons will try to scatter back (with scattering rate $1/\tau$) to lower kinetic energy and shift the Fermi sea back to its original configuration with zero drift velocity. We can then view the Drude transport equation (Eq. 3.18) as describing the motion of the average velocity (momentum) of the entire Fermi sea.

One can think about how this scattering actually occurs in the Sommerfeld model. Here, most electrons have nowhere to scatter to, since all of the available $\mathbf{k}$ states with lower energy (same $|\mathbf{k}|$ are already filled. However, the few electrons near the Fermi surface in the thin crescent between the shifted and unshifted Fermi sea have available states can scatter into the thin unfilled crescent on the other side of the unfilled Fermi sea (See Fig. 3.3). Although these scattering processes happen only to a very few of the electrons, the scattering events are extremely violent in that the change in momentum is exceedingly large (scattering all the way across the Fermi sea[18]).

3.2.5 Shortcomings of Free Electron Model

Although the Sommerfeld (Free Electron) Model of a metal explains quite a bit about metals, it remains incomplete. Here are some items that are not well explained within Sommerfeld theory:

- Magnetism: Some metals, such as Iron, are magnetic even without any applied external magnetic field. We will discuss this (at least to some extent) later in the course when we discuss magnetism.

---

[18] Actually, it may be that many small scatterings walking around the edge of these crescents make up this one effective scattering event

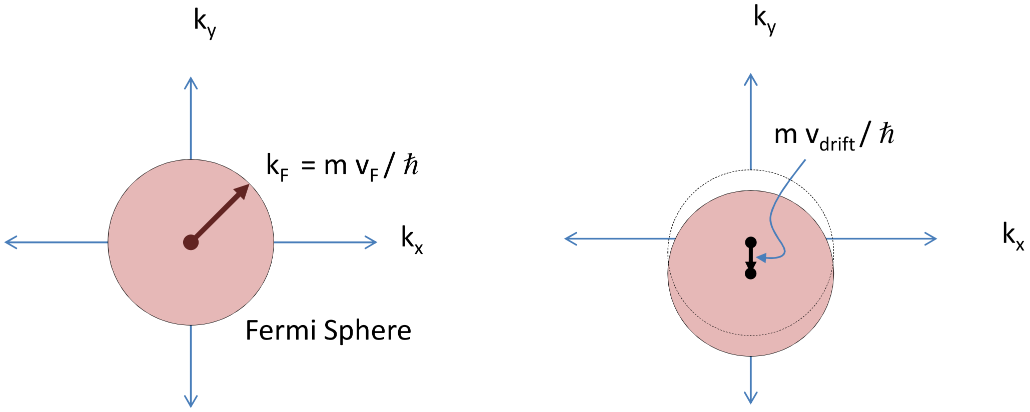


Figure 3.3: Drift Velocity and Fermi Velocity. The Drift momentum is the displacement of the entire fermi sphere (which is generally very very small) whereas the Fermi momentum is the radius of the Fermi sphere, which can be very large. Drude theory makes sense if you think of it as a transport equation for the center of mass of the entire Fermi sphere – i.e., it describes the drift velocity. Scattering of electrons only occurs between the thin crescents that are the difference between the shifted and unshifted Fermi spheres

- Electron interaction: We have treated the electrons as noninteracting fermions. In fact, the typical energy of interaction for electrons ($e^2/4\pi\epsilon_0 r$ with r the typical distance between electrons) is huge, roughly the same scale as the Fermi energy. Yet we have ignored the Coulomb interaction between electrons completely. Understanding why this works is an extremely hard problem that was only understood starting in the late 1950s – again due to the brilliance of Lev Landau (See above footnote 13 in this chapter about Landau). The theory that explains this is frequently known as "Landau Fermi Liquid Theory", but we will not study it in this course.

In addition to these two issues there are a number of shortcomings of the Sommerfeld theory of metals which will be explained later in this course when we study band structure:

- Having discovered now that the typical velocity of electrons v_F is extremely large, and being able to measure the scattering time τ, we obtain a scattering length $\lambda = v_F \tau$ that may be 100 Angstroms or more. One might wonder, if there are atoms every few angstroms in a metal, why do the electrons not scatter from these atoms? This is a fundamentally important issue which we will discuss later in the course (it is known as Bloch's theorem).

- Many of our results depend on the number of electrons in a metal. In order to calculate this number we have always used the chemical valence of the atom. (For example, we assume one free electron per Li atom). However, in fact, except for Hydrogen, there are actually many electrons per atom. Why do core electrons not "count" for calculating the Fermi energy or velocity? What about insulators where there are no electrons free?
- We have still not resolved the question of why the Hall effect sometimes comes out with the incorrect sign, as if the charge carrier were positive rather than negative (the sign of charge of electrons)
- In optical spectra of metals there are frequently many features (higher absorbtion at some frequencies, lower absorbtion at other frequencies). These features give metals there characteristic colors (for example, they make gold yellowish). The Sommerfeld model does not explain these features at all.
- The measured specific heat of electrons is much more correct than in Drude theory, but for some metals is still off by factors as large as 10. Measurements of the mass of the electron in a metal also sometimes gives answers that differ from the actual mass of the electron by similar factors.

All of these issues are a result of electronic *band structure* which we will discuss in chapters 7, 11 and subsequent chapters. In short, we are not taking seriously the periodic structure of atoms in materials.

3.2.6 Summary of (Sommerfeld) Free Electron Model

- Treats properly the fact that electrons are Fermions.
- High density of electrons results in extremely high Fermi energy and Fermi velocity. Thermal and electric excitations are small redistributions of electrons around the Fermi surface.
- Compared to Drude theory, obtains electron velocity ~ 100 times larger, but heat capacity per electron ~ 100 times smaller. Leaves Wiedemann-Franz ratio roughly unchanged from Drude, but fixes problems in predications of thermal properties.
- Specific Heat and (Pauli) paramagnetic susceptibility can be calculated explicitly (know these derivations!) in good agreement with experiment.

References

For free electron (Sommerfeld) theory, good references are:

- Ashcroft and Mermin chapter 2-3.
- Singleton, section 1.5-1.6
- Rosenberg section 7.1-7.9
- Ibach and Luth section 6-6.5
- Kittel chapter 6
- Burns chapter 9B (excluding 9.14 and 9.16)

Part II

What Holds Solids Together

Chapter 4

Chemical Bonding

In chapter 2 we found that the Debye model gave a reasonably good description of the specific heat of solids. However, we also found a number of shortcomings of the theory. These shortcomings basically stemmed from not taking seriously the fact that solids are actually made up of individual atoms assembled in a periodic structure.

Similarly in chapter 3 we found that the Sommerfeld model of metals described quite a bit about metals, but had a number of shortcomings as well — most of these (the ones we listed as being due to "band structure") are similarly due to not realizing that the solids are made up of individual atoms assembled in periodic structures.

As such, a large amount of this course will actually be devoted to understanding the effects of these atoms on electrons and on the vibrations of the solid. However, first it is worth backing up and asking ourselves why atoms stick together to form solids in the first place!

4.1 General Considerations about Bonding

To determine why atoms stick together to form solids, we are in some sense trying to describe the solution to a many particle Schroedinger equation describing the many electrons and many nuclei in a solid. We can at least write down the equation

$$H\Psi = E\Psi$$

where Ψ is the wavefunction describing the positions and spin states of all the electrons and nuclei in the system. The terms in the Hamiltonian include kinetic term (with inputs of the electron and nucleon mass) as well as a Coulomb interaction term between all the electrons and nuclei.[1]

While this type of description of chemical bonding is certainly true, it is also mostly useless. No one ever even tries to solve the Schroedinger equation for more than a few particles at a time. Trying to solve it for 10^{23} electrons simultaneously is completely absurd. One must try to extract useful information about the behavior from simplified models in order to obtain a qualitative understanding. (This is a great example of what I was ranting about in chapter 1 – reductionism

---

[1] To have a fully functioning "theory of everything" as far as all of chemistry, biology, and most of everything that matters to us (besides the sun and atomic energy) is concerned, one needs only Coulomb interaction plus the Kinetic term in the Hamiltonian, plus spin-orbit (relativistic effects) for some of the heavy atoms.

does not work — saying that the Schroedinger equation is the whole solution is misguided). More sophisticated techniques try to turn these qualitative understandings into quantitative predictions.

In fact, what we are trying to do here is to try to understand a whole lot of chemistry from the point of view of a physicist. Instead of learning empirical chemistry rules, we will look at simplified models that show roughly how these rules arise. However at the end of the day, we cannot trust our simplified models too much and we really should learn more chemistry to try to decide if Yttrium really will form a carbonate salt or some similar question.

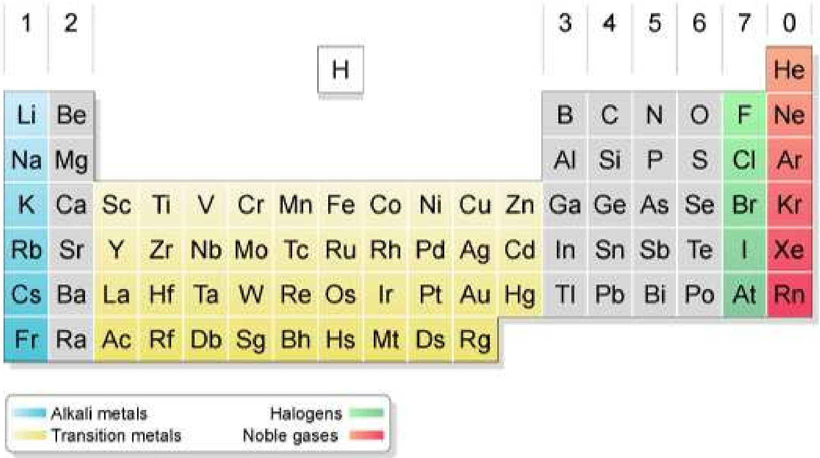


Figure 4.1: The periodic table of the elements.

From a chemist's point of view one frequently thinks about different *types* of chemical bonds depending on the types of atoms involved, and in particular, depending on the atom's position on the periodic table (And in particular, on the atom's *electronegativity* — which is its tendency to attract electrons) . Below we will discuss Ionic Bonds, Covalent Bonds, van der Waals (fluctuating dipole, or molecular) bonds, Metallic Bonds, and Hydrogen Bonds. Of course, they are all different aspects of the Schroedinger equation, and any given material may exhibit aspects of several of these types of bonding. Nonetheless, qualitatively it is quite useful to discuss these different types of bonds to give us intuition about how chemical bonding can occur. A brief description of the many types of bonding and their properties is shown in table 4.1. Note that this table should be considered just as rules-of-thumb, as many materials have properties intermediate between the categories listed.

In this section we will try to be a bit more quantitative about how some of these types of bonding come about. Remember, underneath it is all the Schroedinger equation and the coulomb interaction between electrons and nuclei that is holding materials together!

4.2 Ionic Bonds

The general idea of an ionic bond is that for certain compounds (for example, binary compounds, such as NaCl, made of one element in group I (or II) and one element in group VII), it is energetically favorable for an electron to be physically trasferred from one atom to the other, leaving two

Type of Bonding	Description	Typical of which compounds	Typical Properties
Ionic	Electron is transferred from one atom to another, and the resulting ions attract each other	Binary compounds made of constituents with very different electronegativity: Ex, group 1-7 such as NaCl or group 2-7 compounds.	• Hard, Very Brittle • High Melting Temperature • Electrical Insulator • Water Soluble
Covalent	Electron is shared equally between two atoms forming a bond. Energy lowered by delocalization of wavefunction	Compounds made of constituents with similar electronegativities (ex, 3-5 compounds such as GaAs), or solids made of one element only such as Diamond (C)	• Very Hard (Brittle) • High Melting Temperature • Electrical Insulators or Semiconductors
Metallic Bonds	Electrons delocalized throughout the solid forming a glue between positive ions.	Metals. Left and Middle of Periodic Table.	• Ductile, Maleable (due to non-directional nature of bond. Can be hardened by preventing dislocation motion with impurities) • Lower Melting Temperature • Good electrical and thermal conductors.
Molecular (van der Waals or Fluctuating Dipole)	No transfer of electrons. Dipole moments on constituents align to cause attraction. Bonding strength increases with size of molecule or polarity of constituent.	Nobel Gas Solids, Solids made of Non-Polar (or slightly polar) Molecules Binding to Each Other (Wax)	• Soft, Weak • Low Melting Temperature • Electrical Insulators
Hydrogen	Involves Hydrogen ion bound to one atom but still attracted to another. Special case because H is so small.	Important in organic and biological materials	• Weak Bond (stronger than VdW though) • Important for maintaining shape of DNA and proteins

Table 4.1: Types of Bonds in Solids. This table should be thought of as providing rough rules. Many materials show characteristics intermediate between two (or more!) classes.

oppositely charged ions which then attract each other. One writes a chemical "reaction" of the form

$$\mathrm{Na} + \mathrm{Cl} \rightarrow \mathrm{Na}^+ + \mathrm{Cl}^- \rightarrow \mathrm{NaCl}$$

To find out if such a reaction happens, one must look at the energetics associated with the transfer of the electron.

At least in principle it is not *too* hard to imagine solving the Schroedinger equation for a single atom and determining the energy of the neutral atom, of the positive ion, and of the negative ion. We define:

Ionizaton Energy = Energy required to remove one electron from an atom to create a positive ion
Electron Affinity = Energy gain for creating negative ion from the atom by adding an electron

(To be precise, in both cases we are comparing the energy of having an electron either at position infinity, or on the atom).

Ionization energy is smallest on the left (group I and II) of the periodic table and largest on the right (group VII, and VIII). To a lesser extent the ionization energy also tends to decrease towards the bottom of the periodic table. Similarly electron Affinity is also largest on the right and top of the periodic table (not including the group VIII nobel gases which roughly do not attract electrons measurably at all).

Thus the total energy change of transferring an electron from atom A to atom B is

$$\Delta E_{A+B\rightarrow A^+ + B^-} = (\text{IonizationEnergy})_A - (\text{Electron Affinity})_B$$

(Note carefully the sign. The ionization energy is a positive energy that must be put in, the electron affinity is an energy that comes out).

However this ΔE is the energy to transfer an electron between two atoms very far apart. In addition, there is also

$$\text{Cohesive Energy} = \text{Energy gain from } A^+ + B^- \rightarrow AB$$

This cohesive energy is mostly a classical effect of the Coulomb interaction between the ions as one lets the ions come close together.

The total energy gain for forming a molecule from the two individual atoms is given by

$$\Delta E_{A+B\rightarrow AB} = (\text{Ionization Energy})_A - (\text{Electron Affinity})_B - \text{Cohesive Energy of A-B}$$

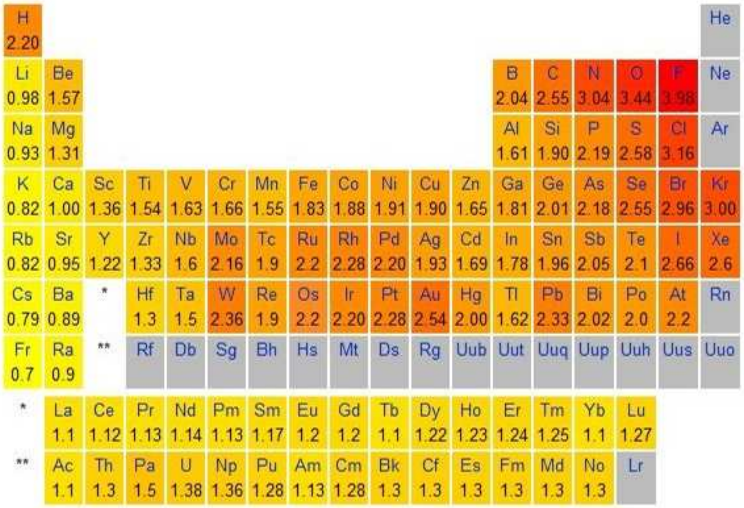
One obtains ionic bond if the total ΔE for this process is less than zero.

In order to determine whether an electron is likely to be transferred between one atom and another, it is convenient to use the a so-called *electronegativity*, which roughly describes how much an atom "wants" electrons, or how much an atom attracts electrons to itself. While there are various definitions of electronegativity that are used, a simple and useful definition is known as the Mulliken Electronegativity[2]

$$\text{(Mulliken) Electronegativity} = \frac{(\text{Electron Affinity}) + (\text{Ionization Energy})}{2}$$

---

[2] Both Robert Mulliken and Linus Pauling won Nobel Prizes in Chemistry for their work understanding chemical bonding. Pauling won a second Nobel prize, in Peace, for his work towards banning nuclear weapons testing. (Only four people have ever won two Nobels: Marie Curie, Linus Pauling, John Bardeen, and Fredrick Sanger. We should all know these names!). , Pauling was criticized later in his life for promoting high doses of vitamin C to prevent cancer and other ailments, sometimes apparently despite scientific evidence to the contrary.



H 2.20																	He
Li 0.98	Be 1.57											B 2.04	C 2.55	N 3.04	O 3.44	F 3.98	Ne
Na 0.93	Mg 1.31											Al 1.61	Si 1.90	P 2.19	S 2.58	Cl 3.16	Ar
K 0.82	Ca 1.00	Sc 1.36	Ti 1.54	V 1.63	Cr 1.66	Mn 1.55	Fe 1.83	Co 1.88	Ni 1.91	Cu 1.90	Zn 1.65	Ga 1.81	Ge 2.01	As 2.18	Se 2.55	Br 2.96	Kr 3.00
Rb 0.82	Sr 0.95	Y 1.22	Zr 1.33	Nb 1.6	Mo 2.16	Tc 1.9	Ru 2.2	Rh 2.28	Pd 2.20	Ag 1.93	Cd 1.69	In 1.78	Sn 1.96	Sb 2.05	Te 2.1	I 2.66	Xe 2.6
Cs 0.79	Ba 0.89	*	Hf 1.3	Ta 1.5	W 2.36	Re 1.9	Os 2.2	Ir 2.20	Pt 2.28	Au 2.54	Hg 2.00	Tl 1.62	Pb 2.33	Bi 2.02	Po 2.0	At 2.2	Rn
Fr 0.7	Ra 0.9	**	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Uub	Uut	Uuq	Uup	Uuh	Uus	Uuo
*	La 1.1	Ce 1.12	Pr 1.13	Nd 1.14	Pm 1.13	Sm 1.17	Eu 1.2	Gd 1.2	Tb 1.1	Dy 1.22	Ho 1.23	Er 1.24	Tm 1.25	Yb 1.1	Lu 1.27		
**	Ac 1.1	Th 1.3	Pa 1.5	U 1.38	Np 1.36	Pu 1.28	Am 1.13	Cm 1.28	Bk 1.3	Cf 1.3	Es 1.3	Fm 1.3	Md 1.3	No 1.3	Lr		

Periodic table of electronegativity using the Pauling scale

Figure 4.2: The Pauling scale of Electronegativities (Very similar to the Mulliken Electronegativity scale. I couldn't find a nice table of Mulliken. But they are very similar). The trend in electronegativities is that the upper right of the periodic table is very electronegative, whereas the lower left is not electronegative (or is electropositive).

In bonding, the electron is always transferred from the atom of lower electronegativity to higher electronegativity. The greater the difference in electronegativities between two atoms the more completely the electron is transferred from one atom to the other. If the difference in electronegativities is small, then the electron is only partially transferred from one atom to the other. We will see below that one can have co-valent bonding even between two identical atoms where there is no difference in electronegativities, and therefore no net transfer of electrons. Before leaving the topic of ionic bonds, it is worth discussing some of the typical physics of ionic solids. First of all, the materials are typically hard, as the Coulomb interaction between oppositely charged ions is strong. However, since water is extremely polar, it can dissolve an ionic solid. This happens (See Fig) by arranging the water molecules such that the negative side of the molecule is close to the positive ions and the positive side of the molecule is close to the negative ions.

4.3 Covalent Bond

Roughly, a covalent bond is a bond where electrons are shared equally between two atoms. There are several pictures that can be used to describe the covalent bond.

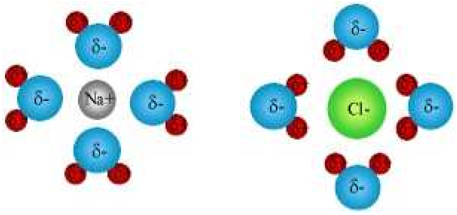


Figure 4.3: Salt, NaCl, dissolved in water. Ionic compounds typically dissolve easily in water since the polar water molecules can screen the highly charged, but otherwise stable, ions.

4.3.1 Particle in a Box Picture

Let us model a hydrogen atom as a box of size L for an electron (for simplicity, let us think about a one dimensional system). The energy of a single electron in a box is (I hope this looks familiar!)

$$E = \frac{\hbar^2 \pi^2}{2mL^2}$$

Now suppose two such atoms come close together. An electron that is shared between the two atoms can now be delocalized over the positions of both atoms, thus it is in a box of size $2L$ and has lower energy

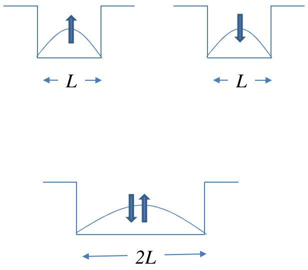
$$E = \frac{\hbar^2 \pi^2}{2m(2L)^2}$$

This reduction in energy that occurs by delocalizing the electron is the driving force for forming the chemical bond. If each atom starts with a single electron (i.e., it is a hydrogen atom) then when the two atoms come together to form a lower energy orbital, both electrons can go into this same ground state orbital since they can take opposite spin states. (Of course the reduction in energy of the two electrons must compete against the coulomb repulsion between the two nuclei, and the coulomb repulsion of the two electrons with each other, which is a much more complicated calculation). The new ground state orbital is known as a *bonding* orbital.

However if we had started with two Helium atoms, where each atom has two electrons, then when the two atoms come together there is not enough room in the single ground state wavefunction. In this case, two of the four electrons must occupy the first excited orbital – which in this case turns out to be exactly the same electronic energy as the original ground state orbital of the original atoms – since no energy is gained by these electrons when the two atoms come together these are known as *antibonding* orbitals. (In fact it requires energy to push the two atoms together if one includes Coulomb repulsions between the nuclei)

4.3.2 Molecular Orbital or Tight-Binding Theory

In this section we make slightly more quantitative some of the idea of the previous section. Let us write a Hamiltonian for two Hydrogen atoms. Since the nuclei are heavy compared to the electrons,



$$E = \frac{\hbar^2\pi^2}{2mL^2}$$

$$E_{bonding} = \frac{\hbar^2\pi^2}{2m(2L)^2}$$

$$E_{anti-bonding} = \frac{\hbar^2(2\pi)^2}{2m(2L)^2}$$

Figure 4.4: Particle in a box picture of covalent bonding. Two separated hydrogen atoms are like two different boxes each with one electron in the lowest eigenstate. When the two boxes are pushed together, one obtains a larger box – thereby lowering the energy of the lowest eigenstate – which is known as the *bonding* orbital. The two electrons can take opposite spin states and can thereby both fit in the bonding orbital. The first excited state is known as the *antibonding* orbital

we will fix their positions and solve the Schroedinger equation for the electrons as a function of the distance between the nuclei. This is known as a "Born-Oppenheimer" approximation[3,4]. We hope to calculate the eigenenergies of the system as a function of the distance between the positively charged nuclei.

For simplicity, let us consider a single electron and two identical positive nuclei[5]. We write the Hamiltonian as

$$H = K + V_1 + V_2$$

with

$$K = \frac{\mathbf{p}^2}{2m}$$

being the kinetic energy of the electron and

$$V_i = \frac{e^2}{4\pi\epsilon_0|\mathbf{r} - \mathbf{R}_i|}$$

is the Coulomb interaction energy between the electron position $\mathbf{r}$ and the position of nuclei $\mathbf{R}_i$.

---

[3] Max Born (also the same guy from Born-Von Karmen boundary conditions) was one of the founders of quantum physics, winning a Nobel Prize in 1954. His daughter, and biographer, Irene, married into the Newton-John family, and had a daughter named Olivia, who became a pop icon and film star in the 1970s. Her most famous role was in the movie of *Grease* playing Sandra-Dee opposite John Travolta. When I was a kid, she was every teenage guy's dream-girl (her, or Farrah Fawcett).

[4] J. Robert Oppenheimer later became the head scientific manager of the American atomic bomb project during the second world war. After this giant scientific and military triumph, he pushed for control of nuclear weapons leading to his being accused of being a communist sympathizer during the "Red" scares of the 1950s and he ended up having his security clearance revoked.

[5] The adventurous student might wish to think about what happens if the two positive nuclei are not identical. Using the tight binding picture, we can allow a different atomic potential ϵ_0 for the two different atoms in Eq. 4.1 below. Depending on how different these potentials are, the resulting bond can be anywhere between covalent and ionic.

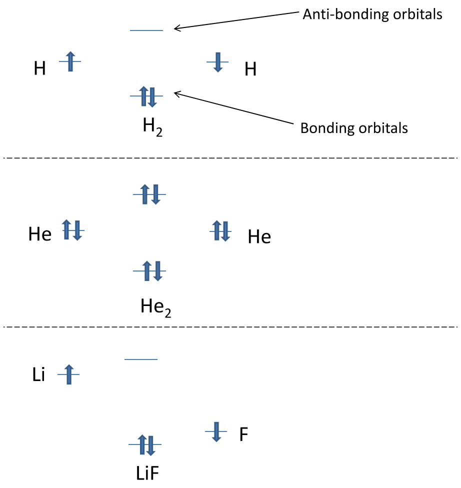


Figure 4.5: Molecular Orbital Picture of Bonding. In this type of picture, on the far left and far right are the orbital energies of the individual atoms well separated from each other. In the middle are the orbital energies when the atoms come together to form a molecule. Top: Two Hydrogen atoms come together to form a H_2 molecule. As mentioned above in the particle-in-a-box picture, the lowest energy eigenstate is reduced in energy when the atoms come together and both electrons go into this bonding orbital. Middle: In the case of Helium, since there are two electrons per atom, the bonding orbitals are filled, and the antibonding orbitals must be filled as well. The total energy is not reduced by the two Helium atoms coming together (thus Helium does not form He_2). Bottom: In the case of LiF, the energies of the Lithium and the Flourine orbitals are different. As a result, the bonding orbital is mostly composed of the orbital on the Li atom – meaning that the bonding electrons are mostly transferred from Li to F — forming a more ionic bond.

Generally this type of Schroedinger equation is hard to solve exactly. (In fact it can be solved exactly in this case, but it is not particularly enlightening to do so). Instead, we will attempt a variational solution. Let us write a trial wavefunction as

$$|\psi\rangle = \phi_1|1\rangle + \phi_2|2\rangle$$

where ϕ_i are complex coefficients, and the kets $|1\rangle$ and $|2\rangle$ are known as "atomic orbitals" or "tight-binding" orbitals. These can be taken as the ground state solution of the Schroedinger equation when there is only one nucleus present. I.e.

$$\begin{aligned}(K+V_1)|1\rangle &= \epsilon_0|1\rangle \\ (K+V_2)|2\rangle &= \epsilon_0|2\rangle \end{aligned} \tag{4.1}$$

where ϵ_0 is the ground state energy of the single atom.

For simplicity, we will now make a bad approximation that $|1\rangle$ and $|2\rangle$ are orthogonal so we can then choose a normalization such that

$$\langle i|j\rangle = \delta_{ij} \tag{4.2}$$

When the two nuclei get very close together, this is clearly not correct. We then have to decide: Either we keep our definition of the atomic orbitals being the solution to the Schroedinger equation for a single nucleus, but we give up on the two atomic orbitals being orthogonal; or we can give up on the orbitals being solutions to the Schroedinger equation for a single nucleus, but we keep orthonormality. For a homework problem, we will consider what happens when we give up orthonormality, and we will see that most of what we learn does not depend too much on whether the orbitals are orthogonal or not. However for now, let us stick with orthonormal orbitals.

An effective Schroedinger equation can be written down for our variational wavefunction which (unsuprisingly) takes the form of an eigenvalue problem

$$\sum_j H_{ij}\phi_j = E\phi_i$$

where

$$H_{ij} = \langle i|H|j\rangle$$

is a two by two matrix in this case. (The equation generalizes in the obvious way to the case where there are more than 2 orbitals). The proof that this is the correct equation to solve is also a homework problem.

Recalling our definition of $|1\rangle$ as being the ground state energy of $K + V_1$, we can write

$$\begin{aligned}
H_{11} &= \langle 1|H|1\rangle = \langle 1|K+V_1|1\rangle + \langle 1|V_2|1\rangle = \epsilon_0 + V_{cross} && (4.3)\\
H_{22} &= \langle 2|H|2\rangle = \langle 2|K+V_2|2\rangle + \langle 2|V_1|2\rangle = \epsilon_0 + V_{cross} && (4.4)\\
H_{12} &= \langle 1|H|2\rangle = \langle 1|K+V_2|2\rangle + \langle 1|V_2|2\rangle = 0 \;-\; t && (4.5)\\
H_{21} &= \langle 2|H|1\rangle = \langle 2|K+V_2|1\rangle + \langle 2|V_1|1\rangle = 0 \;-\; t^* && (4.6)
\end{aligned}$$

In the first two lines

$$V_{cross} = \langle 1|V_2|1\rangle = \langle 2|V_2|2\rangle$$

is the Coulomb potential felt by orbital $|1\rangle$ due to nucleus 2, or equivalently the Coulomb potential felt by orbital $|2\rangle$ due to nucleus 1. In the second two lines (Eqs. 4.5 and 4.6) we have also defined the so-called *hopping* term[6]

$$t = -\langle 1|V_2|2\rangle$$

(the reason for the name hopping will become clear below). Also in the second two lines (Eqs. 4.5 and 4.6) the first term vanishes because of orthogonality of $|1\rangle$ and $|2\rangle$.

Thus our Schroedinger equation is reduced to a two by two matrix equation of the form

$$\begin{pmatrix} \epsilon_0 + V_{cross} & -t \\ -t^* & \epsilon_0 + V_{cross} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = E \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$$

The interpretation of this equation is roughly that orbitals $|1\rangle$ and $|2\rangle$ both have energies ϵ_0 which is shifted by V_{cross} due to the presence of the other nuclei. In addition the electron can "hop" from one orbital to the other by the off-diagonal t term. To understand this interpretation more fully, we realize that in the *time dependent* Schroedinger equation, if the matrix were diagonal an wavefunction that started completely in orbital $|1\rangle$ would stay on that orbital for all time. However, with the off-diagonal term, the time wavefunction can oscillate between the two orbitals.

---

[6] The minus sign is a convention for the definition of t. For many cases of interest, this definition makes t positive, although it can actually have either sign depending on the structure of the orbitals in question and the details of the potential.

Diagonalizing this two-by-two matrix we obtain eigenenergies

$$E_\pm = \epsilon_0 + V_{cross} \pm |t|$$

the lower energy orbital is the "bonding" whereas the higher energy orbital is the anti-bonding. The corresponding wavefunctions are then

$$\begin{aligned}\psi_{bonding} &= \frac{1}{\sqrt{2}}(\phi_1 \pm \phi_2) \quad (4.7)\\ \psi_{anti-bonding} &= \frac{1}{\sqrt{2}}(\phi_1 \mp \phi_2) \quad (4.8)\end{aligned}$$

I.e., these are the symmetric and antisymmetric superposition of orbitals. The signs ± and ∓ depend on the sign of t, where the lower energy one is always called the bonding orbital and the higher energy one is called antibonding. To be precise $t > 0$ makes $(\phi_1 + \phi_2)/\sqrt{2}$ the lower energy bonding orbital. Roughly one can think of these two wavefunctions as being the lowest two "particle-in-a-box" orbitals — the lowest energy wavefunction does not change sign, whereas the first excited state changes sign once (for the case of $t > 0$ the analogy is precise).

Aside: In section 19.4 below, we will consider a more general tight binding model with more than one electron in the system and with Coulomb interactions between electrons as well. That calculation is more complicated, but shows very similar results. That calculation is also non-examinable, but might be fun to read for the adventurous.

Note again that V_{cross} is the energy that the electron on orbital 1 feels from nucleus 2. However, we have not included the fact that the two nuclei also interact, and to a first approximation, this Coulomb repulsion between the two nuclei will precisely cancel the attractive energy between the nucleus and the electron on the opposite orbital. Thus, including this energy we will obtain

$$\tilde{E}_\pm \approx \epsilon_0 \pm |t|$$

As the nuclei get closer together, the hopping term $|t|$ increases, giving an energy level diagram as shown in Fig. 4.6. This picture is obviously unrealistic, as it suggests that two atoms should bind together at zero distance between the nuclei. The problem here is that our assumptions and approximations begin to break down as the nuclei get closer together (for example, our orbitals are no longer orthogonal, V_{cross} does not exactly cancel the coulomb energy between nuclei, etc.).

A more realistic energy level diagram for the bonding and antibonding states is given in Fig. 4.7. Note that the energy diverges as the nuclei get pushed together (this is from the coulomb repulsion between nuclei). As such there is a minimum energy of the system when the nuclei are at some nonzero distance apart from each other, which then becomes the ground state distance of the nuclei in the resulting molecule.

Aside: In Fig. 4.7 there is a minimum of the bonding energy when the nuclei are some particular distance apart. This optimal distance will be the distance of the bond between two atoms. However, at finite temperature, the distance will fluctuate around this minimum (think of a particle in a potential well at finite temperature). Since the potential well is steeper on one side than on the other, at finite temperature, the "particle" in this well will be able to fluctuate to larger distances a bit more than it is able to fluctuate to

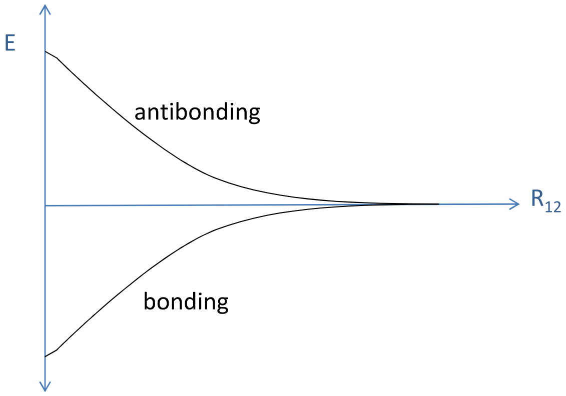


Figure 4.6: Model Tight Binding Energy Levels as a distance between the nuclei of the atoms.

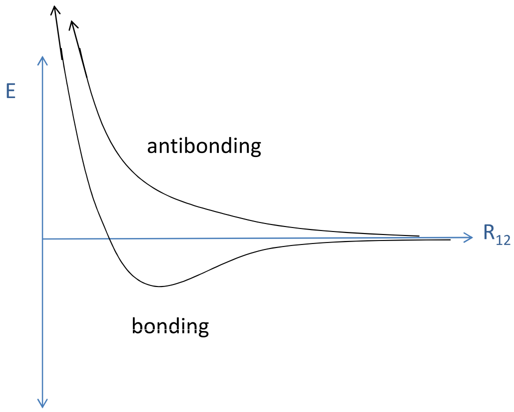


Figure 4.7: More realistic energy levels as a distance between the nuclei of the atoms.

smaller distances. As a result, the average bond distance will increase at finite temperature. This thermal expansion will be explored again in the next chapter and in detail as a homework problem.

Covalently bonded materials tend to be strong and tend to be electrical semiconductors or insulators (since electrons are tied up in the local bonds). The directionality of the orbitals makes these materials retain their shape well (non-ductile) so they are brittle. They do not dissolve in polar solvents such as water in the same way that ionic materials do.

4.4 Van der Waals, Fluctuating Dipole Forces, or Molecular Bonding

When two atoms (or two molecules) are very far apart from each other, there remains an attraction between them due to what is known as van der Waals[7] forces, sometimes known as fluctuating dipole forces, or molecular bonding. In short, both atoms have a dipole moment, which may be zero on average, but can fluctuate momentarily due to quantum mechanics. If the first atom obtains a momentary dipole moment, the second atom can polarize – also obtaining a dipole moment to lower its energy. As a result, the two atoms will attract each other.

This type of bonding between atoms is very typical of inert atoms (such as Noble gases: He, Ne, Kr, Ar, Xe) whose electrons do not participate in covalent bonds or ionic bonds. It is also typical of bonding between molecules that are already formed such as N_2 where there is little possibility for the electrons in this molecule to form covalent or ionic bonds between molecules. This bonding is weak compared to covalent or ionic bonds, but it is also long ranged in comparison since the electrons do not need to hop between atoms.

To be more quantitative, let us consider an electron orbiting a nucleus (say, a proton). If the electron is at a fixed position, there is a dipole moment $\mathbf{p} = e\mathbf{r}$ where $\mathbf{r}$ is the vector from the electron to the proton. With the electron "orbiting" (i.e, in an eigenstate), the average dipole moment is zero. However, if an electric field is applied to the atom, the atom will develop a polarization (i.e., it will be more likely for the electron to be found on one side of the nucleus than on the other). We write

$$\mathbf{p} = \chi \mathbf{E}$$

where χ is known as the polarizability (also known as electric susceptibility). This polarizability can be calculated, for, say a hydrogen atom explicitly. At any rate, it is some postive quantity.

Now, let us suppose we have two such atoms, separated a distance r in the $\hat{x}$ direction. Suppose one atom momentarily has a dipole moment $\mathbf{p}_1$ (for definiteness, suppose this dipole moment is in the $\hat{z}$ direction). Then the second atom will feel an electric field

$$E = \frac{p_1}{4\pi\epsilon_0 r^3}$$

in the negative $\hat{z}$ direction. The second atom then, due to its polarizability, develops a dipole moment $p_2 = \chi E$ which in turn is attracted to the first atom. The potential energy between these two dipoles is

$$U = \frac{-|p_1||p_2|}{4\pi\epsilon_0 r^3} = \frac{-p_1 \chi E}{(4\pi\epsilon_0) r^3} = \frac{-|p_1|^2 \chi}{(4\pi\epsilon_0 r^3)^2} \tag{4.9}$$

corresponding to a force $-dU/dr$ which is attractive and proportional to $1/r^7$.

You can check that independent of the direction of the original dipole moment, the force is always attractive and proportional to $1/r^7$. Although there will be a (nonnegative) prefactor which depends on the angle between the dipole moment $\mathbf{p}_1$ and $\mathbf{x}$ the direction between the two atoms.

Note. This argument appears to depend on the fact that the dipole moment of the first atom is nonzero. On average the atom's dipole moment will be zero. However in Eq. 4.9 in fact what

---

[7] J. D. van der Waals was awarded the Nobel prize in Physics in 1910 for his work on the structure of Liquids and Gases. You may remember the van der Waals equation of state from your thermodynamics course last year. There is a crater named after him on the far side of the moon.

enters is $|p_1|^2$ which has a nonzero expectation value. (In fact this is precisely the calculation that $\langle x \rangle$ for an electron in a hydrogen atom is zero, but $\langle x^2 \rangle$ is nonzero). A problem on using quantum mechanics to estimate the strength of fluctuating dipole bonds is provided as an additional exercise.

While these fluctuating dipolar forces are generally weak, they are the only forces that occur when electrons cannot be shared between atoms — either in the case where the electrons are not chemically active or when the atoms are far apart. In fact, the forces can be quite strong. A well known example of a van der Waals force is the force that allows lizards, such as Geckos to climb up walls. They appear to have hair on their feet that makes very close contact with the atoms of the wall, and they can climb up the walls with only van der Waals forces!

4.5 Metallic Bonding

It is sometimes hard to distinguish metallic bonding from covalent bonding. Roughly, however, one defines a metallic bond to be the bonding that occurs in metal. Similar to covalent bonds in the sense that electrons are shared between atoms, the actual electrons become delocalized throughout the crystal (we will discuss how this occurs in section 7.2 below).

In a metal, electrons can be completely delocalized, and these free electrons hold together the positive ions that they have left behind.

Since the electrons are completely delocalized, the bonds in metals tend not to be directional. Metals are thus often ductile and malleable. Since the electrons are free, metals are good conductors of electricity as well as of heat.

4.6 Hydrogen bonds

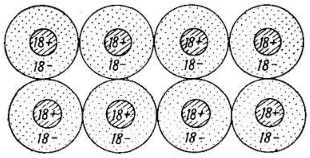
The hydrogen atom is extremely special due to its very small size. As a result, the bonds formed with hydrogen atoms are qualitatively different from other bonds. When the hydrogen atom forms a covalent or ionic bond with a larger atom, being small, it simply sits on the surface of its partner. This then makes the molecule (hydrogen and its partner) into a dipole. These dipoles can then attract each other as usual.

A very good example of the hydrogen bond is water, H_2O. Each oxygen atom is bound to two hydrogens (however because of the atomic orbital structure, these atoms are not collinear). The hydrogens, with their positive charge remain attracted to oxygens of other water molecules. In ice, these attractions are strong enough to form a weak, but stable bond between water molecules, thus forming a crystal. Sometimes one can think of the the hydrogen atom as forming "half" a bond with two oxygen atoms, thus holding the two oxygen atoms together.

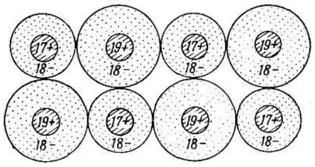
Hydrogen bonding is extremely important in biological molecules where, for example, hydrogen bonds hold together strands of DNA.

4.7 Summary of Bonding (Pictoral)

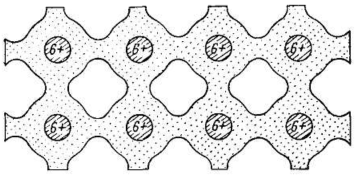
See also the table 4.1 for a summary of bonding types.



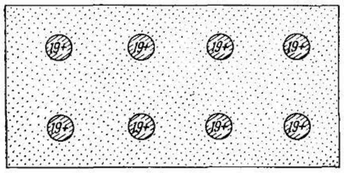
Argon Crystal. Molecular (Flucuating dipole) Bonds. No transfer of charge.



Potassium Chloride (KCl). Ionic Crystal. An electron is transferred from the K to the Cl to make K+ and Cl- which then attract each other



Carbon Crystal (Covalent Bonds). Each Atom shares electrons with its neighbors.



Metallic bonding in Potassium. Electrons are completely delocalized into an electron gas which holds together the ions.

Figure 4.8: Cartoons of Bonding Types

References on Chemical Bonding

- Rosenberg, section 1.11-1.19
- Ibach and Luth, chapter 1
- Hook and Hall, section 1.6
- Kittel, chapter 3 up to elastic strain
- Ashcroft and Mermin, chapters 19-20
- Burns, section 6.2-6.6 and also chapters 7 and 8.

Probably Ashcroft and Mermin as well as Burns chapters 7, and 8 are too much information.

4.8 Types of Matter

Once we understand how it is that atoms bond together, we can examine what types of matter can be formed. An obvious thing that can happen is that atoms can bond together the form regular crystals. A crystal is made of small units reproduced many times and built into a regular array. The macroscopic morphology of a crystal can reflect its underlying structure (See Fig. 4.9) We will spend much of the remainder of the term studying crystals.

Another form of matter is liquid. Here, atoms are attracted to each other, but not so strongly that they form permanent bonds (or the temperature is high enough to make the bonds unstable). Liquids (and gases) [8] are disordered configurations of molecules where the molecules

---

[8] As we learned in stat-mech and thermo courses, there is no “fundamental” difference between a liquid and a gas.

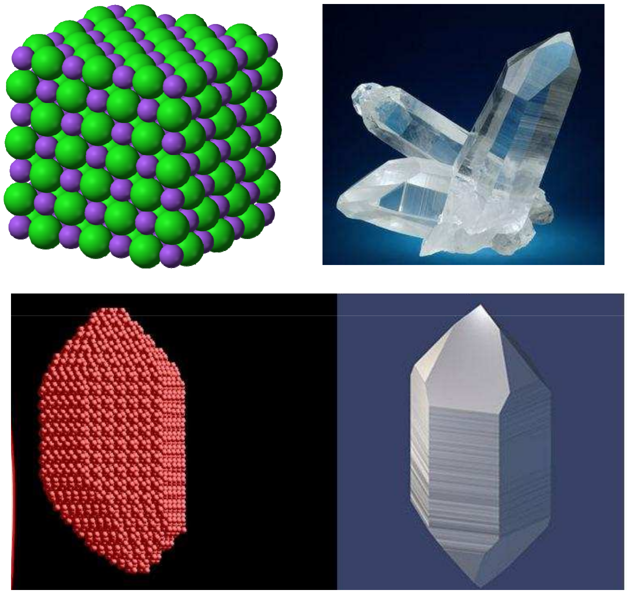


Figure 4.9: Crystals: Top left: Small units (One green one blue) reproduced periodically to form a crystal. Top right: A crystal of quartz. Bottom: The macroscopic morphology of a crystal reflects its underlying structure.

are free to move around into new configurations.

Somewhere midway between the idea of a crystal and the idea of a liquid is the possibility of amorphous solids and glasses. In this case the atoms are bonded into position in a disordered configuration. Unlike a liquid, the atoms cannot flow freely.

Many more possibilities exist. For example, one may have so-called liquid-crystals, where the system orders in some ways but remains disordered in other ways. For example, in figure 4.12 the system is crystalline (ordered) in one direction, but remains disordered within each plane. One can also consider cases where the molecules are always oriented the same way but are at completely random positions (known as a “nematic”). There are a huge variety of possible liquid crystal phases of matter. In every case it is the interactions between the molecules (“bonding” of some type, whether it be weak or strong) that dictates the configurations.

One should also be aware of polymers, which are long chains of atoms (such as DNA)[9].

Generally liquids are high density and not very compressible, whereas gases are low density and very compressible. A single substance (say, water) may have a phase transition between its gas and liquid phase (boiling), but one can also go continuously from the gas to liquid phase without boiling by going to high pressure and going around the critical point (becoming “supercritical”).

[9] Here is a really cool experiment to do in your kitchen. Cornstarch is a polymer — a long chain of atoms. Take a box of cornstarch and make a mixture of roughly half cornstarch and half water (you may have to play with the proportions). The concoction should still be able to flow. And if you put your hand into it, it will feel like a liquid and be gooey. But if you take a tub of this and hit it with a hammer very quickly, it will feel as hard as a brick, and

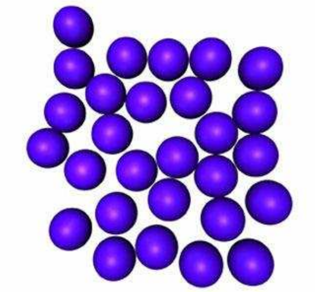


Figure 4.10: In Liquids, molecules are not in an ordered configuration and are free to move around (i.e, the liquid can flow). However, the liquid molecules do attract each other and at any moment in time you can typically define neighbors.

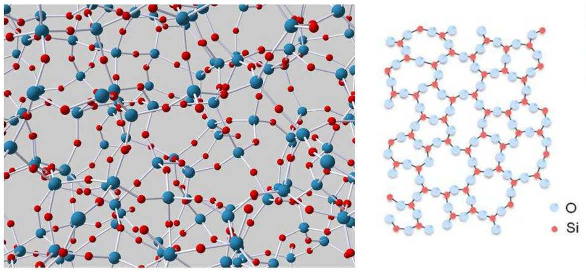


Figure 4.11: Silica can be an amorphous solid, or a glass. Left is a three dimensional picture, and right is a 2d cartoon. Here the atoms are disordered, but are bonded together and cannot flow.

And there are many more types of condensed matter systems that we simply do not have time to discuss. One can even engineer artificial types of order which do not occur naturally. Each one of these types of matter has its own interesting properties and if we had more time we would discuss them all in depth!

References

• Dove, chapter 2 gives discussion of many types of matter.

For an even more complete survey of the types of condensed matter see "Principles of

---

it will even crack (then it turns back to goo). In fact, you can make a deep tub of this stuff and although it feels completely like a fluid, you can run across the top of it (If you are too lazy to try doing this try Googling "Ellen cornstarch" to see a youtube video of the experiment). This mixture is a "non-Newtonian" fluid — its effective viscosity depends on how fast the force is applied to the material. The reason that polymers have this property is that the long polymer strands get tangled with each other. If a force is applied slowly the strands can unentangle and flow past each other. But if the force is applied quickly they cannot unentangle fast enough and the material acts just like a solid.

Figure 4.12: A liquid crystal has some of the properties of a solid and some of the properties of a liquid. In this picture of a smectic-C liquid crystal the system is crystalline in the vertical direction (forming discrete layers) but remains liquid (random positions) within each plane. Like a crystal, in this case, the individual molecules all have the same orientation.

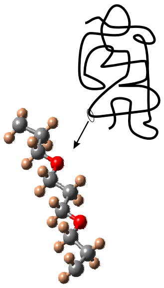


Figure 4.13: A polymer is a long chain of atoms.

Condensed Matter Physics", by Chaikin and Lubensky (Cambridge).

Part III

Toy Models of Solids in One Dimension: Mainly Vibrations

Chapter 5

The one-dimensional monatomic solid

In the first few chapters we found that our simple models of solids, and electrons in solids, were insufficient in several ways. In order to improve our understanding, we now need to take the periodic microstructure of crystals more seriously. To get a qualitative understanding of the effects of the periodic lattice, it is frequently sufficient to think in terms of simple one dimensional systems.

5.1 Forces between atoms: Compressibility, Sound, and Thermal Expansion

In the last chapter we discussed bonding between atoms. We found, particularly in the discussion of covalent bonding, that the lowest energy configuration would have the atoms at some optimal distance between (See figure 4.7, for example). Given this shape of the energy as a function of distance between atoms we will be able to come to some interesting conclusions.

For simplicity, let us imagine a 1-dimensional system of atoms (atoms in a single line). The potential $V(x)$ between the atoms is drawn in the Figure 5.1.

The classical equilibrium position is the position at the bottom of the well (marked x_{eq} in the figure). The distance between atoms at low temperature should then be x_{eq}. (As a homework assignment we will consider how quantum mechanics can change this value and increase it a little bit!).

Now, let us Taylor expand the potential around its minimum.

$$V(x) \approx V(x_{eq}) + \frac{\kappa}{2}(x - x_{eq})^2 + \frac{\kappa_3}{3!}(x - x_{eq})^3 + \ldots$$

Note that there is no linear term (if there were a linear term, then the position x_{eq} would not be the minimum). If there are only small deviations from the position x_{eq} the higher terms are much much smaller than the leading quadratic term and we can throw these terms out. This is a rather crucial general principle that *any* potential, close enough to its minimum, is quadratic.

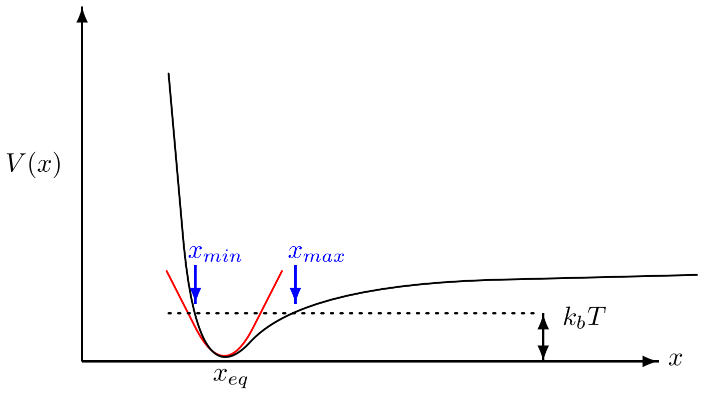


Figure 5.1: Potential Between Neighboring Atoms (black). The red curve is a quadratic approximation to the minimum (it may look crooked but in fact the red curve is symmetric and the black curve is asymmetric). The equilibrium position is x_{eq}. At finite temperature T, the system can oscillate between x_{max} and x_{min} which are not symmetric around the minimum. Thus as T increases the *average* position moves out to larger distance and the system expands.

Compressibility (or Elasticity)

We thus have a simple Hooke's law quadratic potential around the minimum. If we apply a force to compress the system (i.e., apply a pressure to our model one dimensional solid) we find

$$-\kappa(\delta x_{eq}) = F$$

where the sign is so that a positive (compressive) pressure reduces the distance between atoms. This is obviously just a description of the compressibility (or elasticity) of a solid. The usual description of compressibility is

$$\beta = -\frac{1}{V}\frac{\partial V}{\partial P}$$

(one should ideally specify if this is measured at fixed T or at fixed S. Here, we are working at $T = S = 0$ for simplicity). In the 1d case, we write the compressibility as

$$\beta = -\frac{1}{L}\frac{\partial L}{\partial F} = \frac{1}{\kappa\, x_{eq}} = \frac{1}{\kappa\, a} \tag{5.1}$$

with L the length of the system and x_{eq} is the spacing between atoms. Here we make the conventional definition that the equilibrium distance between identical atoms in a system (the so-called lattice constant) is written as a.

Sound

You may recall from your fluids course that in an isotropic compressible fluid, one predicts sound waves with velocity

$$v = \sqrt{\frac{B}{\rho}} = \sqrt{\frac{1}{\rho\beta}} \tag{5.2}$$

Here ρ is the mass density of the fluid, B is the bulk modulus, which is $B = 1/\beta$ with β the (adiabatic) compressibility.

While in a real solid the compressibility is anisotropic and the speed of sound depends in detail on the direction of propagation, in our model 1d solid this is not a problem and we can calculate that the the density is m/a with m the mass of each particle and a the equilibrium spacing between particles.

Thus using our result from above, we predict a sound wave with velocity

$$v = \sqrt{\frac{\kappa a^2}{m}} \tag{5.3}$$

Shortly (in section 5.2.2) we will re-derive this expression from a microscopic equations of motion for the atoms in the 1d solid.

Thermal Expansion

So far we have been working at zero temperature, but it is worth thinking at least a little bit about thermal expansion. This will be fleshed out more completely in a homework assignment. (In fact even in the homework assignment the treatment of thermal expansion will be very crude, but that should still be enough to give us the general idea of the phenomenon[1]).

Let us consider again figure 5.1 but now at finite temperature. We can imagine the potential as a function of distance between atoms as being like a ball rolling around in a potential. At zero energy, the ball sits at the the minimum of the distribution. But if we give the ball some finite temperature (i.e, some energy) it will oscillate around the minimum. At fixed energy k_bT the ball rolls back and forth between the points x_{min} and x_{max} where $V(x_{min}) = V(x_{max}) = k_bT$. But away from the minimum the potential is asymmetric, so $|x_{max} - x_{eq}| > |x_{min} - x_{eq}|$ so on average the particle has a position $\langle x \rangle > x_{eq}(T = 0)$. This is in essence the reason for thermal expansion! We will obtain positive thermal expansion for any system where $\kappa_3 < 0$ (i.e., at small x the potential is steeper) which almost always is true for real solids.

Summary

- Forces between atoms determine ground state structure.
- These same forces, perturbing around the ground state, determine elasticity, sound velocity, and thermal expansion.
- Thermal expansion comes from the non-quadratic part of the interatomic potential.

5.2 Micoscopic Model of Vibrations in 1d

In chapter 2 we considered the Boltzmann, Einstein, and Debye models of vibrations in solids. In this section we will consider a detailed model of vibration in a solid, first classically, and then quantum mechanically. We will be able to better understand what these early attempts to understand vibrations achieved and we will be able to better understand their shortcomings.

---

[1] Although this description is an annoyingly crude discussion of thermal expansion, we are mandated by the IOP to teach something on this subject. Explaining it more correctly, is unfortunately, rather messy!

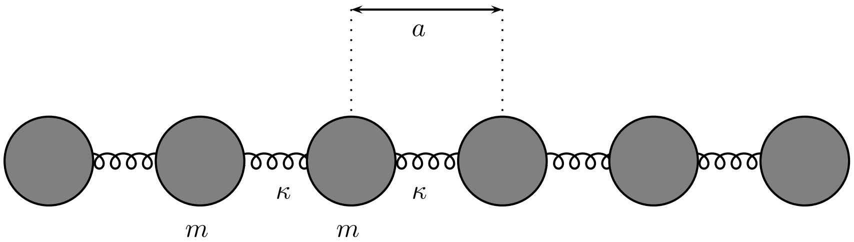
Let us consider a chain of identical atoms of mass m where the equilibrium spacing between atoms is a. Let us define the position of the n^{th} atom to be x_n and the equilibrium position of the n^{th} atom to be $x_n^{eq} = na$.

Once we allow motion of the atoms, we will have x_n deviating from its equilibrium position, so we define the small variable

$$\delta x_n = x_n - x_n^{eq}$$

Note that in our simple model we are allowing motion of the masses only in one dimension (i.e., we are allowing longitudinal motion of the chain, not transverse motion).

As discussed in the previous section, if the system is at low enough temperature we can consider the potential holding the atoms together to be quadratic. Thus, our model of a solid is chain of masses held together with springs as show in this figure



With this quadratic interatomic potential, we can write the total potential energy of the chain to be

$$\begin{aligned} V_{tot} &= \sum_i V(x_i - x_{i+1}) \\ &= V_{eq} + \sum_i \frac{\kappa}{2}(\delta x_i - \delta x_{i+1})^2 \end{aligned}$$

The force on the n^{th} mass on the chain is then given by

$$F_n = \frac{\partial V_{tot}}{\partial x_n} = \kappa(\delta x_{n+1} - \delta x_n) + \kappa(\delta x_{n-1} - \delta x_n)$$

Thus we have Newton's equation of motion

$$m(\ddot{\delta x_n}) = F_n = \kappa(\delta x_{n+1} + \delta x_{n-1} - 2\delta x_n) \tag{5.4}$$

To remind the reader, for any coupled system system, a *normal mode* is defined to be a collective oscillation where all particles move at the same frequency. We now attempt a solution to Newton's equations by using an ansatz that describes the normal modes as waves

$$\delta x_n = A e^{i\omega t - ikx_n^{eq}} = A e^{i\omega t - ikna}$$

where A is an amplitude of oscillation.

Now the reader might be confused about how it is that we are considering complex values of δx_n. Here we are using complex numbers for convenience but actually we implicitly mean to take the real part. (This is analogous to what one does in circuit theory with oscillating currents!). Since we are taking the real part, it is sufficient to consider only $\omega \geqslant 0$, however, we must be careful that k can then have either sign, and these are inequivalent once we have specified that ω is positive.

Plugging our ansatz into Eq. 5.4 we obtain

$$-m\omega^2 A e^{i\omega t - ikna} = \kappa A e^{i\omega t}\left[e^{-ika(n+1)} + e^{-ika(n-1)} - 2e^{-ikan}\right]$$

or

$$m\omega^2 = 2\kappa[1-\cos(ka)] = 4\kappa\sin^2(ka/2) \tag{5.5}$$

We thus obtain the result

$$\omega = 2\sqrt{\frac{\kappa}{m}}\left|\sin\left(\frac{ka}{2}\right)\right| \tag{5.6}$$

In general a relationship between a frequency (or energy) and a wavevector (or momentum) is known as a *dispersion relation*. This particular dispersion relation is shown in Fig. 5.2

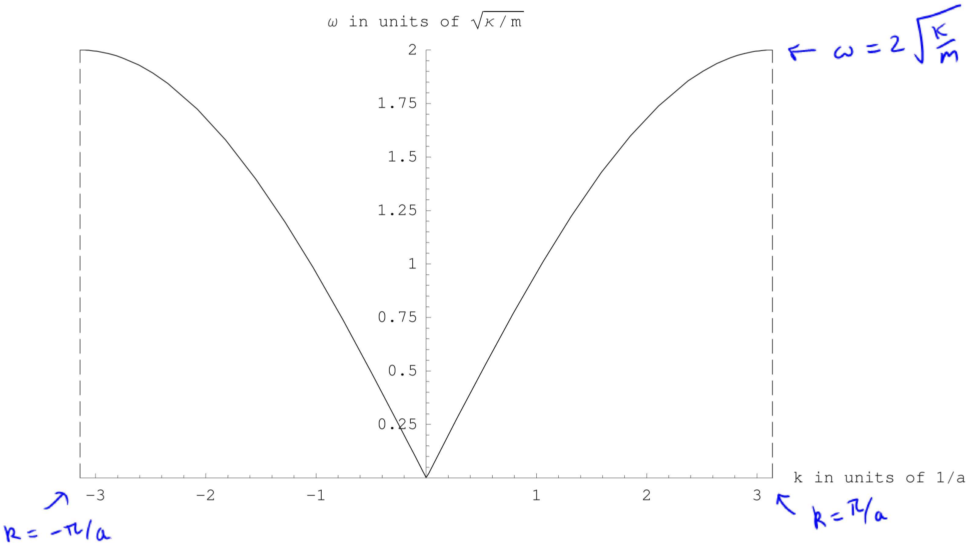


Figure 5.2: Dispersion relation of the 1d harmonic chain. The dispersion is periodic in $k \to k+2\pi/a$

5.2.1 First Exposure to the Reciprocal Lattice

Note that in the figure Fig. 5.2 we have only plotted the dispersion for $-\pi/a \leqslant k \leqslant \pi/a$. The reason for this is obvious from Eq. 5.6 — the dispersion relation is actually periodic in $k \to k+2\pi/a$. In fact this is a very important general principle:

> **Principle 5.2.1:** A system which is periodic in real space with a periodicity a will be periodic in reciprocal space with periodicity $2\pi/a$.

In this principle we have used the word *reciprocal* space which means k-space. In other words this principle tells us that if a system looks the same when $x \to x+a$ then in k-space the dispersion will look the same when $k \to k + 2\pi/a$. We will return to this principle many times in later chapters.

The periodic unit (the "unit cell") in k-space is conventionally known as the *Brillouin Zone*[2,3]. This is your first exposure to the concept of a Brillouin zone, but it again will play a very central role in later chapters. The "First Brillouin Zone" is a unit cell in k-space centered around the point $k = 0$. Thus in Fig. 5.2 we have shown only the first Brillouin zone, with the understanding that the dispersion is periodic for higher k. The points $k = \pm\pi/a$ are known as the *Brillouin-Zone boundary* and are defined in this case as being points which are symmetric around $k = 0$ and are separated by $2\pi/a$.

It is worth pausing for a second and asking why we expect that the dispersion curve should be periodic in $k \to k + 2\pi/a$. Recall that we defined our vibration mode to be of the form

$$\delta x_n = A e^{i\omega t - ikna} \tag{5.7}$$

If we take $k \to k + 2\pi/a$ we obtain

$$\delta x_n = A e^{i\omega t - i(k+2\pi/a)na} = A e^{i\omega t - ikna} e^{-i2\pi n} = A e^{i\omega t - ikna}$$

where here we have used

$$e^{-i2\pi n} = 1$$

for any integer n. What we have found here is that shifting $k \to k + 2\pi/a$ gives us back exactly the same oscillation mode the we had before we shifted k. The two are physically exactly equivalent!

In fact, it is similarly clear that shifting k by any $k + 2\pi p/a$ with p an integer since, will give us back exactly the same wave also since

$$e^{-i2\pi np} = 1$$

as well. We can thus define a set of points in k-space (reciprocal space) which are all physically equivalent to the point $k = 0$. This set of points is known as the *reciprocal lattice*. The original periodic set of points $x_n = na$ is known as the *direct lattice* or *real-space lattice* to distinguish it from the reciprocal lattice, when necessary.

The concept of the reciprocal lattice will extremely important later on. We can see the analogy between the direct lattice and the reciprocal lattice as follows:

Direct lattice: $x_n = \ldots \quad -2a, \quad -a, \quad 0, \quad a, \quad 2a, \quad \ldots$
Reciprocal lattice: $G_n = \ldots \quad -2\left(\frac{2\pi}{a}\right), \quad -\frac{2\pi}{a}, \quad 0, \quad \frac{2\pi}{a}, \quad 2\left(\frac{2\pi}{a}\right), \quad \ldots$

Note that the defining property of the reciprocal lattice in terms of the points in the real lattice can be given as

$$e^{iG_m x_n} = 1 \tag{5.8}$$

A point G_m is a member of the reciprocal lattice iff Eq. 5.8 is true for all x_n in the real lattice.

---

[2] Leon Brillouin was one of Sommerfeld's students. He is famous for many things including for being the "B" in the "WKB" approximation. I'm not sure if WKB is on your syllabus, but it really should be if it is not already!

[3] The pronunciation of "Brillouin" is something that gives English speakers a great deal of difficulty. If you speak French you will probably cringe at the way this name is butchered. (I did badly in French in school, so I'm probably one of the worst offenders.) According to online dictionaries it is properly pronounced somewhere between the following words: brēwan, bree-wah, breel-wahn, bree(y)wa(n), and bree-l-(uh)-wahn. At any rate, the "l" and the "n" should both be very weak.

5.2.2 Properties of the Dispersion of the 1d chain

We now return to more carefully examine the properties of the dispersion we calculated (Eq. 5.6).

Sound Waves:

Recall that sound wave[4] is a vibration that has a long wavelength (compared to the inter-atomic spacing). In this long wavelength regime, we find the dispersion we just calculated to be linear with wavevector $\omega = v_{sound} k$ as expected for sound with

$$v_{sound} = a\sqrt{\frac{\kappa}{m}}.$$

(To see this, just expand the sin in Eq. 5.6). Note that this sound velocity matches the velocity predicted from Eq. 5.3.

However, we note that at larger k, the dispersion is no longer linear. This is in disagreement with what Debye assumed in his calculation in section 2.2. So clearly this is a shortcoming of the Debye theory. In reality the dispersion of normal modes of vibration is linear only at long wavelength.

At shorter wavelength (larger k) one typically defines two different velocities: The *group velocity*, the speed at which a wavepacket moves, is given by

$$v_{group} = d\omega/dk$$

And the *phase velocity*, the speed at which the individual maxima and minima move, is given by

$$v_{phase} = \omega/k$$

These two match in the case of a linear dispersion, but otherwise are different. Note that the group velocity becomes zero at the Brillouin zone boundaries $k = \pm\pi/a$ (i.e., the dispersion is flat). As we will see many times later on, this is a general principle!

Counting Normal Modes:

Let us now ask how many normal modes there are in our system. Naively it would appear that we can put any k such that $-\pi/a \leqslant k < \pi/a$ into Eq. 5.6 and obtain a new normal mode with wavevector k and frequency $\omega(k)$. However this is not precisely correct.

Let us assume our system has exactly N masses in a row, and for simplicity let us assume that our system has periodic boundary conditions i.e., particle x_0 has particle x_1 to its right and particle x_{N-1} to its left. Another way to say this is to let, $x_{n+N} = x_n$, i.e., this one dimensional system forms a big circle. In this case we must be careful that the wave ansatz Eq. 5.7 makes sense as we go all the way around the circle. We must therefore have

$$e^{i\omega t - ikna} = e^{i\omega t - ik(N+n)a}$$

Or equivalently we must have

$$e^{ikNa} = 1$$

---

[4] For reference it is good to remember that humans can hear sound wavelengths roughly between 1cm and 10m. Both of these are very long wavelength compared to interatomic spacings.

This requirement restricts the possible values of k to be of the form

$$k = \frac{2\pi p}{Na} = \frac{2\pi p}{L}$$

where p is an integer and L is the total length of the system. Thus k becomes quantized rather than a continuous variable. This means that the k-axis in Figure 5.2 is actually a discrete set of many many individual points; the spacing between two of these consecutive points being $2\pi/(Na)$.

Let us now count how many normal modes we have. As mentioned above in our discussion of the Brillouin zone, adding $2\pi/a$ to k brings one back to exactly the same physical wave. Thus we only ever need consider k values within the first Brillouin zone (i.e., $-\pi/a \leqslant k < pi/a$ (since π/a is the same as $-\pi/a$ we choose to count one but not the other). Thus the total number of normal modes is

$$\text{Total Number of Modes} = \frac{\text{Range of } k}{\text{Spacing between neigboring } k} = \frac{2\pi/a}{2\pi/(Na)} = N \qquad (5.9)$$

There is precisely one normal mode per mass in the system — that is, one normal mode per degree of freedom in the whole system. This is what Debye insightfully predicted in order to cut off his divergent integrals in section 2.2.3 above!

5.2.3 Quantum Modes: Phonons

We now make a rather important leap from classical to quantum physics.

> **Quantum Correspondence:** If a classical harmonic system (i.e., any quadratic hamiltonian) has a normal oscillation mode at frequency ω the corresponding quantum system will have eigenstates with energy
>
> $$E_n = \hbar\omega(n + \frac{1}{2}) \qquad (5.10)$$

Presumably you know this well in the case of a single harmonic oscillator. The only thing different here is that the harmonic oscillator is a collective normal mode not just motion of a single particle. This quantum correspondence principle will be the subject of a homework assignment.

Thus at a given wavevector k, there are many possible eigenstates, the ground state being the $n = 0$ eigenstate which has only the zero-point energy $\hbar\omega(k)/2$. The lowest energy excitation is of energy $\hbar\omega(k)$ greater than the ground state corresponding to the excited $n = 1$ eigenstate. Generally all excitations at this wavevector occur in energy units of $\hbar\omega(k)$, and the higher values of energy correspond classically to oscillations of increasing amplitude.

Each excitation of this "normal mode" by step up the harmonic oscillator excitation ladder (increasing the quantum number n) is known as a "phonon".

Definition 5.2.1. A *phonon* is a discrete quanta of vibration[5]

This is entirely analogous to defining a single quanta of light as a photon. As is the case with the photon, we may think of the phonon as actually being a particle, or we can think of the phonon as being a quantized wave.

---

[5] I do not like the definition of a phonon as "a quanta of vibrational energy" which many books use. The vibration does carry indeed energy, but it carries other quantum numbers (such as crystal momentum) as well, so why specify energy only?

If we think about the phonon as being a particle (as with the photon) then we see that we can put many phonons in the same state (ie., the quantum number n in Eq. 5.10 can be increased to any value), thus we conclude that phonons, like photons, are bosons. As with photons, at finite temperature there will be a nonzero number of phonons (i.e., n will be on average nonzero) as given by the Bose occupation factor.

$$n_B(\beta\omega) = \frac{1}{e^{\beta\omega} - 1}$$

with $\beta = 1/(k_b T)$ and ω the oscillation frequency.

Thus, the expectation of the phonons at wavevector k is given by

$$E_k = \hbar\omega(k)\left(n_B(\beta\omega(k)) + \frac{1}{2}\right)$$

We can use this type of expression to calculate the heat capacity of our 1d model[6]

$$U_{total} = \sum_k \hbar\omega(k)\left(n_B(\beta\omega(k)) + \frac{1}{2}\right)$$

where the sum over k here is over all possible normal modes, i.e, $k = 2\pi p/(Na)$ such that $-\pi/a \leqslant k < \pi/a$. Thus we really mean

$$\sum_k \rightarrow \sum_{p=-N/2}^{N/2-1} |_{k=(2\pi p)/(Na)}$$

Since for a large system, the k points are very close together, we can convert the discrete sum into an integral (something we should be very familiar with by now) to obtain

$$\sum_k \rightarrow \frac{Na}{2\pi}\int_{-\pi/a}^{\pi/a} dk$$

Note that we can use this continuum integral to count the total number of modes in the system

$$\frac{Na}{2\pi}\int_{-\pi/a}^{\pi/a} dk = N$$

as predicted by Debye.

Using this integral form of the sum we have the total energy given by

$$U_{total} = \frac{N}{2\pi}\int_{-\pi/a}^{\pi/a} dk\, \hbar\omega(k)\left(n_B(\beta\omega(k)) + \frac{1}{2}\right)$$

from this we could calculate specific heat as dU/dT.

These two previous expressions look exactly like what Debye would have obtained from his calculation (for a 1d version of his model)! The only difference lies in our expression for $\omega(k)$. Debye only knew about sound where $\omega = vk$, is linear in wavevector. We, on the other hand, have

---

[6] The observant reader will note that we are calculating $C_V = dU/dT$ the heat capacity at constant volume. Why constant volume? As we saw above when we studied thermal expansion, the crystal does not expand unless we include third order (or higher) order terms in the interatomic potential, which are not in this model!

just calculated that for our microscopic ball and spring model ω is not linear in k. In fact, Einstein's calculation of specific heat can be phrased in exactly the same language. Only for Einstein's model $\omega = \omega_0$ is constant for all k. We thus see Einstein's model, Debye's model, and our microscopic harmonic model in a very unified light. The only difference between the three is what we use for a dispersion relation.

One final comment is that it is frequently useful to further replace integrals over k with integrals over frequency (we did this when we studied the Debye model above). We obtain generally

$$\frac{Na}{2\pi}\int_{-\pi/a}^{\pi/a} dk = \int d\omega g(\omega)$$

where[7]

$$g(\omega) = 2\frac{Na}{2\pi}|dk/d\omega|$$

Note that in the (1d) Debye model this density of states is constant from $\omega = 0$ to $\omega = \omega_{Debye} = v\pi/a$. In our model, as we have calculated above, the density of states is not a constant, but becomes zero at frequency above the maximum frequency $2\sqrt{\kappa/m}$. (In a homework problem we calculate this density of states explicitly). Finally in the Einstein model, this density of states is a delta-function at frequency ω_0.

5.2.4 Comments on Reciprocal Space II: Crystal Momentum

As mentioned above, the wavevector of a phonon is defined only modulo the reciprocal lattice. In other words, k is the same as $k + G_m$ where $G_m = 2\pi m/a$ is a point in the reciprocal lattice. Now we are supposed to think of these phonons as particles – and we like to think of our particles as having energy $\hbar\omega$ and a momentum $\hbar k$. But we cannot define a phonon's momentum this way because physically it is the same phonon whether we describe it as $\hbar k$ or $\hbar(k + G_m)$. We thus instead define a concept known as the *crystal momentum* which is the momentum modulo the reciprocal lattice — or equivalently we agree that we must always describe k within the first Brillouin zone.

In fact, this idea of crystal momentum is extremely powerful. Since we are thinking about phonons as being particles, it is actually possible for two (or more) phonons to bump into each other and scatter from each other — the same way particles do[8]. In such a collision, energy is conserved and *crystal momentum is conserved*! For example three phonons each with crystal momentum $\hbar(2/3)\pi/a$ can scatter off of each other to produce three phonons each with crystal momentum $-\hbar(2/3)\pi/a$. This is allowed since the initial and final states have the same energy and

$$3 \times (2/3)\pi/a = 3 \times (-2/3)\pi/a \mod (2\pi/a)$$

During these collisions although momentum $\hbar k$ is not conserved, crystal momentum is[9]. In fact, the situation is similar when, for example, phonons scatter from electrons in a periodic lattice –

---

[7] The factor of 2 out front comes from the fact that each ω occurs for the two possible values of $\pm k$.

[8] In the harmonic model we have considered phonons do not scatter from each other. We know this because the phonons are eigenstates of the system, so their occupation does not change with time. However, if we add anharmonic (cubic and higher) terms to the inter-atomic potential, this corresponds to perturbing the phonon hamiltonian and can be interpreted as allowing phonons to scatter from each other.

[9] This thing we have defined $\hbar k$ has dimensions of momentum, but is not conserved. However, as we will discuss below in chapter 10, if a particle, like a photon, enters a crystal with a given momentum and undergoes a process that conserves crystal momentum but not momentum, when the photon exits the crystal we will find that total momentum of the system is indeed conserved, with the momentum of the entire crystal accounting for any momentum that is missing from the photon. See footnote 5 in section 10.1.1

crystal momentum becomes the conserved quantity rather than momentum. This is an extremely important principle which we will encounter again and again. In fact, it is a main cornerstone of solid-state physics.

Aside: There is a very fundamental reason for the conservation of crystal momentum. Conserved quantities are results of symmetries (this is a deep and general statement known as Noether's theorem[10]). For example, conservation of momentum is a result of the translational invariance of space. If space is not the same from point to point, for example if there is a potential $V(x)$ which is different at different places, then momentum is not conserved. The conservation of crystal momentum correspondingly results from space being invariant under translations of a, giving us momentum that is conserved modulo $2\pi/a$.

5.3 Summary of the one-dimensional chain

A number of very crucial new ideas have been introduced in this section. Many of these will return again and again in later chapters.

- Normal modes are collective oscillations where all particles move at the same frequency.
- If a system is periodic in space with periodicity $\Delta x = a$, then in reciprocal space (k-space) the system is periodic with periodicity $\Delta k = 2\pi/a$.
- Values of k which differ by multiples of $2\pi/a$ are physically equivalent. The set of points in k-space which are equivalent to $k = 0$ are known as the reciprocal lattice.
- Any values of k is equivalent to some k in the first Brillouin-zone, $-\pi/a \leqslant k < \pi/a$ (in 1d).
- The sound velocity is the slope of the dispersion in the small k limit (group = phase velocity in this limit).
- A classical normal mode of frequency ω gets translated into quantum mechanical eigenstates $E_n = \hbar\omega(n + \frac{1}{2})$. If the system is in the n^{th} eigenstate, we say that it is occupied by n phonons.
- Phonons can be thought of as particles, like photons, that obey Bose statistics.

References

Sound and compressibility:

- Goodstein, section 3.2b.
- Ibach an Luth, beginning of section 4.5
- Hook and Hall, section 2.2

Thermal Expansion (Most references go into *way* too much depth on thermal expansion):

- Kittel chapter 5, section on thermal expansion.

---

[10] Emmy Noether has been described by Einstein, among others, as the most important woman in the history of mathematics.

Normal Modes of Monatomic Chain and introduction to phonons:

- Kittel, beginning of chapter 4
- Goodstein, beginning of section 3.3
- Hook and Hall, section 2.3.1
- Burns, section 12.1-12.2
- Ashcroft and Mermin, beginning of chapter 22.

Chapter 6

The one-dimensional diatomic solid

In the previous chapter we studied in detail a one dimensional model of a solid where every atom is identical to every other atom. However, in real materials not every atom is the same (for example, in Sodium Chloride, we have two types of atoms!). We thus intend to generalize our previous discussion of the one dimension solid to a one dimensional solid with two types of atoms. Much of this will follow the outline set in the previous chapter, but we will see that several fundamentally new features will now emerge.

6.1 Diatomic Crystal Structure: Some useful definitions

Consider the following model system

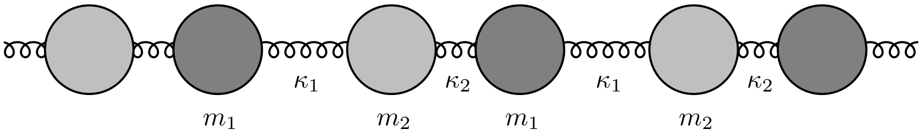


Fig. 6.1.1

Which represents a periodic arrangement of two different types of atoms. Here we have given them two masses m_1 and m_2 and the springs which alternate along the 1-dimensional chain. The springs connecting the atoms have spring constants κ_1 and κ_2 and also alternate.

In this circumstance with more than one type of atom, we first would like to identify the so-called *unit cell* which is the repeated motif in the arrangement of atoms. In this picture, we have put a box around the unit cell. The length of the unit cell in 1d is known as the *lattice constant* and it is labeled a.

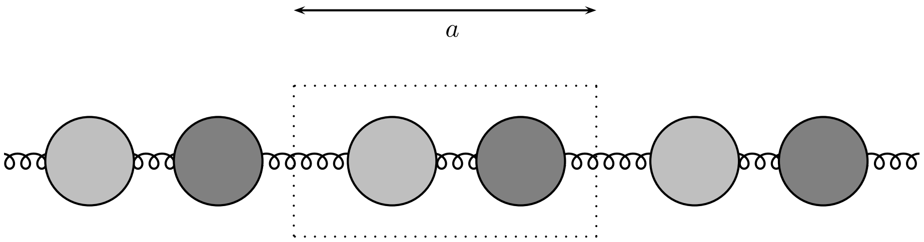


Fig. 6.1.2

Note however, that the definition of the unit cell is extremely non-unique. We could just as well have chosen (for example) the unit cell to be as follows.

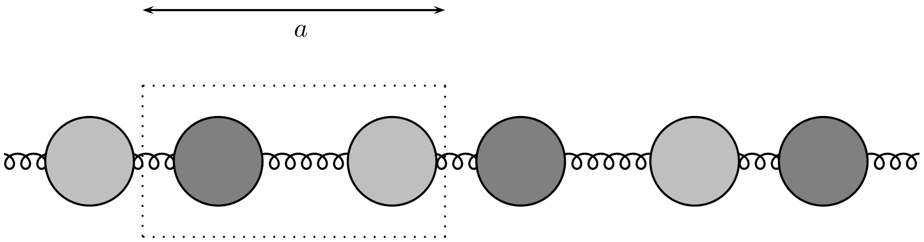


Fig. 6.1.3

The important thing in defining a periodic system is to choose *some* unit cell and then construct the full system by reproducing the same unit cell over and over. (In other words, make a definition of the unit cell and stick with that definition!).

It is sometimes useful to pick some references point inside each unit cell. This set of reference points makes a simple *lattice* (we will define the term "lattice" more closely in later chapters – but for now the point is that a lattice has only one type of point in it – not two different types of points). So in this figure, we have marked our reference point in each unit cell with an `X` (again, the choice of this reference point is arbitrary).

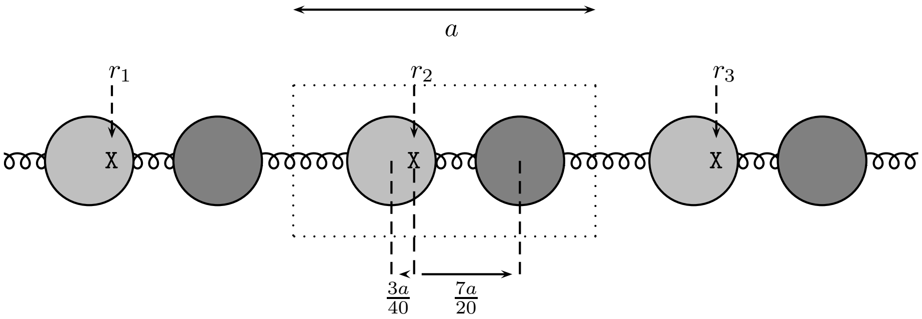


Fig. 6.1.4

Given the reference lattice point in the unit cell, the description of all of the atoms in the unit cell with respect to this reference point is known as a *basis*. In this case we might describe our basis as

light gray atom	centered at position	$3a/40$	to left of reference lattice point
dark gray atom	centered at position	$7a/20$	to right of reference lattice point

Thus if the reference lattice point in unit cell n is called r_n (and the spacing between the lattice points is a) we can set

$$r_n = an$$

with a the size of the unit cell. Then the (equilibrium) position of the light gray atom in the n^{th} unit cell is

$$x_n^{eq} = an - 3a/40$$

whereas the (equilibrium) position of the dark gray atom in the n^{th} unit cell is

$$y_n^{eq} = an + 7a/20$$

6.2 Normal Modes of the Diatomic Solid

For simplicity, let us focus on the case where all of the masses along our chain are the same $m_1 = m_2 = m$ but the two spring constants κ_1 and κ_2 are different. (For homework we will consider the case where the masses are different, but the spring constants are the same!).

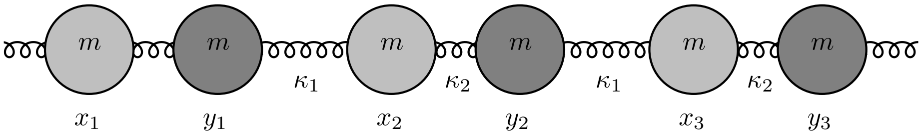


Fig. 6.2.1

Given the spring constants in the picture, we can write down Newton's equations of of motion for the deviations of the positions of the masses from their equilibrium positions. We obtain

$$\begin{aligned} m\,\delta\ddot{x}_n &= \kappa_2(\delta y_n - \delta x_n) + \kappa_1(\delta y_{n-1} - \delta x_n) && (6.1)\\ m\,\delta\ddot{y}_n &= \kappa_1(\delta x_{n+1} - \delta y_n) + \kappa_2(\delta x_n - \delta y_n) && (6.2) \end{aligned}$$

Analogous to the one dimensional case we propose ansätze[1] for these quantities that have the form of a wave

$$\begin{aligned} \delta x_n &= A_x e^{i\omega t - ikna} && (6.3)\\ \delta y_n &= A_y e^{i\omega t - ikna} && (6.4) \end{aligned}$$

where, as in the previous chapter, we implicitly mean to take the real part of the complex number. As such, we can always choose to take $\omega > 0$ as long as we consider k to be either positive and negative

As we saw in the previous chapter, values of k that differ by $2\pi/a$ are physically equivalent. We can thus focus our attention to the first Brillouin zone $-\pi/a \leqslant k < \pi/a$. Note that the important length here is the unit cell length or lattice constant a. Any k outside the first Brillouin zone is redundant with some other k inside the zone.

As we found in the previous chapter, if our system has N unit cells, then (putting periodic boundary conditions on the system) k will be is quantized in units of $2\pi/(Na) = 2\pi/L$. Note that here the important quantity is N, the number of unit cells, not the number of atoms ($2N$).

Dividing the range of k in the first Brillouin zone by the spacing between neighboring k's, we obtain exactly N different possible values of k exactly as we did in Eq. 5.9. In other words, we have exactly one value of k per unit cell.

We might recall at this point the intuition that Debye used — that there should be exactly one possible excitation mode per degree of freedom of the system. Here we obviously have two degrees of freedom per unit cell, but we obtain only one possible value of k per unit cell. The resolution, as we will see in a moment, is that there will be two possible oscillation modes for each wavevector k.

We now proceed by plugging in our ansätze (Eq. 6.3 and 6.4) into our equations of motion (Eq. 6.1 and 6.2). We obtain

$$\begin{aligned} -\omega^2 m A_x e^{i\omega t - ikna} &= \kappa_2 A_y e^{i\omega t - ikna} + \kappa_1 A_y e^{i\omega t - ik(n-1)a} - (\kappa_1 + \kappa_2) A_x e^{i\omega t - ikna}\\ -\omega^2 m A_y e^{i\omega t - ikna} &= \kappa_1 A_x e^{i\omega t - ik(n+1)a} + \kappa_2 A_x e^{i\omega t - ikna} - (\kappa_1 + \kappa_2) A_y e^{i\omega t - ikna} \end{aligned}$$

which simplifies to

$$\begin{aligned} -\omega^2 m A_x &= \kappa_2 A_y + \kappa_1 A_y e^{ika} - (\kappa_1 + \kappa_2) A_x\\ -\omega^2 m A_y &= \kappa_1 A_x e^{-ika} + \kappa_2 A_x - (\kappa_1 + \kappa_2) A_y \end{aligned}$$

This can be rewritten conveniently as an eigenvalue equation

$$m\omega^2 \begin{pmatrix} A_x \\ A_y \end{pmatrix} = \begin{pmatrix} (\kappa_1 + \kappa_2) & -\kappa_2 - \kappa_1 e^{ika} \\ -\kappa_2 - \kappa_1 e^{-ika} & (\kappa_1 + \kappa_2) \end{pmatrix} \begin{pmatrix} A_x \\ A_y \end{pmatrix} \tag{6.5}$$

The solutions of this are obtained by finding the zeros of the secular determinant

$$0 = \begin{vmatrix} (\kappa_1 + \kappa_2) - m\omega^2 & -\kappa_2 - \kappa_1 e^{ika} \\ -\kappa_2 - \kappa_1 e^{-ika} & (\kappa_1 + \kappa_2) - m\omega^2 \end{vmatrix} = \left|(\kappa_1 + \kappa_2) - m\omega^2\right|^2 - |\kappa_2 + \kappa_1 e^{ika}|^2$$

---

[1] I believe this is the proper pluralization of ansatz.

The roots of which are clearly given by

$$m\omega^2 = (\kappa_1 + \kappa_2) \pm |\kappa_1 + \kappa_2 e^{ika}|$$

The second term needs to be simplified

$$|\kappa_1 + \kappa_2 e^{ika}| = \sqrt{(\kappa_1 + \kappa_2 e^{ika})(\kappa_1 + \kappa_2 e^{-ika})} = \sqrt{\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2 \cos(ka)}$$

So we finally obtain

$$\omega_\pm = \sqrt{\frac{\kappa_1 + \kappa_2}{m} \pm \frac{1}{m}\sqrt{\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2\cos(ka)}} \tag{6.6}$$

Note in particular that for each k we find two normal modes — thus since there are N different k values, we obtain $2N$ modes total (if there are N unit cells in the entire system). This is in agreement with our above discussion that we should have exactly one normal mode per degree of freedom in our system.

The dispersion of these two modes is shown in Figure 6.1.

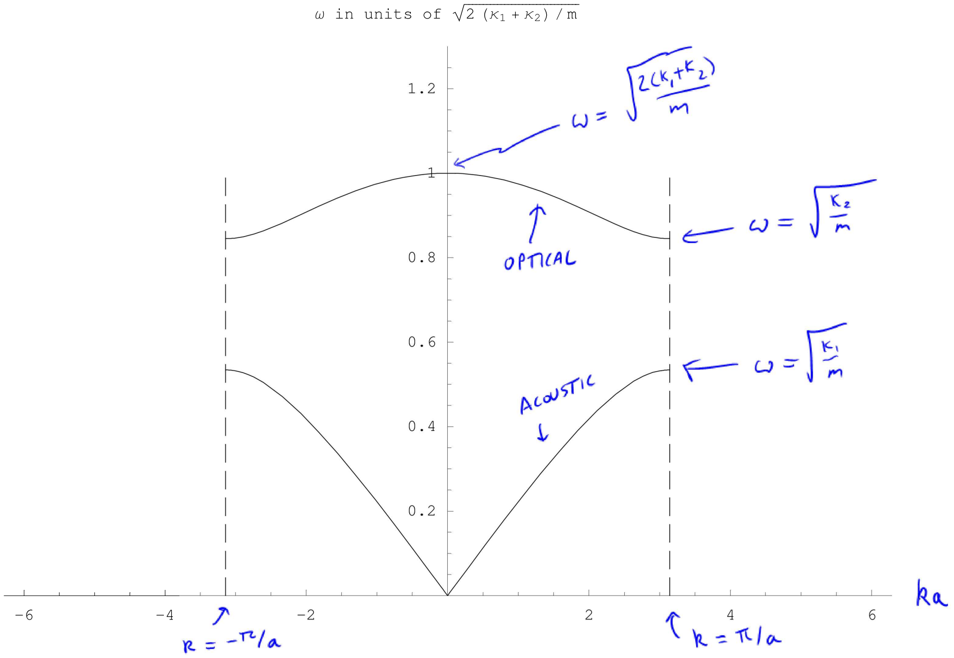


Figure 6.1: Dispersion relation of the 1d diatomic chain. The dispersion is periodic in $k \to k + 2\pi/a$. Here the dispersion is shown for the case of $\kappa_2 = 1.4\kappa_1$. This scheme of plotting dispersions, putting all normal modes within the first Brillouin zone, is the *reduced zone scheme*. Compare this to Fig. 6.2 below.

A few things to note about this dispersion. First of all we note that there is a long wavelength low energy mode with linear dispersion (corresponding to ω_- in Eq. 6.6). This is the sound wave, or acoustic mode. Generally the definition of an *acoustic mode* is any mode that has linear dispersion as $k \to 0$. These correspond to sound waves.

By expanding Eq. 6.6 for small k it is easy to check that the sound velocity is

$$v_{sound} = \frac{d\omega_-}{dk} = \sqrt{\frac{a^2\kappa_1\kappa_2}{2m(\kappa_1+\kappa_2)}} \tag{6.7}$$

In fact, we could have calculated this sound velocity on general principles analogous to what we did in Eq. 5.2. The density of the chain is $2m/a$. The effective spring constant of two springs κ_1 and κ_2 in series is $\tilde{\kappa} = (\kappa_1+\kappa_2)/(\kappa_1\kappa_2)$ so the compressibility of the chain is $\beta = 1/(\tilde{\kappa}a)$ (See Eq. 5.1). Then plugging into Eq. 5.2 gives exactly the same sound velocity as we calculate here in Eq. 6.7.

The higher energy branch of excitations is known as the *optical* mode. It is easy to check that in this case the optical mode goes to frequency $\sqrt{2(\kappa_1+\kappa_2)/m}$ at $k=0$, and also has zero group velocity at $k=0$. The reason for the nomenclature "optical" will become clearer later in the course when we study scattering of light from solids. However, here is a very simplified description of why it is named this way: Consider a solid being exposed to light. It is possible for the light to be absorbed by the solid, but energy and momentum must both be conserved. However, light travels at a very high velocity c, so $\omega = ck$ is a very large number. Since phonons have a maximum frequency, this means that photons can only be absorbed for very small k. However, for small k, acoustic phonons have energy $vk \ll ck$ so that energy and momentum cannot be conserved. On the other hand, optical phonons have energy $\omega_{optical}$ which is finite for small k so that at some value of small k, we have $\omega_{optical} = ck$ and one can match the energy and momentum of the photon to that of the phonon.[2] Thus, whenever phonons interact with light, it is inevitably the optical phonons that are involved.

Let us examine a bit more closely the acoustic and the optical mode as $k \to 0$. Examining our eigenvalue problem Eq. 6.5, we see that in this limit the matrix to be diagonalized takes the simple form

$$\omega^2 \begin{pmatrix} A_x \\ A_y \end{pmatrix} = \frac{\kappa_1+\kappa_2}{m} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} A_x \\ A_y \end{pmatrix} \tag{6.8}$$

The acoustic mode (which has frequency 0) is solved by the eigenvector

$$\begin{pmatrix} A_x \\ A_y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

This tells us that the two masses in the unit cell (at positions x and y) move together for the case of the acoustic mode in the long wavelength limit. This is not surprising considering our understanding of sound waves as being very long wavelength compressions and rarifactions. This is depicted in Figure 6.2.2. Note in the figure that the amplitude of the compression is slowly modulated, but always the two atoms in the unit cell move almost exactly the same way.

---

[2] From this naive argument, one might think that the process where one photon with frequency $\omega_{optical}$ is absorbed while emitting a phonon is an allowed process. This is not true since the photons carry spin and spin must also be conserved. Much more typically the interaction between photons and phonons is one where a photon is absorbed and then re-emitted. I.e., it is inelastically scattered. We will discuss this later on.

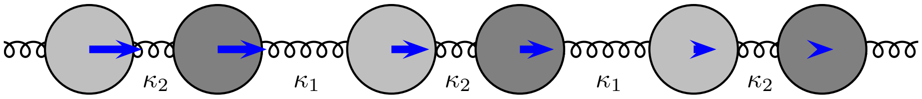


Fig. 6.2.2

A long wavelength acoustic mode

On the other hand, the optical mode at $k = 0$, having frequency $\omega^2 = \frac{2(\kappa_1+\kappa_2)}{m}$, has the eigenvector

$$\begin{pmatrix} A_x \\ A_y \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

which described the two masses in the unit cell moving in *opposite* directions, for the optical mode. This is depicted in Figure 6.2.3. Note in the figure that the amplitude of the compression is slowly modulated, but always the two atoms in the unit cell move almost exactly the opposite way.

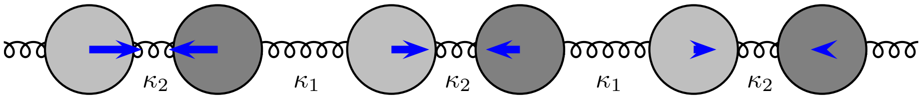


Fig. 6.2.3

A long wavelength optical mode

In order to get a better idea of how motion occurs for both the optical and acoustic modes, it is useful to see animations, which you can find on the web. Another good resource is to download the program "ChainPlot" from Professor Mike Glazer's web site (http://www.amg122.com/programs)[3]

In this example we had two atoms per unit cell and we obtained two modes per distinct value of k. One of these modes is acoustic and one is optical. More generally if there are M atoms per unit cell (in one dimension) we will have M modes per distinct value of k, of which one mode will be acoustic (goes to zero energy at $k = 0$) and all of the remaining modes are optical (do not go to zero energy at $k = 0$).

Caution: We have been careful to discuss a true one dimensional system, where the atoms are allowed to move only along the one dimensional line. Thus each atom has only one degree of freedom. However, if we allow atoms to move in other directions (transverse to the 1d line) we will have more degrees of freedom per atom. When we get to the 3d solid we should expect 3 degrees of freedom per atom. And there should be 3 different acoustic modes at each k at long wavelength. (In 3d, if there are n atoms per unit cell, there will be $3(n-1)$ optical modes but always 3 acoustic modes totalling $3n$ degrees of freedom per unit cell.

---

[3] Note in particular the comment on this website about most books getting the form of the acoustic mode incorrect!

One thing that we should study closely is the behavior at the Brillouin zone boundary. It is also easy to check that the frequencies $\omega_\pm$ at the zone boundary ($k = \pm\pi/a$) are $\sqrt{\kappa_1/m}$ and $\sqrt{\kappa_2/m}$, the larger of the two being ω_+. We can also check that the group velocity $d\omega/dk$ of both modes goes to zero at the zone boundary (Similarly the optical mode has zero group velocity at $k = 0$).

In Fig. 6.1 above, we have shown both modes at each value of k, such that we only need to show k within the first Brillouin zone. This is known as the *reduced zone scheme.* Another way to plot exactly the same dispersions is shown in Fig. 6.2 and is known as the *extended zone scheme.* Essentially you can think of this as "unfolding" the dispersions such that there is only one mode at each value of k. In this picture we have defined (for the first time) the *second Brillouin zone.*

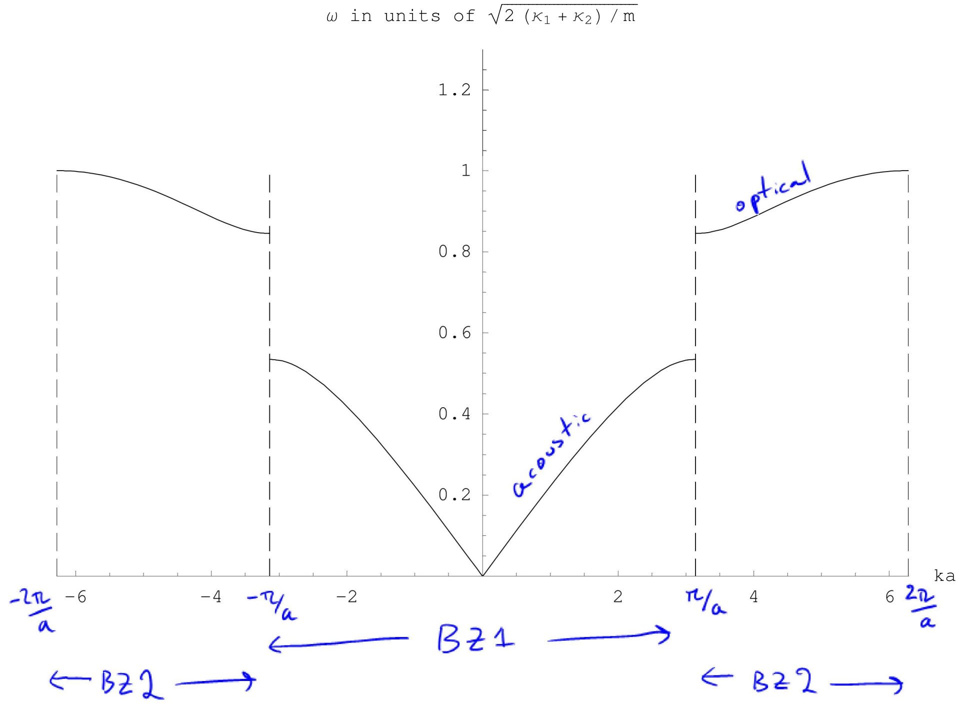


Figure 6.2: Dispersion relation of the 1d diatomic chain in the extended zone scheme (Again choosing $\kappa_2 = 1.4\kappa_1$). Compare this to Fig. 6.1 above. The first Brillouin zone is labeled BZ1 and the second Brillouin zone is labeled BZ2.

Recall the first zone in 1d is defined as $|k| \leqslant \pi/a$. Analogously the second Brillouin zone is now $\pi/a \leqslant |k| \leqslant 2\pi/a$. In later chapters we will define the Brillouin zones more generally.

Here is an example where it is very useful to think using the extended zone scheme. We have been considering with $\kappa_2 > \kappa_1$, now let us consider what would happen if we take the limit of $\kappa_2 \to \kappa_1$. When the two spring constants become the same, then in fact the two atoms in the unit cell become identical, and we have a simple monotomic chain (which we discussed at length in the previous chapter). As such we should define a new smaller unit cell with lattice constant $a/2$, and the dispersion curve is now just a simple $|\sin|$ as it was in chapter Ref. 5.2.

Thus it is frequently useful if the two atoms in a unit cell are not too different from each

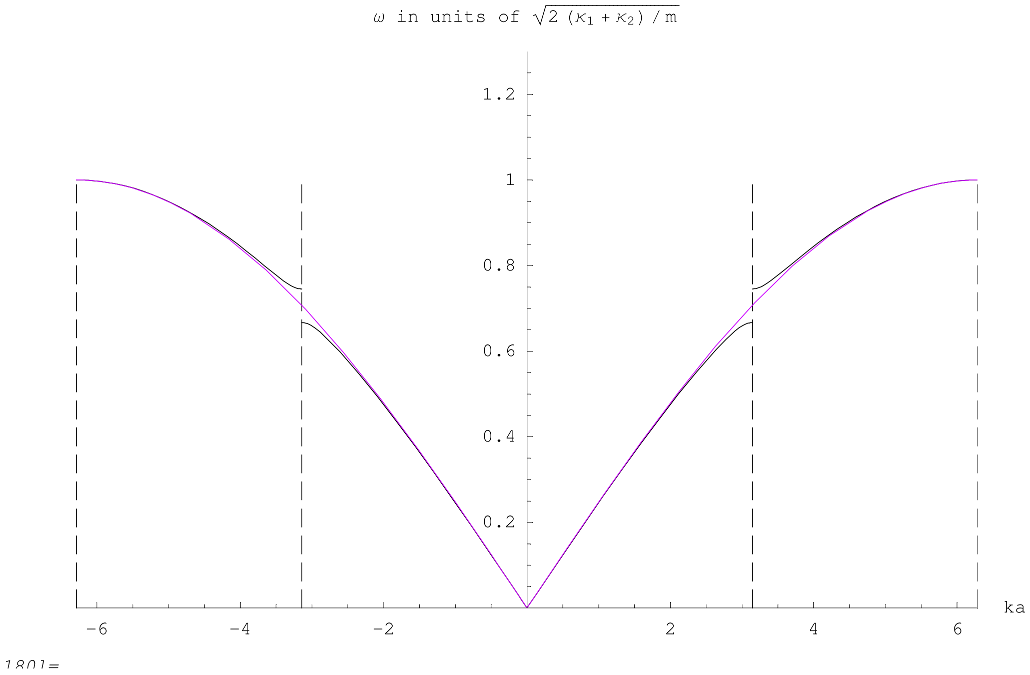


Figure 6.3: (black) Dispersion relation of the 1d diatomic chain in the extended zone scheme with κ_2 not too different from κ_1. (blue) Dispersion relation when $\kappa_2 = \kappa_1$. In this case, the two atoms become exactly the same, and we have a monatomic chain with lattice spacing $a/2$. This single band dispersion precisely matches that calculated in chapted 5.2 above, only with the lattice constant redefined to $a/2$.

other, to think about the dispersion as being a small perturbation to a situation where all atoms are identical. When the atoms are made slightly different, a small gap opens up at the zone boundary, but the rest of the dispersion continues to look mostly as if it is the dispersion of the monatomic chain.

6.3 Summary of the diatomic solid

A number of key concepts are introduced in this chapter as well

- A unit cell is the repeated motif that comprises a crystal
- The basis is the description of the unit cell with respect to a reference lattice
- The lattice constant is the size of the unit cell (in 1d)
- If there are M atoms per unit cell we will find M normal modes at each wavevector k
- One of these modes is an acoustic mode, meaning that it has linear dispersion at small k, whereas the remaining $M - 1$ are optical meaning they have finite frequency at $k = 0$.

- For the acoustic mode, all atoms in the unit cell move in-phase with each other, whereas for optical modes, they move out of phase with each other
- Except for the acoustic mode, all other excitation branches have zero group velocity for $k = n\pi/a$ for any n.
- If all of the dispersion curves are plotted within the first Brillouin zone $|k| \leqslant \pi/a$ we call this the reduced zone scheme. If we "unfold" the curves such that there is only one excitation plotted per k, but we use more than one Brillouin zone, we call this the extended zone scheme.

- If the two atoms in the unit cell become identical, the new unit cell is half the size of the old unit cell. It is convenient to describe this limit in the extended zone scheme.

References

- Ashcroft and Mermin, chapter 22 (but not the 3d part)
- Ibach and Luth, section 4.3
- Kittel, chapter 4
- Hook and Hall, sections 2.3.2, 2.4, 2.5
- Burns, section 12.3

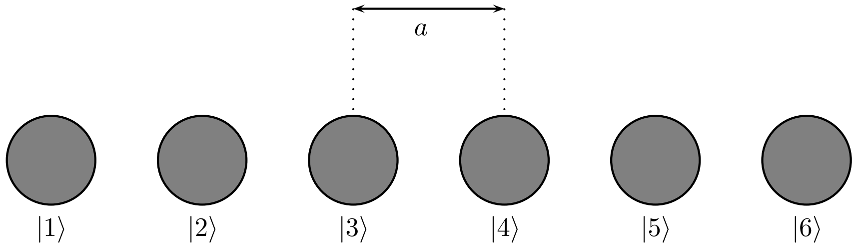
Chapter 7

One-dimensional tight binding model (interlude and preview)

In the previous two chapters we have considered the properties of vibrational waves (phonons) in a one dimensional system. At this point, we are going to make a bit of an excursion to consider electrons in solids again. The point of this excursion, besides being a preview of much of the physics that will re-occur later on, is to make the point that all waves in periodic environments (in crystals) are similar. In the previous two chapters we considered vibrational waves. In this chapter we will consider electron waves (Remember that in quantum mechanics particles are just as well considered to be waves!)

7.1 Tight binding model in 1d

We described the molecular orbital, or tight binding, picture for molecules previously in section 4.3.2. We also met the equivalent picture, or LCAO (linear combination of atomic orbitals) model of bonding for homework. What we will do here is consider a chain of such molecular orbitals to represent orbitals in a macroscopic (one dimensional) solid.



In this picture, there is a single orbital on atom n which we call $|n\rangle$. For convenience we will assume that the system has periodic boundary conditions (i.e, there are N sites, and site N

is the same as site 0). Further we will assume that all of the orbitals are orthogonal to each other.

$$\langle n|m\rangle = \delta_{n,m} \tag{7.1}$$

Let us now take a general trial wavefunction of the form

$$|\Psi\rangle = \sum_n \phi_n |n\rangle$$

As we showed for homework, the effective Schrödinger equation for this type of tight-binding model can be written as

$$\sum_m H_{nm}\phi_m = E\phi_n \tag{7.2}$$

where H_{nm} is the matrix element of the Hamiltonian

$$H_{nm} = \langle n|H|m\rangle$$

As mentioned previously when we studied the molecular orbital model, this Schrödinger equation is actually a variational approximation. For example, instead of finding the exact ground state, it finds the best possible ground state made up of the orbitals that we have put in the model.

One can make variational approach increasingly better by expanding the Hilbert space and putting more orbitals into the model. For example, instead of having only one orbital $|n\rangle$ at a given site, one could consider many $|n,\alpha\rangle$ where α runs from 1 to some number p. As p is increased the approach becomes increasingly more accurate and eventually is essentially exact. This method of using tight-binding like orbitals to increasingly well approximate the exact Schrödinger equation is known as LCAO (linear combination of atomic orbitals). However, one complication (which we treat only in one of the additional homework assignments) is that when we add many more orbitals we typically have to give up our nice orthogonality assumption, i.e., $\langle n,\alpha|m,\beta\rangle = \delta_{nm}\delta_{\alpha\beta}$ no longer holds. This makes the effective Schrödinger equation a bit more complicated, but not fundamentally different.

At any rate, in the current chapter we will work with only one orbital per site and we assume the orthogonality Eq. 7.1.

We write the Hamiltonian as

$$H = K + \sum_j V_j$$

where $K = p^2/(2m)$ is the kinetic energy and V_j is the coulomb interaction of the electron with the nucleus at site j.

$$V_j = V(r - r_j)$$

where r_j is the position of the j^{th} nucleus.

With these definitions we have

$$H|m\rangle = (K + V_m)|m\rangle + \sum_{j\neq m} V_j|m\rangle$$

Now, we should recognize that $K + V_m$ is the Hamiltonian which we would have if there were only a single nucleus (the m^{th} nucleus) and no other nuclei in the system. Thus, if we take the tight-binding orbitals $|m\rangle$ to be the atomic orbitals, then we have

$$(K + V_m)|m\rangle = \epsilon_{atomic}|m\rangle$$

where ϵ_{atomic} is the energy of an electron on nucleus m in the absence of any other nuclei.

Thus we can write

$$H_{n,m} = \langle n|H|m\rangle = \epsilon_{atomic}\, \delta_{n,m} + \sum_{j \neq m} \langle n|V_j|m\rangle$$

We now have to figure out what the final term of this equation is of this equation. The meaning of this term is that, via the interaction with some nucleus which is not the m^{th} an electron on the m^{th} atom can be transferred to the n^{th}. Generally this can only happen if n and m are very close to each other. Thus, we write

$$\sum_{j \neq m} \langle n|V_j|m\rangle = \begin{cases} V_0 & n = m \\ -t & n = m \pm 1 \\ 0 & \text{otherwise} \end{cases}$$

which defines both V_0 and t. (Note by translational invariance of the system, we expect that the result should depend only on $n - m$, which this form does).

Thus we obtain

$$H_{n,m} = \epsilon_0 \delta_{n,m} - t\,(\delta_{n+1,m} + \delta_{n-1,m}) \tag{7.3}$$

where we have now defined[1]

$$\epsilon_0 = \epsilon_{atomic} + V_0$$

This Hamiltonian is a very well known model. Here t is known as the hopping term, as it allows the Hamiltonian (which generates time evolution) to move the electron from one site to another, and it has dimensions of energy[2]. It stands to reason that the magnitude of t depends on how close together the orbitals are - becoming large when the orbitals are close together and decaying exponentially when they are far apart.

7.2 Solution of the tightbinding model in 1d

The solution of the tight binding model is very analogous to what we did to study vibrations (and hence the point of presenting the tight binding model at this point!). We propose an ansatz solution

$$\phi_m = \frac{e^{-ikna}}{\sqrt{N}} \tag{7.4}$$

where the denominator is included for normalization where there are N sites in the system. We now plug this ansatz into the Schrödinger equation Eq. 7.2. Note that in this case, there is no frequency in the exponent of our ansatz. This is simply because we are trying to solve the time-independent Schrödinger equation. Had we used the time dependent equation, we would need a factor of $e^{i\omega t}$ as well!

As with vibrations, it is obvious that $k \to k + 2\pi/a$ gives the same solution. Further, if we consider the system to have periodic boundary conditions with N sites (length $L = Na$), the allowed values of k are quantized in units of $2\pi/L$. As with Eq. 5.9 there are precisely N possible different solutions of the form of Eq. 7.4.

---

[1] Here ϵ_0 is *not* a dielectric constant or the permittivity of free space, but rather just the energy of having an electron sit on a site. (At some point we just run out of new symbols to use for new quantities!)

[2] The notation t is conventional in the field – except among atomic physicists who like to call it J.

Plugging the ansatz into the left side of the Schrödinger equation 7.2 and then using Eq. 7.3 gives us

$$\sum_m H_{n,m}\phi_m = \epsilon_0 \frac{e^{-ikna}}{\sqrt{N}} - t\left(\frac{e^{-ik(n+1)a}}{\sqrt{N}} + \frac{e^{-ik(n-1)a}}{\sqrt{N}}\right)$$

which we set equal to the right side of the Schrödinger equation

$$E\phi_m = E\,\frac{e^{-ikna}}{\sqrt{N}}$$

To obtain the spectrum

$$E = \epsilon_0 - 2t\cos(ka) \tag{7.5}$$

which looks rather similar to the phonon spectrum of the 1d monatomic chain which was (See Eq. 5.5)

$$\omega^2 = 2\frac{\kappa}{m} - 2\frac{\kappa}{m}\cos(ka)$$

Note however, that in the electronic case one obtains the energy whereas in the phonon case one obtains the *square* of the frequency.

This dispersion curve is shown in Fig. 7.1. Analogous to the phonon case, it is periodic in $k \to k + 2\pi/a$. Further, analogous to the phonon case, the dispersion always has zero group velocity (is flat) for $k = n\pi/a$ for n any integer.

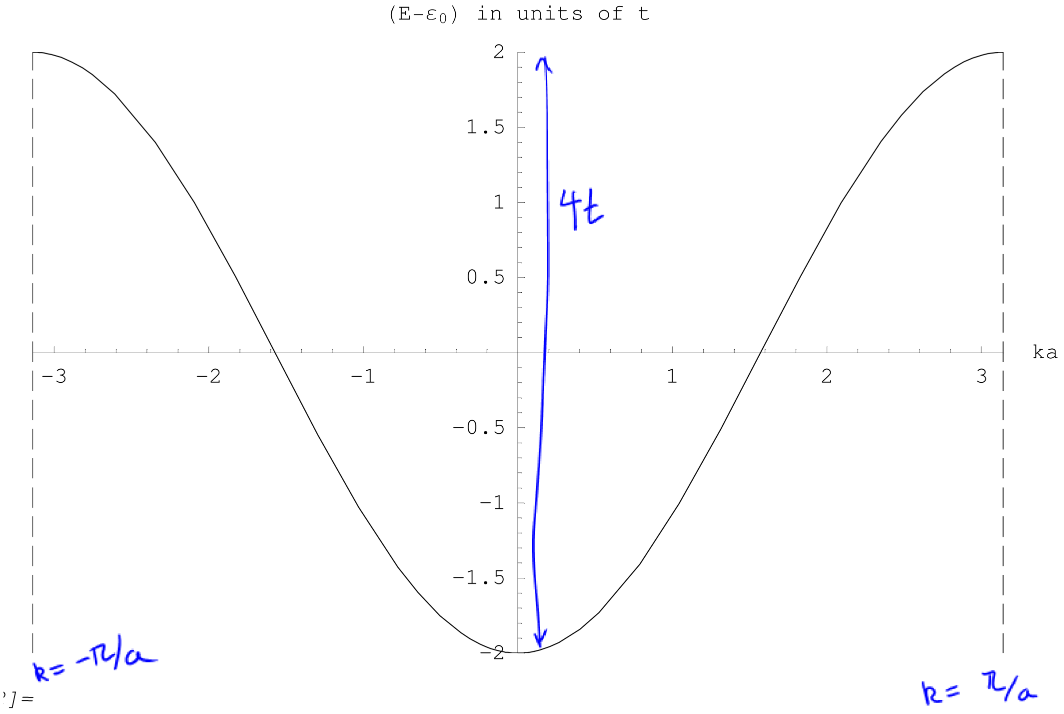


Figure 7.1: Tight binding dispersion band.

Note that unlike free electrons, the electron dispersion here has a maximum energy as well as a minimum energy. Electrons only have eigenstates within a certain energy *band*. The word "band" is used both to describe the energy range for which eigenstates exist, as well as to describe one connected piece of the dispersion curve (In this picture there is only a single piece of the dispersion curve, hence a single band).

The energy difference from the bottom of the band is known as the *bandwidth.* Within this bandwidth (between the top and bottom of the band) for any energy there exists (at least one) k state with that energy. Outside of the bandwidth there are no k-states with that energy.

The bandwidth (which in this model is $4t$) is determined by the magnitude of the hopping, which, as mentioned above depends on the distance between nuclei. As a function of the interatomic spacing then the bandwidth increases as shown in Fig 7.2. On the right of this diagram there are N states, each one being an atomic orbital $|n\rangle$. On the left of the diagram these N states form a band, yet as discussed above, there remain precisely N states. (This should not surprise us, being that we have not changed the dimension of the Hilbert state, we have just expressed it in terms of the complete set of eigenvalues of the Hamiltonian). Note that the average energy of a state in this band remains always zero.

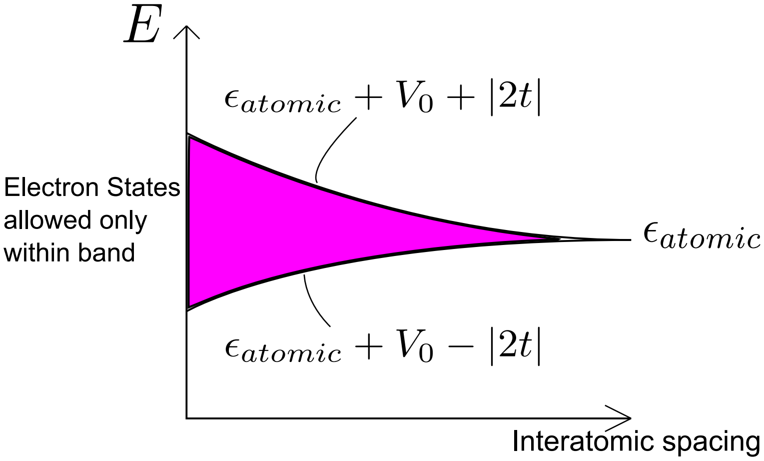


Figure 7.2: Caricature of spread of bandwidth with interatomic spacing.

Aside: Note that if the band is not completely filled, the *total* energy of all of the electrons *decreases* as the atoms are moved together and the band width increases. (Since the average energy remains zero, but some of the higher energy states are not filled). This decrease in energy is precisely the binding force of a "metallic bond" which we discussed in section 4.5. Of course we have not considered the repulsive force between neighboring nuclei, so the nuclei do not get too close together.

Near the bottom of the band, the dispersion is parabolic. For our above dispersion (Eq. 7.5), expanding for small k, we obtain

$$E(k) = \text{Constant} + ta^2k^2$$

[Note that for $t < 0$, the energy minimum is at the Brillouin zone boundary $k = \pi/a$. In this case we would expand for k close to π/a instead of for k close to 0]. The resulting parabolic behavior is similar to that of free electrons which have a dispersion

$$E_{free}(k) = \frac{\hbar^2k^2}{2m}$$

We can therefore view the bottom of the band as being almost like free electrons, except that we have to define a new *effective mass* which we call m^* such that

$$\frac{\hbar^2 k^2}{2m^*} = ta^2k^2$$

which gives us

$$m^* = \frac{\hbar^2}{2ta^2}$$

In other words, the *effective mass* m^* is defined such that the dispersion of the bottom of the band is exactly like the dispersion of free particles of mass m^*. Note that this mass has nothing to do with the actual mass of the electron, but rather depends on the hopping matrix element t. Further we should keep in mind that the k that enters into the dispersion relationship is actually the crystal momentum, not the actual momentum of the electron (recall that crystal momentum is defined only modulo $2\pi/a$). However, so long as we stay at very small k, then there is no need to worry about the periodicity of k which occurs. Nonetheless, we should keep in mind that if electrons scatter off of other electrons, or off of phonons, it is crystal momentum that is conserved. (See the discussion in section 5.2.4).

7.3 Introduction to electrons filling bands

We now imagine that our tight binding model is actually made up of atoms and each atom "donates" one electron into the band (i.e., the atom has *valence* one). Since there are N possible k-states in the band, and electrons are fermions, you might guess that this would precisely fill the band. However, there are two possible spin states for an electron at each k, so in fact, this then only half-fills the band. This is depicted in the left of Fig. 7.3. The filled states (shaded) in this picture are filled with both up and down spins.

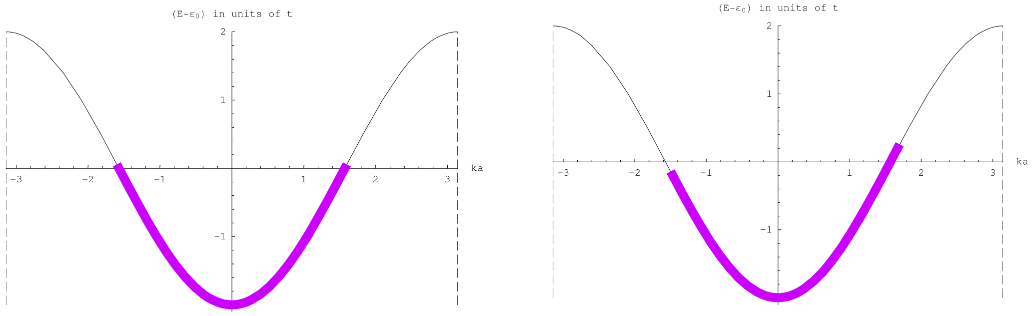


Figure 7.3: Left: If each atom has valence 1, then the band is half-filled. The states that are shaded are filled with both up and down spin electrons. The fermi surface is the boundary between the filled and unfilled states. Right: When a small electric field is applied, at only a small cost of energy, the fermi sea can shift slightly thus allowing current to run.

It is crucial in this picture that there is a Fermi surface — the points where the shaded meets the unshaded region. If a small electric field is applied to the system, it only costs a very small amount of energy to shift the fermi surface as shown in the right of Fig. 7.3, populating a few k-states moving right and de-populating some k-states moving left. In other words, the state of the system responds by changing a small bit and a current is induced. As such, this system is a *metal* in that it conducts electricity. Indeed, crystals of atoms that are mono-valent are very frequently metals!

On the other hand, if each atom in our model were di-valent (donates two electrons to the band) then the band would be entirely full of electrons. In fact, it does not matter if we think about this as being a full band where every k-state $|k\rangle$ is filled with two electrons (one up and one down), or a filled band where every site $|n\rangle$ is filled — these two statements describe the same multi-electron wavefunction. In fact, there is a single unique wavefunction that describes this completely filled band.

In the case of the filled band, were one to apply a small electric field to this system, the system cannot respond at all. There is simply no freedom to repopulate the occupation of k-states because every state is already filled. Thus we conclude an important principle

Principle: A filled band carries no current.

Thus our example of a di-valent tightbinding model is an insulator. (This type of insulator is known as a *band insulator*). Indeed, many systems of di-valent atoms are indeed insulators (although we will discuss how this principle can fail below as well).

7.4 Multiple bands

In the above model, we considered only the case where there is a single atom in the unit cell and a single orbital per atom. However, more generally we might consider a case where we have multiple orbitals per unit cell.

One possibility is to consider one atom per unit cell, but several orbitals per atom[3]. Analogous to what we found with the above tight binding model, when the atoms are very far apart, one has only the atomic orbitals on each atom. However, as the atoms are moved closer together, the orbitals merge together and the energies spread to form bands[4]. Analogous to Fig. 7.2 we have shown how this occurs for the two band case in Fig. 7.4.

A very similar situation occurs when we have two atoms per unit cell but only one orbital per atom. We will do a problem like this for homework. However, the general result will be quite analogous to what we found for vibrations of a diatomic chain in chapter 6.

In Fig. 7.5 we show the spectrum of a tight-binding model with two different atoms per unit cell – each having a single orbital. We have shown results here in both the reduced and extended zone schemes.

As for the case of vibrations, we see that there are now two possible energy eigenstates at each value of k. In the language of electrons, we say that there are two bands (we do not use the words "acoustic" and "optical" for electrons, but the idea is similar). Note that there is a gap between the two bands where there are simply no energy eigenstates.

Let us think for a second about what should result in this situation. If each atom (of either

---

[3] Each atom actually has an infinite number of orbitals to be considered. But only a small number of them are filled, and within our level of approximation, we can only consider very few of them.

[4] This picture of atomic orbitals in the weak hopping limit merging together to form bands does not depend on the fact that the crystal of atoms is ordered. Glasses and amorphous solids can have this sort of band structure as well!

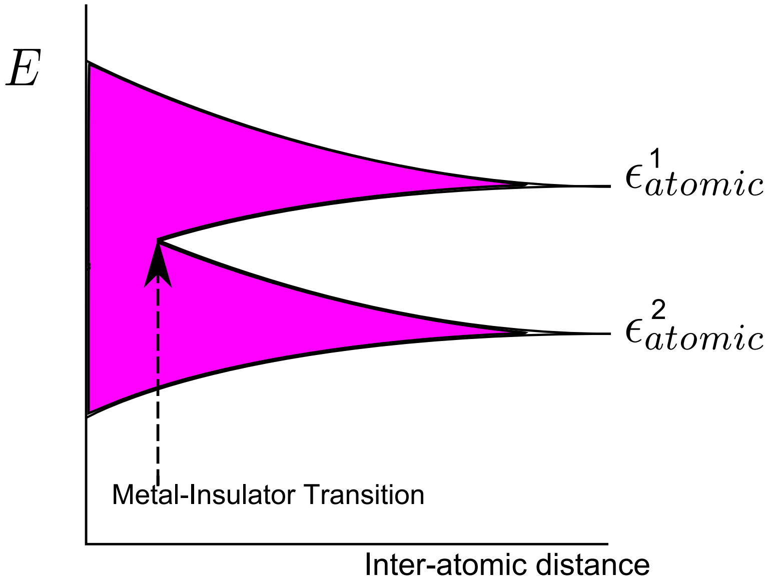


Figure 7.4: Charicature of spread of bandwidth with interatomic spacing for a two-band model. In the atomic limit, the orbitals have energies ϵ^1 and ϵ^2

type) were divalent, then the two electrons donated would completely fill the single orbital on each site. In this case, both bands would be completely filled with both spin-up and spin-down electrons.

On the other hand, if each atom (of either type) is monovalent, then this means exactly half of the states of the system should be filled. However, here, when one fills half of the states of the system, then all of the states of the lower band are completely filled (with both spins) but all of the states in the upper band are completely empty. In the extended zone scheme it appears that a gap has opened up precisely where the Fermi surface is! (at the Brillouin zone boundary!)

In the situation where a lower band is completely filled but an upper band is completely empty, if we apply a weak electric field to the system can current flow? In this case, one cannot rearrange electrons within the lower band, but one can remove an electron from the lower band and put it in the upper band in order to change the overall (crystal) momentum of the system. However, moving an electron from the lower band requires a finite amount of energy — one must overcome the gap between the bands. As a result, for small enough electric fields, this cannot happen. We conclude that a filled band is an insulator as long as there is a finite gap to any higher empty bands.

As with the single band case, one can imagine the magnitude of hopping changing as one changes the distance between atoms. When the atoms are far apart, then one is in the atomic limit, but these atomic states spread into bands as the atoms get closer together as shown in Fig. 7.4.

For the case where each atom is mono-valent, in the atomic limit, half of the states are filled – that is the lower energy atomic orbital is filled with both spin-up and spin down electrons whereas the higher energy orbital is completely empty. (I.e., an electron is transferred from the higher energy atom to the lower energy atom and this completely fills the lower energy band). As the atoms are brought closer together, the atomic orbitals spread into bands (the hopping t

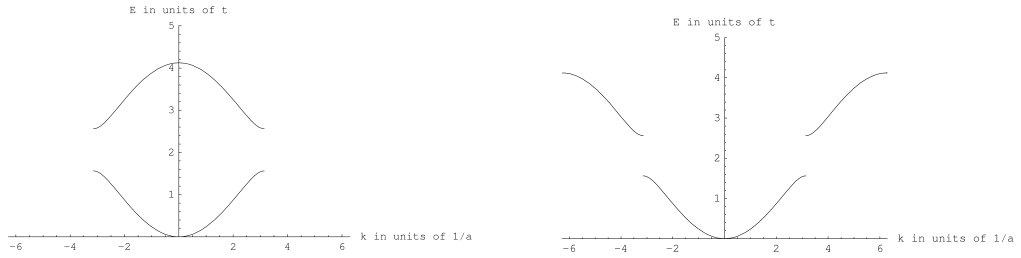


Figure 7.5: Diatomic Tight Binding Problem in 1D. Left: Reduced Zone scheme. Right: Extended Zone scheme.

increases). However, at some point the bands get so wide that their energies overlap — in which case there is no gap to transfer electrons between bands, and the system becomes a metal as marked in Fig. 7.4.

7.5 Summary of 1d tight binding model

- Solving tight-binding Shroedinger equation for electron waves is very similar to solving equations for vibrational (phonon) waves. The structure of the reciprocal lattice and the Brillouin zone remains the same.
- Obtain energy bands where energy eigenstates exist, and gaps between bands.
- Zero hopping is the atomic limit, as hopping increases, atomic orbitals spread into bands.
- Energies are parabolic in k near bottom of band — looks like free electrons, but with a modified effective mass.
- A filled band with a gap to the next band is an insulator (a *band* insulator), a partially filled band has a fermi surface and is a metal.
- Whether a band is filled depends on the valence of the atoms.
- As we found for phonons, gaps open at Brillouin zone boundaries. Group velocities are also zero at zone boundaries.

References

No book has an approach to tight binding that is exactly like what we have here. The books that come closest do essentially the same thing, but in three dimensions (which complicates life a bit). These books are:

- Ibach and Luth, section 7.3
- Kittel, chapter 9, section on tight-binding

- Burns, section 10-9, and 10-10.
- Singleton, chapter 4.

Possibly the nicest (albeit short) description is given by

- Dove, section 5.5.5

Also a nice short description of the physics (without any detail is given by)

- Rosenberg, section 8.19.

Finally, an alternative approach to tight binding is given by

- Hook and Hall, section 4.3.

The discussion of Hook and Hall is good (and they consider one dimension, which is nice), but they insist on using time dependent Schrödinger equation, which is annoying.

Part IV

Geometry of Solids

Chapter 8

Crystal Structure

Having introduced a number of important ideas in one dimension, we must now deal with the fact that our world is actually spatially three dimensional. While this adds a bit of complication, really the important concepts are no harder in three dimensions than they were in one dimension. Some of the most important ideas we have already met in 1d, but we will reintroduce more generally here.

There are two parts that might be difficult here. First, we do need to wrestle with a bit of geometry. Hopefully most will not find this too hard. Secondly will also need to establish a language in order to describe structures in 2 and 3d intelligently. As such, much of this chapter is just a list of definitions to be learned, but unfortunately this is necessary in order to be able to carry further at this point.

8.1 Lattices and Unit Cells

Definition 8.1.1. A *Lattice* is an infinite set of points defined by integer sums of a set of linearly independent primitive basis[1] vectors[2]

For example, in two dimensions, as shown in figure 8.1 the lattice points are described as

$$\mathbf{R}_{[n_1,n_2]} = n_1\mathbf{a_1} + n_2\mathbf{a_2} \qquad\qquad n_1, n_2 \in \mathbb{Z} \qquad\qquad (2d)$$

with $\mathbf{a_1}$ and $\mathbf{a_2}$ being the primitive basis vectors and n_1 and n_2 being integers. In three dimensions points of a lattice are analogously indexed by three integers

$$\mathbf{R}_{[n_1,n_2,n_3]} = n_1\mathbf{a_1} + n_2\mathbf{a_2} + n_3\mathbf{a_3} \qquad\qquad n_1, n_2, n_3 \in \mathbb{Z} \qquad\qquad (3d) \qquad\qquad (8.1)$$

Note that in 1d this definition of a lattice fits with our previous description of a lattice as being the points $R = na$ with n an integer.

---

[1] We have already used the word "basis" before in chapter 6.1, and unfortunately, here this is a different use of the same word. In this definition we mean the usual mathematical definition of a basis vector. I.e., a choice of linearly independent vectors.

[2] Warning: Some books (Ashcroft and Mermin in particular) refer to this as a *Bravais Lattice*. This enables them to use the term *Lattice* to describe other things that we would not call a lattice (cf, the honeycomb). However, the definition we use here is more common, and more correct mathematically as well. [Thank you, Mike Glazer, for catching this].

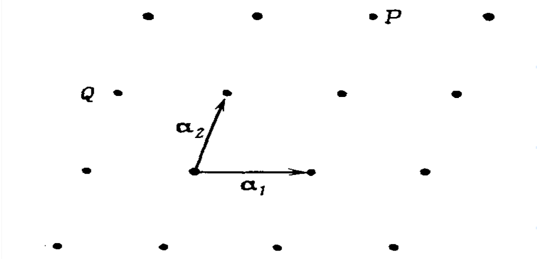


Figure 8.1: A lattice is defined as integer sums of a set of primitive basis vectors.

It is important to point out that in 2 and 3 dimensions, the choice of primitive basis vectors is not unique[3] as show in figure 8.2. (In 1d, the single primitive basis vector is unique up to the sign (direction) of a).

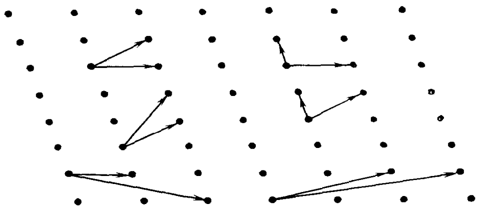


Figure 8.2: The choice of primitive basis vectors for a lattice is not unique.

It turns out that there are several definitions that are entirely equivalent to the one we have just given:

Equivalent Definition 8.1.1.1. A *Lattice* is an infinite set of vectors where addition of any two vectors in the set gives a third vector in the set.

It is easy to see that our above first definition 8.1.1 implies the second one 8.1.1.1. Here is a less crisply defined, but sometimes more useful definition.

Equivalent Definition 8.1.1.2. A *Lattice* is a set of points where the environment of any given point is equivalent to the environment of any other given point.

It turns out that any periodic structure can be expressed as a lattice of repeating motifs. A cartoon of this statement is shown in Fig. 8.3.

---

[3] Given a set of primitive basis vectors $\mathbf{a_i}$ a new set of primitive basis vectors may be constructed as $\mathbf{b_i} = \sum_j m_{ij} \mathbf{a_j}$ so long as m_{ij} is an invertible matrix with integer entries *and* the inverse matrix $[m^{-1}]_{ij}$ also has integer entries.

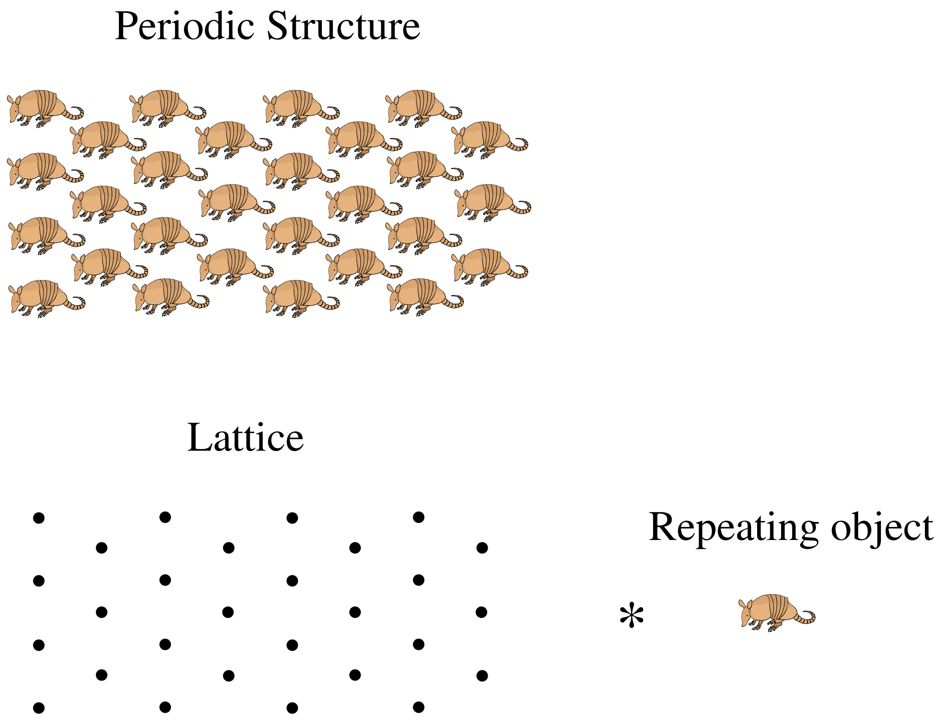


Figure 8.3: Any periodic structure can be represented as a lattice of repeating motifs.

One should be cautious however, that not all periodic arrangements of points are lattices. The honeycomb[4] shown in Fig. 8.4 is *not* a lattice. This is obvious from the third definition 8.1.1.2: The environment of point P and point R are actually different — point P has a neighbor directly above it (the point R), whereas point R has no neighbor directly above.

In order to describe a honeycomb (or other more complicated arrangements of points) we have the idea of a unit cell, which we have met before in section 6.1 above. Generally we have

Definition 8.1.2. A *unit cell* is a region of space, such that when many identical units are stacked together it tiles (completely fills) all of space and reconstructs the full structure

An equivalent (but less rigorous) definition is that the

Equivalent Definition 8.1.2.1. A *unit cell* is the repeated motif which is the elementary building block of the periodic crystal.

To be more specific we frequently want to work with the smallest possible unit cell

Definition 8.1.3. A *primitive unit cell* for a periodic crystal is a unit cell containing only a single lattice point.

---

[4] One should be careful *not* to call this a hexagonal lattice. First of all, by our definition it, is not a lattice at all since all points do not have the same environment. Secondly, some people use the term “hexagonal” to mean what the rest of us call a triangular lattice: a lattice of triangles where each point has six nearest neighbor points. (See Fig 8.6 below)

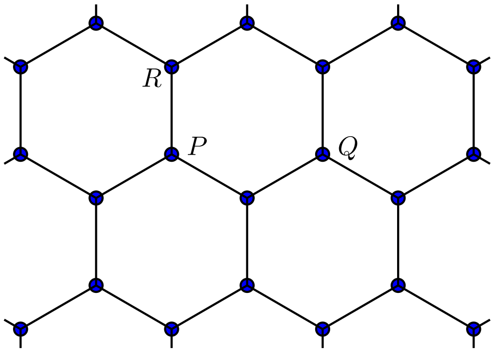


Figure 8.4: The honeycomb is not a lattice. Points P and R are inequivalent. (Points P and Q are equivalent)

As mentioned above in section 6.1 the definition of the unit cell is never unique. This is shown, for example, for a lattice in Fig. 8.5

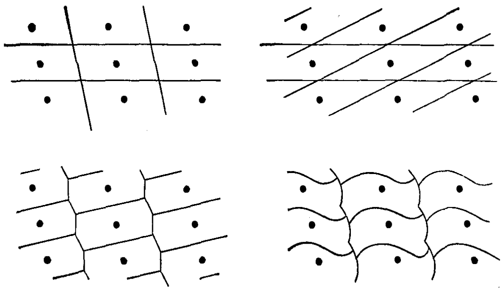


Figure 8.5: The choice of a unit cell is not unique. All of these unit cells reconstruct the same crystal.

Sometimes it is useful to define a unit cell which is not primitive in order to have it simpler to work with. This is known as a *conventional unit cell.* Almost always these conventional unit cells are chosen so as to have orthogonal axes.

Some examples of possible unit cells are shown for the triangular lattice in fig 8.6. In this figure the conventional unit cell (upper left) is in this case is chosen to have orthogonal axes – which is often easier to work with than axes which are non-orthogonal.

A note about counting the number of lattice points in the unit cell: It is frequently the case

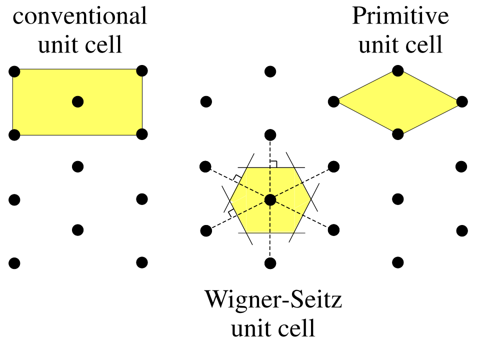


Figure 8.6: Some unit cells for the triangular lattice.

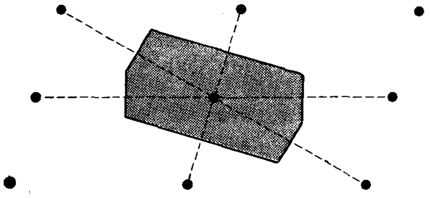


Figure 8.7: The Wigner-Seitz construction for a lattice in 2d.

that we will work with unit cells where the lattice points live at the corners (or edges) of the cells. When a lattice point is on the boundary of the unit cell, it should only be counted fractionally depending on what fraction of the point is actually in the cell. So for example in the conventional unit cell shown in Fig. 8.6, there are two lattice points within this cell. Obviously there is one point in the center, then four points at the corners — each of which is one quarter inside the cell, so we obtain $2 = 1 + 4(\frac{1}{4})$ points in the cell. (Since there are two points in this cell, it is by definition, not primitive. Similarly for the primitive cell shown in this figure (upper right), the two lattice points at the left and the right have a 60^o degree slice (which is $1/6$ of a circle) inside the cell. The two points at the top and the bottom have $1/3$ of the point inside the unit cell. Thus this unit cell contains $2(\frac{1}{3}) + 2(\frac{1}{6}) = 1$ point, and is thus primitive. Note however, that we can just imagine shifting the unit cell a tiny amount in almost any direction such that a single lattice point is completely inside the unit cell and the others are completely outside the unit cell. This sometimes makes counting much easier.

Also shown in Fig. 8.6 is a so-called *Wigner-Seitz unit cell*[5].

Definition 8.1.4. Given a lattice point, the set of all points in space which are closer to that given

---

[5] Eugene Wigner was yet another Nobel laureate who was another one of the truly great minds of the last century of physics. Fredrick Seitz was far less famous, but gained notoriety in his later years by being a consultant for the tobacco industry, a strong proponent of the Regan-era Star-Wars missile defense system, and a prominent sceptic of global warming. He passed away in 2007.

lattice point than to any other lattice point constitute the *Wigner-Seitz cell* of the given lattice point.

There is a rather simple scheme for constructing such a Wigner-Seitz cell: choose a lattice point and draw lines to all of its possible near neighbors (not just its nearest neighbors). Then draw perpendicular bisectors of all of these lines. The perpendicular bisectors bound the Wigner-Seitz cell[6]. It is always true that the Wigner-Seitz construction for a lattice gives a primitive unit cell. In figure 8.7 we show another example of the Wigner-Seitz construction for a 2d lattice. A similar construction can be performed in 3d in which case one must construct perpendicular-bisecting planes to bound the Wigner-Seitz cell.

The description of objects in the unit cell in terms of the reference point in the unit cell is known as a "basis". (This is the same definition of "basis" we used in section 6.1 above).

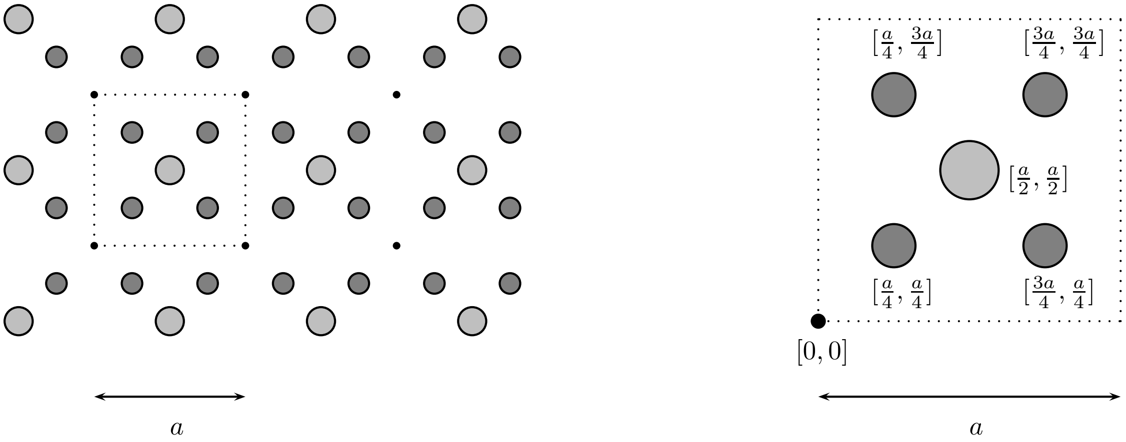


Figure 8.8: Left: A periodic structure in two dimensions. A unit cell is marked with the dotted lines. Right: A blow-up of the unit cell with the coordinates of the particles in the unit cell with respect to the reference point in the lower left hand corner. The basis is the description of the atoms along with these positions.

In Fig. 8.8 we show a periodic structure in two dimension made of two types of atoms On the right we show a primitive unit cell (expanded) with the position of the atoms given with respect to the reference point of the unit cell which is taken to be the lower left-hand corner. We can describe the basis of this crystal as follows:

---

[6] This Wigner-Seitz construction can be done on an irregular lattice as well as on a periodic lattice. For an irregular lattice the resulting object is known as a Voronoi cell.

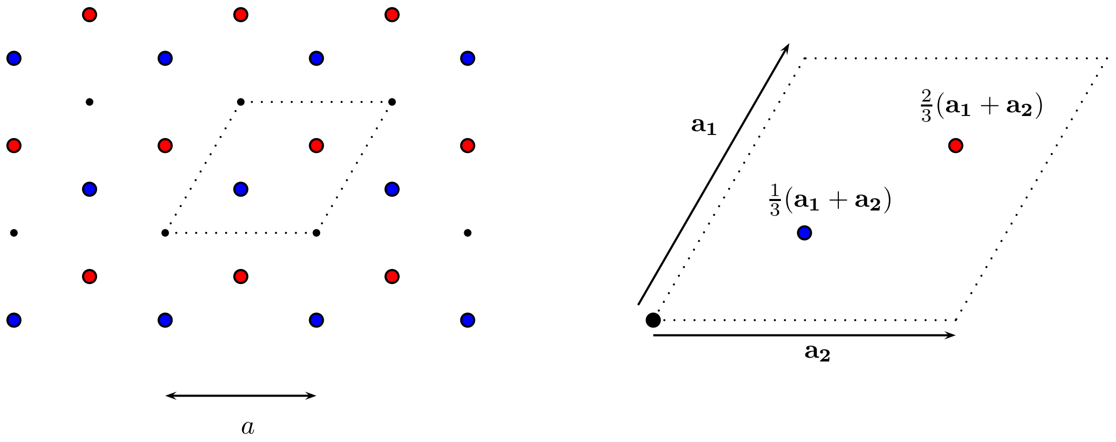


Figure 8.9: The honeycomb from Fig. 8.4 with the two inequivalent points of the unit cell given different colors. The unit cell is outlined dotted on the left and the corners of the unit cell are marked with small black dots. On the right the unit cell is expanded with coordinates with respect to the reference point written.

Basis for crystal in Fig. 8.8 =		
Large Light Gray Atom	Position=	$[a/2, a/2]$
Small Dark Gray Atoms	Position=	$[a/4, a/4]$ $[a/4, 3a/4]$ $[3a/4, a/4]$ $[3a/4, 3a/4]$

The reference points forming the square lattice have positions

$$\mathbf{R}_{n_1,n_2} = [a\,n_1, a\,n_2] = a\,n_1\hat{\boldsymbol{x}} + a\,n_2\hat{\boldsymbol{y}}$$

with n_1, n_2 integers so that the large light gray atoms have positions

$$\mathbf{R}^{light-gray}_{n_1,n_2} = [a\,n_1, a\,n_2] + [a/2, a/2]$$

whereas the small dark gray atoms have positions

$$\begin{aligned}
\mathbf{R}^{dark-gray1}_{n_1,n_2} &= [a\,n_1, a\,n_2] + [a/4, a/4] \\
\mathbf{R}^{dark-gray2}_{n_1,n_2} &= [a\,n_1, a\,n_2] + [a/4, 3a/4] \\
\mathbf{R}^{dark-gray3}_{n_1,n_2} &= [a\,n_1, a\,n_2] + [3a/4, a/4] \\
\mathbf{R}^{dark-gray4}_{n_1,n_2} &= [a\,n_1, a\,n_2] + [3a/4, 3a/4]
\end{aligned}$$

In this way you can say that the positions of the atoms in the crystal are “the lattice plus the basis”.

We can now return to the case of the honeycomb shown in Fig. 8.4 above. Here, the reference points (small black dots) form a (triangular) lattice, where we can write the primitive lattice vectors as

$$\begin{aligned} \mathbf{a_1} &= a\,\hat{\boldsymbol{x}} \\ \mathbf{a_2} &= (a/2)\,\hat{\boldsymbol{x}} + (a\sqrt{3}/2)\,\hat{\boldsymbol{y}} \end{aligned}$$

In terms of the reference points of the lattice, the basis for the primitive unit cell, i.e., the coordinates of the two blue dots with respect to the reference point, are given by $\frac{1}{3}(\mathbf{a_1} + \mathbf{a_2})$ and $\frac{2}{3}(\mathbf{a_1} + \mathbf{a_2})$.

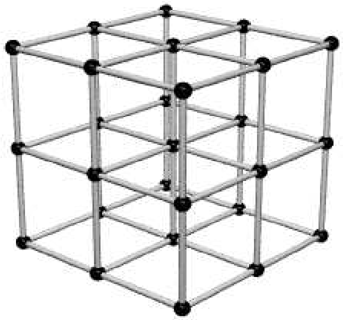


Figure 8.10: A simple cubic lattice

8.2 Lattices in 3D

The simplest lattice in three dimensions is the simple cubic lattice shown in Fig. 8.10 (sometimes known as cubic "P" or cubic-primitive lattice). The primitive unit cell in this case can most conveniently be taken to be single cube – which includes 1/8 of each of its eight corners.

In fact, real crystals of atoms are rarely simple cubic[7]. If you think of an atom as a small sphere, if you assemble spheres into a simple cubic lattice you find that it is a very inefficient way to pack the spheres together – in that you are left with a lot of empty space in the center of the unit cells.

Only slightly more complicated than the simple cubic lattice are the *tetragonal* and *orthorhombic* lattices where the axes remain perpendicular, but the primitive lattice vectors may be of different lengths (shown in Fig 8.11). The orthorhombic unit cell has three different lengths of its perpendicular lattice primitive basis vectors, whereas the tetragonal unit cell has two lengths the same and one different.

Conventionally to represent a given vector amongst the infinite number of possible lattice vectors in a lattice, one writes

$$[uvw] = u\mathbf{a_1} + v\mathbf{a_2} + w\mathbf{a_3} \tag{8.2}$$

where u,v and w are integers. For cases where the lattice vectors are orthogonal, the basis vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are assumed to be in the $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$ directions. (We have seen this notation before in the subscripts of the Eqns. after Definition 8.1.1).

---

[7] Of all of the chemical elements, Polonium is the only one which forms a simple cubic lattice.

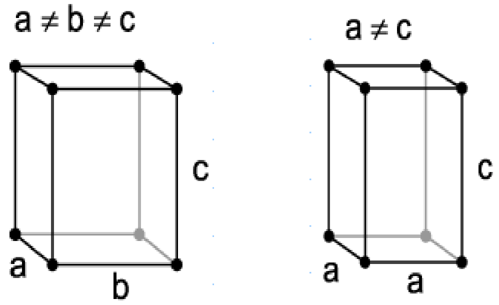


Figure 8.11: Unit cells for orthorhombic (left) and tetragonal (right) lattices.

Lattices in 3d certainly exist where axes are non-orthogonal, but... *you will not be held responsible for any 3d crystal system where coordinate axes are not orthogonal.*

Two further lattice systems that you will need to know are the Face Centered Cubic (fcc) and Body Centered Cubic (bcc) lattices.

The Body Centered Cubic (bcc) Lattice

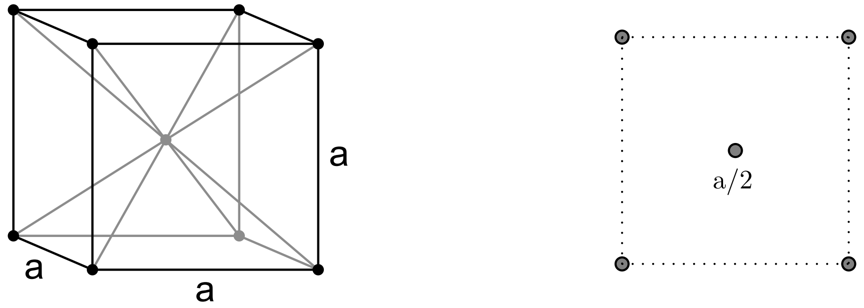


Figure 8.12: Conventional unit cell for the body centered cubic (I) lattice. Left: 3D view. Right: A plan view of the conventional unit cell. Unlabeled points are both at heights 0 and a.

The Body centered (bcc) lattice is a simple cubic lattice where there is an additional point in the very center of the cube (this is sometimes known[8] as cubic-I). The unit cell is shown in the left of Fig. 8.12. Another way to show this unit cell, which does not rely on showing a three-dimensional picture, is to use a so-called *plan view* of the unit cell, shown in the right of

---

[8] Actually, I don't know where this notation comes from.

Fig. 8.12. A plan view (a term used in engineering and architecture) is a 2d projection from the top of an object where heights are labeled to show the third dimension. In the picture of the bcc unit cell, there are eight lattice points on the corners of the cell (each of which is 1/8 inside of the conventional unit cell) and one point in the center of the cell. Thus the conventional unit cell contains exactly two ($= 8 \times 1/8 + 1$) lattice points.

Packing together these unit cells to fill space, we see that the lattice points of a full bcc lattice can be described as being points having coordinates $[x, y, z]$ where either all three coordinates are integers $[uvw]$ times the lattice constant a, or all three are odd-half-integers times the lattice constant a.

It is often convenient to think of the bcc lattice as a simple cubic lattice with a basis of two atoms per conventional cell. The simple cubic lattice contains points $[x, y, z]$ where either all three coordinates are integers in units of the lattice constant. Within the conventional simple-cubic-unit cell we put one point at position $[0, 0, 0]$ and another point at the position $[1/2, 1/2, 1/2]$ in units of the lattice constant. Thus the points of the bcc lattice are written as

$$\begin{aligned}
\mathbf{R}_{corner} &= a[n_1, n_2, n_3] \\
\mathbf{R}_{center} &= a[n_1, n_2, n_3] + a[1/2, 1/2, 1/2]
\end{aligned}$$

as if the two different types of points were two different types of atoms, although all points in this lattice should be considered equivalent (they only look inequivalent because we have chosen a conventional unit cell with two lattice points in it).

Now, we may ask why it is that this set of points forms a lattice. In terms of our first definition of a lattice (Definition 8.1.1) we can write the primitive basis vectors of the bcc lattice as

$$\begin{aligned}
\mathbf{a_1} &= [a,\, 0,\, 0] \\
\mathbf{a_2} &= [0,\, a,\, 0] \\
\mathbf{a_3} &= [\frac{a}{2}, \frac{a}{2}, \frac{a}{2}]
\end{aligned}$$

It is easy to check that any combination

$$\mathbf{R} = n_1\mathbf{a_1} + n_2\mathbf{a_2} + n_3\mathbf{a_3} \tag{8.3}$$

with n_1, n_2 and n_3 integers gives a point within our definition of the bcc lattice (that the three coordinates are either all integer or all half-odd integer times the lattice constant). Further one can check that any point satisfying the conditions for the bcc lattice can be written in the form of Eq. 8.3.

We can also check that our description of a bcc lattice satisfies our second description of a lattice (definition 8.1.1.1) that addition of any two points of the lattice (given by Eq. 8.3) gives another point of the lattice.

More qualitatively we can consider definition 8.1.1.2 of the lattice — that the local environment of every point in the lattice should be the same. Examining the point in the center of the unit cell, we see that it has precisely 8 nearest neighbors in each of the possible diagonal directions. Similarly any of the points in the corners of the unit cells will have 8 nearest neighbors corresponding to the points in the center of the 8 adjacent unit cells.

The *coordination number* of a lattice (frequently called Z or z) is the number of nearest neighbors any point of the lattice has. For the bcc lattice the coordination number is $Z = 8$.

As in two dimensions, a Wigner-Seitz cell can be constructed around each lattice points which encloses all points in space that are closer to that lattice point than to any other point in the lattice. This Wigner-Seitz unit cell for the bcc lattice is shown in Figure 8.13. Note that this cell is bounded by the perpendicular bisecting planes between lattice points.

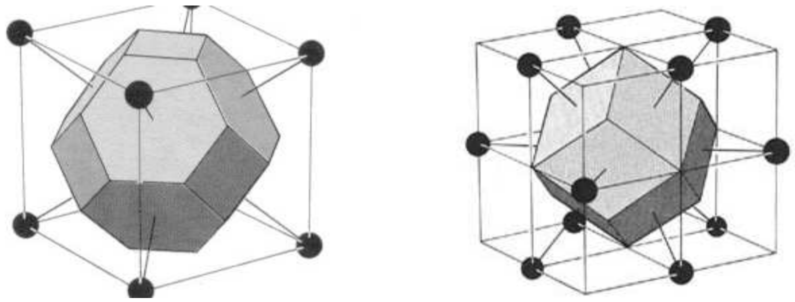


Figure 8.13: Wigner-Seitz unit cell for the bcc lattice (left) and the fcc lattice (right).

The Face Centered Cubic (fcc) Lattice

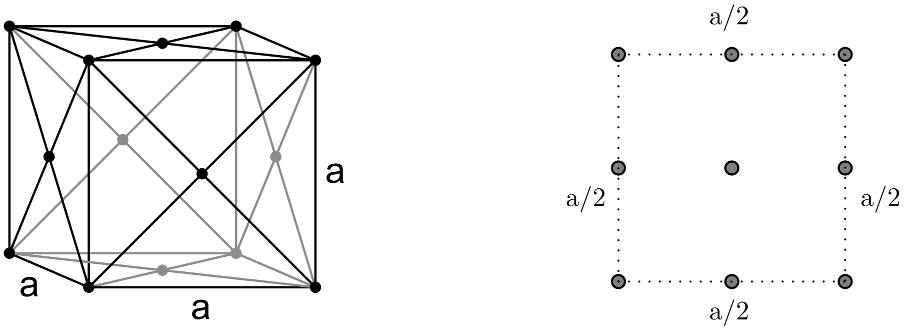


Figure 8.14: Conventional unit cell for the face centered cubic (F) lattice. Left: 3D view. Right: A plan view of the conventional unit cell. Unlabeled points are both at heights 0 and a.

The face centered (fcc) lattice is a simple cubic lattice where there is an additional point in the center of every face of every cube (this is sometimes known as cubic-F, for "face centered"). The unit cell is shown in the left of Fig. 8.14. A plan view is of the unit cell is shown on the right of Fig. 8.14.

In the picture of the fcc unit cell, there are eight lattice points on the corners of the cell (each of which is 1/8 inside of the conventional unit cell) and one point in the center of each of the 6 faces, which is 1/2 inside the cell. Thus the conventional unit cell contains exactly four ($= 8 \times 1/8 + 6 \times 1/2$) lattice points. Packing together these unit cells to fill space, we see that the lattice points of a full fcc lattice can be described as being points having coordinates (x, y, z) where either all three coordinates are integers times the lattice constant a, or two of the three coordinates are odd-half-integers times the lattice constant a and the remaining one coordinate is an integer times the lattice constant a. Analogous to the bcc case, it is sometimes convenient to think of the fcc lattice as a simple cubic lattice with a basis of four atoms per conventional cell. The simple cubic lattice contains points $[x, y, z]$ where all three coordinates are integers in units of the lattice constant a. Within the conventional simple-cubic-unit cell we put one point at position $[0, 0, 0]$ and another point at the position $[a/2, a/2, 0]$ another at $[a/2, 0, a/2]$ and another at $[0, a/2, a/2]$. Thus the points of the fcc lattice are written as

$$\begin{aligned}
\mathbf{R}_{corner} &= a[n_1, n_2, n_3] \qquad (8.4)\\
\mathbf{R}_{face-xy} &= a[n_1, n_2, n_3] + a[1/2, 1/2, 0]\\
\mathbf{R}_{face-xz} &= a[n_1, n_2, n_3] + a[1/2, 0, 1/2]\\
\mathbf{R}_{face-yz} &= a[n_1, n_2, n_3] + a[0, 1/2, 1/2]
\end{aligned}$$

Again, this expresses the points of the lattice as if they were four different types of points but they only look inequivalent because we have chosen a conventional unit cell with four lattice points in it.

Again we can check that this set of points forms a lattice. In terms of our first definition of a lattice (Definition 8.1.1) we write the primitive basis vectors of the fcc lattice as

$$\begin{aligned}
\mathbf{a_1} &= [\frac{a}{2}, \frac{a}{2}, 0]\\
\mathbf{a_2} &= [\frac{a}{2}, 0, \frac{a}{2}]\\
\mathbf{a_3} &= [0, \frac{a}{2}, \frac{a}{2}]
\end{aligned}$$

Again it is easy to check that any combination

$$\mathbf{R} = n_1\mathbf{a_1} + n_2\mathbf{a_2} + n_3\mathbf{a_3} \qquad (8.5)$$

with n_1, n_2 and n_3 integers gives a point within our definition of the fcc lattice (that either the three coordinates are either all integer, or two of three are half-odd-integer and the remaining is integer in units of the lattice constant a).

We can also similarly check that our description of a fcc lattice satisfies our other two definitions of (definition 8.1.1.1 and 8.1.1.2) of a lattice. The Wigner-Seitz unit cell for the fcc lattice is shown in Figure 8.13.

Other Lattices in 3d

In addition to the simple cubic, orthorhombic, tetragonal, fcc, and bcc lattices, there are nine other types of lattices in 3d. These are known as the fourteen *Bravais lattice types*[9] *You are not responsible for knowing these!* But it is probably a good idea to know that they exist.

---

[9] Named after Auguste Bravais who classified all the lattices in 3d in 1848. Actually they should be named after Moritz Frankenheim who studied the same thing over ten years earlier – although made a minor error in his studies, and therefore missed getting his name associated with them.

Figure 8.15 shows the full variety of Bravais lattice types in 3 dimensions[10]

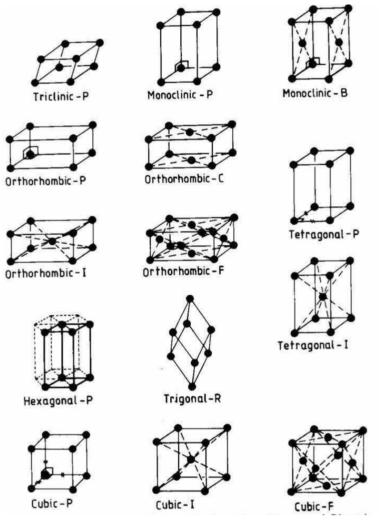


Figure 8.15: Unit cells for all of the 3d Bravais Lattice types.

Real Crystals

Once we have discussed lattices we can combine a lattice with a basis to describe any periodic structure — and in particular, we can describe any crystalline structure.

Several examples of real (and reasonably simple) crystal structures are shown in Fig. 8.16.

8.3 Summary of crystal structure

This chapter introduced a plethora of new definitions, aimed at describing crystal structure in 3 dimensions. Here is a list of some of the concepts that one should know

- Definition of a *lattice* (in three different ways See definitions 8.1.1, 8.1.1.1, 8.1.1.2)
- Definition of a *unit cell* for a periodic structure, and definition of a *primitive unit cell* and a *conventional unit cell*
- Definition and construction of the *Wigner-Seitz (primitive) unit cell.*
- One can write any periodic structure in terms of a lattice and a basis. (See examples in Fig. 8.16)
- In 3d, know the simple cubic lattice, the fcc lattice and the bcc lattice.
- The fcc and bcc lattices can be thought of as simple cubic lattices with a basis.
- Know how to read a *plan view* of a structure.

---

[10] While it is an extremely deep fact that there are only 14 lattice types in 3d, the precise statement of this theorem, as well of the proof of it, are beyond the scope of this course.

References

All books cover this. Some books give *way* too much detail for us. I recommend the following as giving not too much and not too little:

- Kittel, chapter 1
- Ashcroft and Mermin chapter 4 (caution of the nomenclature issue, see footnote 2) of this chapter .

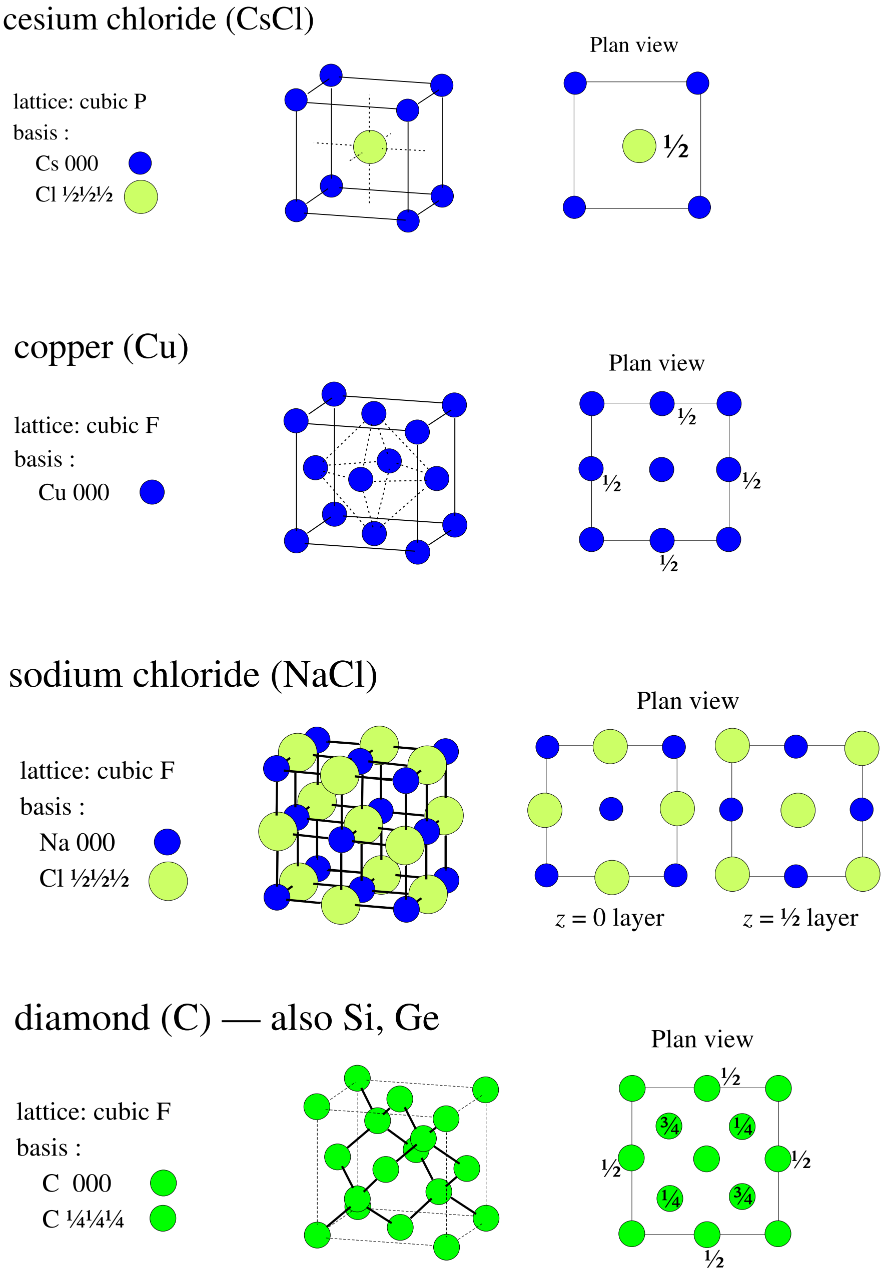


Figure 8.16: Some examples of real crystals with simple structures.

Chapter 9

Reciprocal Lattice, Brillouin Zone, Waves in Crystals

In the last chapter we explored lattices and crystal structure. However as we saw in chapters 5-7, the important physics of waves in solids (whether they be vibrational waves, or electron waves) is best described in reciprocal space. This chapter thus introduces reciprocal space in 3 dimensions. As with the previous chapter, there is some tricky geometry in this chapter, and a few definitions to learn as well. This makes this material a bit tough to slog through, but stick with it because soon we will make substantial use of what we learn here. At the end of this chapter we will finally have enough definitions to describe the dispersions of phonons and electrons in 3d systems.

9.1 The Reciprocal Lattice in 3d

9.1.1 Review of 1d

Let us first recall some results from 1d. We consider a simple lattice in 1d $R_n = na$ with n an integer. Recall that two points in k-space (reciprocal space) were defined to be equivalent to each other if $k_1 = k_2 + G_m$ where $G_m = 2\pi m/a$ with m an integer. The points G_m form the reciprocal lattice.

Recall that the reason that we identified different k values was because we were considering waves of the form

$$e^{ikx_n} = e^{ikna}$$

with n an integer. Because of this form of the wave, we find that shifting $k \to k + G_m$ leaves this functional form unchanged since

$$e^{i(k+G_m)x_n} = e^{i(k+G_m)na} = e^{ikna}e^{i(2\pi m/a)na} = e^{ikx_n}$$

where we have used

$$e^{i2\pi mn} = 1$$

in the last step. Thus, so far as the wave is concerned, k is the same as $k + G_m$.

9.1.2 Reciprocal Lattice Definition

Generalizing the above result from 1d, we make the following definition:

Definition 9.1.1. Given a (direct) lattice of points $\mathbf{R}$, a point $\mathbf{G}$ is a point in the reciprocal lattice iff

$$e^{i\mathbf{G}\cdot\mathbf{R}} = 1 \tag{9.1}$$

for all points $\mathbf{R}$ of the direct lattice.

To construct the reciprocal lattice, let us write the points of the direct lattice in the form (Here we specialize to the 3d case).

$$\mathbf{R} = n_1\mathbf{a_1} + n_2\mathbf{a_2} + n_3\mathbf{a_3} \tag{9.2}$$

with n_1, n_2 and n_3 integers, and with $\mathbf{a_1}$, $\mathbf{a_2}$, and $\mathbf{a_3}$ being primitive basis vectors of the direct lattice.

We now make two key claims:

1. We claim that the reciprocal lattice (defined by Eq. 9.1) is a lattice in reciprocal space (thus explaining its name).
2. We claim that the primitive basis vectors of the reciprocal lattice (which we will call $\mathbf{b_1}$, $\mathbf{b_2}$, and $\mathbf{b_3}$) are defined to have the following property:

   $$\mathbf{a_i}\cdot\mathbf{b_j} = 2\pi\delta_{ij} \tag{9.3}$$

   where δ_{ij} is the Kronecker delta[1].

We can certainly construct vectors $\mathbf{b_i}$ to have the desired property of Eq. 9.3, as follows:

$$\begin{aligned}
\mathbf{b_1} &= \frac{2\pi\,\mathbf{a_2}\times\mathbf{a_3}}{\mathbf{a_1}\cdot(\mathbf{a_2}\times\mathbf{a_3})} \\
\mathbf{b_2} &= \frac{2\pi\,\mathbf{a_3}\times\mathbf{a_1}}{\mathbf{a_1}\cdot(\mathbf{a_2}\times\mathbf{a_3})} \\
\mathbf{b_3} &= \frac{2\pi\,\mathbf{a_1}\times\mathbf{a_2}}{\mathbf{a_1}\cdot(\mathbf{a_2}\times\mathbf{a_3})}
\end{aligned}$$

It is easy to check that Eq. 9.3 is satisfied. For example,

$$\begin{aligned}
\mathbf{a_1}\cdot\mathbf{b_1} &= \frac{2\pi\,\mathbf{a_1}\cdot(\mathbf{a_2}\times\mathbf{a_3})}{\mathbf{a_1}\cdot(\mathbf{a_2}\times\mathbf{a_3})} = 2\pi \\
\mathbf{a_2}\cdot\mathbf{b_1} &= \frac{2\pi\,\mathbf{a_2}\cdot(\mathbf{a_2}\times\mathbf{a_3})}{\mathbf{a_1}\cdot(\mathbf{a_2}\times\mathbf{a_3})} = 0
\end{aligned}$$

Now, given vectors $\mathbf{b_1}$, $\mathbf{b_2}$, and $\mathbf{b_3}$ satisfying Eq. 9.3 we have claimed that these are in fact primitive lattice vectors for the reciprocal lattice.

---

[1] Leopold Kronecker was a mathematician who is famous (among other things) for the sentence "God made the integers, everything else is the work of man". In case you don't already know this, the Kronecker delta is defined as $\delta_{ij} = 1$ for $i = j$ and is zero otherwise. (Kronecker did a lot of other interesting things as well)

Let us write an arbitrary point in reciprocal space as

$$\mathbf{G} = m_1\mathbf{b_1} + m_2\mathbf{b_2} + m_3\mathbf{b_3} \tag{9.4}$$

and for a the moment, let us not require m_1, m_2 and m_3 integers. (We are about to discover that for $\mathbf{G}$ to be a point of the reciprocal lattice, they must be integers, but this is what we want to prove!).

To find points of the reciprocal lattice we must show that Eq. 9.1 is satisfied for all points $\mathbf{R} = n_1\mathbf{a_1} + n_2\mathbf{a_2} + n_3\mathbf{a_3}$ of the direct lattice with n_1, n_2 and n_3 integers. We thus write

$$e^{i\mathbf{G}\cdot\mathbf{R}} = e^{i(m_1\mathbf{b_1}+m_2\mathbf{b_2}+m_3\mathbf{b_3})\cdot(n_1\mathbf{a_1}+n_2\mathbf{a_2}+n_3\mathbf{a_3})} = e^{2\pi i(n_1m_1+n_2m_2+n_3m_3)}$$

In order for $\mathbf{G}$ to be a point of the reciprocal lattice, this must equal unity for all points $\mathbf{R}$ of this direct lattice, i.e., for all integer values of n_1, n_2 and n_3. Clearly this can only be true if m_1, m_2 and m_3 are also integers. Thus, we find that the points of the reciprocal lattice are precisely those of the form of Eq. 9.4 with m_1, m_2 and m_3 integers. This further proves our claim that the reciprocal lattice is in fact a lattice!

9.1.3 The Reciprocal Lattice as a Fourier Transform

Quite generally one can think of the Reciprocal lattice as being a Fourier transform of the direct lattice. It is easiest to start by thinking in one dimension. Here the direct lattice is given again by $R_n = an$. If we think of the "density" of lattice points in one dimension, we might put a delta-function of density at these lattice points so we write the density as[2]

$$\rho(r) = \sum_n \delta(r - an)$$

Fourier transforming this function gives[3]

$$\mathcal{F}[\rho(r)] = \int dr e^{ikr}\rho(r) = \sum_n \int dr e^{ikr}\delta(r - an) = \sum_n e^{ikan} = 2\pi \sum_m \delta(k - 2\pi m/a)$$

The last step here is a bit nontrivial.[4] Here e^{ikan} is clearly unity if $k = 2\pi m/a$, i.e., if k is a point on the reciprocal lattice. In this case, each term of the sum contributes unity to the sum and one obtains an infinite result. If k is not such a reciprocal lattice point, then the terms of the sum oscillate and the sum comes out to be zero.

This principle generalizes to the higher dimensional (2d and 3d) cases. Generally

$$\mathcal{FT}[\rho(\mathbf{r})] = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} = (2\pi)^D \sum_{\mathbf{G}} \delta^D(\mathbf{k} - \mathbf{G}) \tag{9.5}$$

---

[2] Since the sums are over all lattice points they should go from $-\infty$ to $+\infty$. Alternately, one uses periodic boundary conditions and sums over all points.

[3] With Fourier transforms there are many different conventions about where one puts the factors of 2π. Probably in your mathematics class you learned to put $1/\sqrt{2\pi}$ with each k integral and with each r integral. However, in Solid-State physics conventionally $1/(2\pi)$ comes with each k integral, and no factor of 2π comes with each r integral. See section 2.2.1 to see why this is used.

[4] This is sometimes known as the Poisson resummation formula, after Siméon Denis Poisson, the same guy after whom Poisson's equation $\nabla^2\phi = -\rho/\epsilon_0$ is named, as well as other mathematical things such as the Poisson random distribution. His last name means "fish" in French.

where in the middle term, the sum is over lattice points $\mathbf{R}$ of the direct lattice, and in the last term it is a sum over points $\mathbf{G}$ of the reciprocal lattice. Here D is the number of dimensions (1, 2 or 3) and the δ^D is a D-dimensional delta function[5]. This equality is similar to that explained above. As above, if $\mathbf{k}$ is a point of the reciprocal lattice, then $e^{i\mathbf{k}\cdot\mathbf{R}}$ is always unity and the sum is infinite (a delta function). However, if $\mathbf{k}$ is not a point on the reciprocal lattice then the summands oscillate, and the sum comes out to be zero. Thus one obtains delta function peaks precisely at the positions of reciprocal lattice vectors.

Aside: It is an easy exercise to show[6] that the reciprocal lattice of an fcc direct lattice is a bcc lattice in reciprocal space. Conversely, the reciprocal lattice of a bcc direct lattice is an fcc lattice in reciprocal space.

Fourier Transform of Any Periodic Function

In the above section we considered the Fourier transform of a function $\rho(\mathbf{r})$ which is periodic in space being just a set of delta functions at lattice points. However, it is not too different to consider the Fourier transform of any function with the periodicity of the lattice (and this will be quite important below in chapter 10 below). Consider a function $\rho(\mathbf{r})$ which is periodic $\rho(\mathbf{r}) = \rho(\mathbf{r}+\mathbf{R})$ for any lattice vector $\mathbf{R}$. We then have

$$\mathcal{F}[\rho(\mathbf{r})] = \int \mathbf{dr}\; e^{i\mathbf{k}\cdot\mathbf{r}}\rho(\mathbf{r})$$

The integral over all of space can be broken up into a sum of integrals over each unit cell. Here we write any point in space $\mathbf{r}$ a lattice point $\mathbf{R}$ and a vector $\mathbf{x}$ within the unit cell

$$\mathcal{F}[\rho(\mathbf{r})] = \sum_{\mathbf{R}} \int_{unit-cell} \mathbf{dx}\; e^{i\mathbf{k}\cdot(\mathbf{x}+\mathbf{R})}\rho(\mathbf{x}+\mathbf{R}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} \int_{unit-cell} \mathbf{dx}\; e^{i\mathbf{k}\cdot\mathbf{x}}\rho(\mathbf{x})$$

where here we have use the invariance of ρ under lattice translations $\mathbf{x} \rightarrow \mathbf{x}+\mathbf{R}$. The first term, as in Eq. 9.5 just gives a sum of delta functions yielding

$$\mathcal{F}[\rho(\mathbf{r})] = (2\pi)^D \sum_{\mathbf{G}} \delta^D(\mathbf{k}-\mathbf{G})S(\mathbf{k})$$

where

$$S(\mathbf{k}) = \int_{unit-cell} \mathbf{dx}\; e^{i\mathbf{k}\cdot\mathbf{x}}\rho(\mathbf{x}) \tag{9.6}$$

is known as the *structure factor* and will become very important in the next chapter.

9.1.4 Reciprocal Lattice Points as Families of Lattice Planes

Another way to understand the reciprocal lattice is via families of lattice planes of the direct lattice.

Definition 9.1.2. A *lattice plane* (or *crystal plane*) is a plane containing at least three noncolinear (and therefore an infinite number of) points of a lattice.

Definition 9.1.3. A *family of lattice planes* is an infinite set of equally separated lattice planes

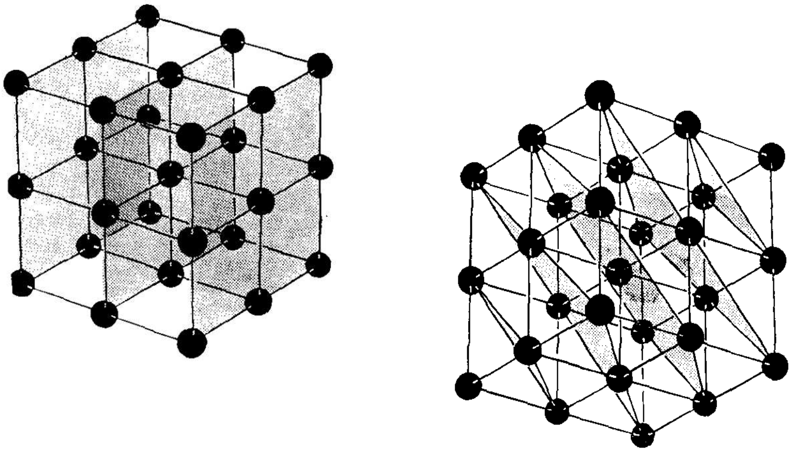


Figure 9.1: Two examples of families of lattice planes on the cubic lattice. Each of these planes is a crystal plane because it intersects an infinite number of lattice points. The left example is (1,0,0) and the right example is (1,1,1) in the Miller Index notation.

In Figure 9.1, two examples of families of lattice planes are shown. Note that the planes are parallel and equally spaced, and every point of the lattice is included in exactly one lattice plane.

I now make the following claim:

Claim: *The families of lattice planes are in one-to-one correspondence*[7] *with the possible directions of reciprocal lattice vectors, to which they are normal. Further the spacing between these lattice planes is* $d = 2\pi/|\mathbf{G_{min}}|$ *where* $\mathbf{G_{min}}$ *is the minimum length reciprocal lattice vector in this normal direction.*

This correspondence is made as follows. First we consider the set of planes defined by points $\mathbf{r}$ such that

$$\mathbf{G} \cdot \mathbf{r} = 2\pi m \tag{9.7}$$

This defines an infinite set of parallel planes normal to $\mathbf{G}$. Since $e^{i\mathbf{G}\cdot\mathbf{r}} = 1$ we know that every lattice point is a member of a one of these planes (since this is the definition of $\mathbf{G}$ in Eq. 9.1). However, for the planes defined by Eq. 9.7, not every plane needs to contain a lattice point (so generically this is a family of parallel equally spaced planes, but not a family of lattice planes).

---

[5] For example, in two dimensions $\delta^2(\mathbf{r} - \mathbf{r_0}) = \delta(x - x_0)\delta(y - y_0)$ where $\mathbf{r} = (x, y)$

[6] Try it!

[7] For this to be precisely true we must define $\mathbf{G}$ and $-\mathbf{G}$ to be the same direction. If this sounds like a cheap excuse, we can say that "oriented" families of lattice planes are in 1-to-1 correspondence with the directions of reciprocal lattice vectors, thus keeping track of the two possible normals of the family of lattice planes.

For this larger family of planes, the spacing between planes is given by

$$d = \frac{2\pi}{|\mathbf{G}|} \tag{9.8}$$

To prove this we simply note that two adjacent planes must have

$$\mathbf{G} \cdot (\mathbf{r_1} - \mathbf{r_2}) = 2\pi$$

Thus in the direction parallel to $\mathbf{G}$, the spacing between planes is $2\pi/|\mathbf{G}|$ as claimed.

Clearly different values of $\mathbf{G}$ that happen to point in the same direction, but have different magnitudes, will define parallel sets of planes. As we increase the magnitude of $\mathbf{G}$, we add more and more lattice planes. For example, examining Eq. 9.7 we see that when we double the magnitude of $\mathbf{G}$ we correspondingly double the density of planes, which we can see from the spacing formula Eq. 9.8. However, whichever $\mathbf{G}$ we choose, all of the lattice points will be included in one of the defined planes. If we choose the maximally possible spaced planes, hence the smallest possible value of $\mathbf{G}$ allowed in any given direction which we call $\mathbf{G_{min}}$, then in fact every defined plane will include lattice points and therefore be lattice planes, and the spacing between these planes is correspondingly $2\pi/|\mathbf{G_{min}}|$.[8] This proves our above claim.

9.1.5 Lattice Planes and Miller Indices

It is convenient to define a notation for describing lattice planes. The conventional notations are known as *Miller Indices*.[9] One writes

$$(h, k, l)$$

with integers h, k and l, to mean a family of lattice planes corresponding to reciprocal lattice vector

$$\mathbf{G}_{(h,k,l)} = h\mathbf{b_1} + k\mathbf{b_2} + l\mathbf{b_3} \tag{9.9}$$

where $\mathbf{b_i}$ are the standard primitive basis vectors of the reciprocal lattice[10]. Note that (h, k, l) as a family of lattice planes, should be the shortest reciprocal lattice vector in that direction, meaning that the integers h, k and l should have no common divisor. One may also write (h, k, l) where h, k and l do have a common divisor, but then one is talking about a reciprocal lattice vector, or a family of planes that is not a family of lattice planes (i.e., there are some planes that do not intersect lattice point).

Important Complication: For fcc and bcc lattices, Miller indices are usually stated using the primitive basis vectors of the *cubic* lattice in Eq. 9.9 rather than the primitive basis vector of the fcc or bcc.

This comment is quite important. For example, the (100) family of planes for the cubic lattice (shown in the right of Fig. 9.1) intersects every corner of the cubic unit cell. However, if we were discussing a bcc lattice, there would also be another lattice point in the center of every

---

[8] More rigorously, if there is a family of lattice planes in direction $\widehat{\mathbf{G}}$ with spacing between planes d, then $\mathbf{G} = 2\pi\widehat{\mathbf{G}}/d$ is necessarily a reciprocal lattice vector. To see this note that $e^{i\mathbf{G}\cdot\mathbf{R}} = 1$ will be unity for all lattice points. Further, in a family of lattice planes, all lattice points are included within the planes, so $e^{i\mathbf{G}\cdot\mathbf{R}} = 1$ for all $\mathbf{R}$ a lattice point, which implies $\mathbf{G}$ is a reciprocal lattice vector. Furthermore, $\mathbf{G}$ is the shortest reciprocal lattice vector in the direction of $\widehat{\mathbf{G}}$ since increasing $\mathbf{G}$ will result in a smaller spacing of lattice planes and some planes will not intersect lattice points $\mathbf{R}$.

[9] These are named after the 19th century mineralogist William Hallowes Miller. It is interesting that the structure of lattice planes was understood long before the world was even certain there was such a thing as an atom.

[10] We have already used the corresponding notation $[uvw]$ to represent lattice points of the direct lattice. See for example, Eq. 8.1 and Eq. 8.2.

Figure 9 This plane intercepts the a, b, c axes at 3a, 2b, 2c. The reciprocals of these numbers are $\frac{1}{3}, \frac{1}{2}, \frac{1}{2}$. The smallest three integers having the same ratio are 2, 3, 3, and thus the Miller indices of the plane are (233).

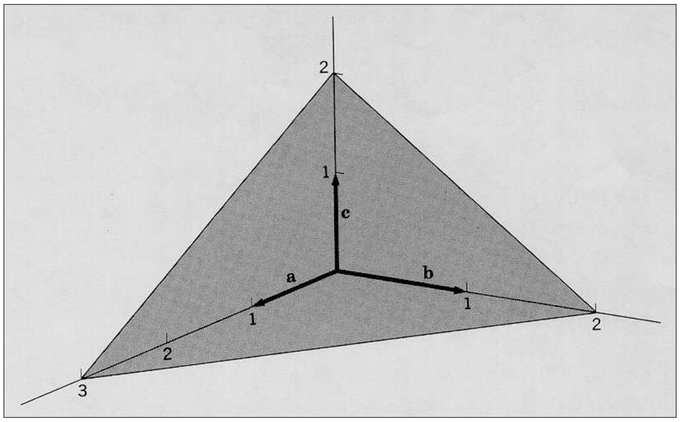


Figure 9.2: Determining miller indices from the intersection of a plane with the coordinate axes. The spacing between lattice planes in this family would be $\frac{1}{|d_{(2,3,3)}|^2} = \frac{2^2}{a^2} + \frac{3^2}{b^2} + \frac{3^2}{c^2}$.

conventional unit cell, and the (100) lattice planes would not intersect. However,the (200) planes would interesect these points as well, so in this case (200) represents a true family of lattice planes whereas (100) does not!

From Eq. 9.8 one can write the spacing between a family of planes specified by Miller indices (h, k, l)

$$d_{(h,k,l)} = \frac{2\pi}{|\mathbf{G}|} = \frac{2\pi}{h^2|\mathbf{b_1}|^2 + k^2|\mathbf{b_2}|^2 + l^2|\mathbf{b_3}|^2} \tag{9.10}$$

Where we have assumed that the coordinate axes of the primative basis vectors $\mathbf{b_i}$ are orthogonal. Recall that $|b_i| = 2\pi/|a_i|$ where a_i are the lattice constants in the three orthogonal directions. Thus we can equivalently write

$$\frac{1}{|d_{(h,k,l)}|^2} = \frac{h^2}{a_1^2} + \frac{k^2}{a_2^2} + \frac{l^2}{a_3^2} \tag{9.11}$$

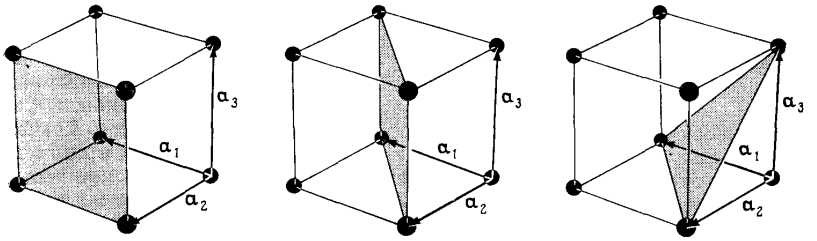
Note that for a cubic lattice this simplifies to

$$d_{(hkl)}^{cubic} = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \tag{9.12}$$

A useful shortcut for figuring out the geometry of lattice planes is to look at the intersection of a plane with the three coordinate axes. The intersections x_1, x_2, x_3 with the three coordinate axes (in units of the three principle lattice constants) are related to the Miller indices via

$$\frac{a_1}{x_1} : \frac{a_2}{x_2} : \frac{a_3}{x_3} = h : k : l$$

In figure Fig. 9.3 we show three more examples of Miller indices.



(010) (110) (111)

Figure 9.3: More examples of Miller Indices.

Note that Miller indices can be negative if the planes intersect the negative axes. We could have, for example, a lattice plane (1,-1,1). Conventionally, the minus sign is denoted with an over-bar rather than a minus sign, so we write $(1\bar{1}1)$ instead[11].

Finally, we note that different lattice planes may be the same under a symmetry of the crystal. For example, in a cubic lattice, (111) looks the same as $(1\bar{1}1)$ after rotation (and possibly reflection) of the axes of the crystal (but would never look like (122) under any rotation or reflection since the spacing between planes is different!). If we want to describe all lattice planes that are equivalent in this way, we write $\{111\}$ instead.

---

[11] How $(1\bar{1}1)$ is pronounced is a bit random. Some people say “one-(bar-one)-one” and others say “one-(one-bar)-one”. I have no idea how the community got so confused as to have these two different conventions. I think in Europe the former is more prevalent whereas in America the latter is more prevalent. At any rate, it is always clear when it is written.

9.2 Brillouin Zones

The whole point of going into such gross detail about the structure of reciprocal space is in order to describe waves in solids. In particular, it will be important to understand the structure of the Brillouin zone.

9.2.1 Review of 1d Dispersions and Brillouin Zones

As we learned in chapters 5-7, the Brillouin zone is extremely important in describing the excitation spectrum of waves in periodic media. As a reminder, in Fig. 9.4 we show the excitation spectrum of vibrations of a diatomic chain (chapter 6 in both the reduced, and extended zone schemes). Since waves are physically equivalent under shifts of the wavevector k by a reciprocal lattice vector $2\pi/a$, we can always express every excitation within the first Brillouin zone, as shown in the reduced zone scheme (left of Fig. 9.4). In this example, since there are two atoms per unit cell, there are precisely two excitation modes per wavevector. On the other hand, we can always unfold the spectrum and put the lowest (acoustic) excitation mode in the first Brillouin zone and the higher energy excitation mode (optical) in the second Brillouin zone, as shown in the extended zone scheme (right of Fig.9.4). Note that there is a jump in the excitation spectrum at the Brillouin zone boundary.

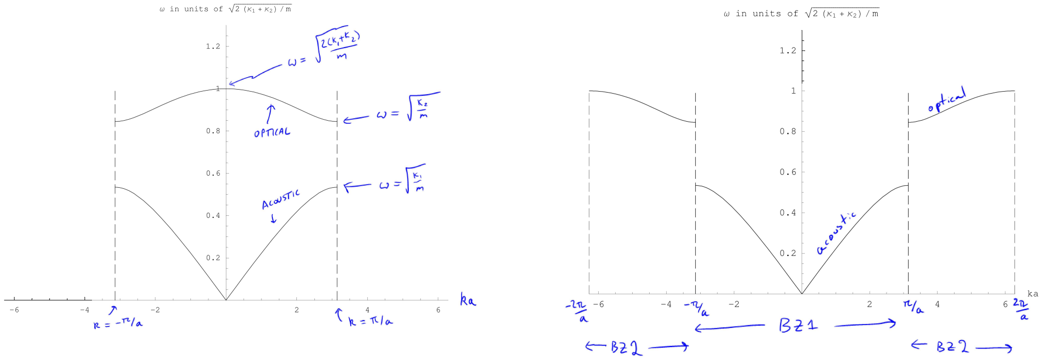


Figure 9.4: Phonon spectrum of a diatomic chain in 1D. Left: Reduced Zone scheme. Right: Extended Zone scheme. (See Figs. 6.1 and 6.2)

9.2.2 General Brillouin Zone Construction

Definition 9.2.1. A *Brillouin zone* is a unit cell of the reciprocal lattice.

Entirely equivalent to the 1d situation, physical waves in crystals are unchanged if their wavevector is shifted by a reciprocal lattice vector $\mathbf{k} \to \mathbf{k} + \mathbf{G}$. Alternately, we realize that the physically relevant quantity is the crystal momentum. Thus, the Brillouin zone has been defined to include each physically different crystal momentum exactly once (Each $\mathbf{k}$ point within the Brillouin zone is physically different and all physically different points occur once within the zone).

While the most general definition of Brillouin zone allows us to choose any shape unit cell for the reciprocal lattice, there are some definitions of unit cells which are more convenient than others.

We define the *first Brillouin zone* in reciprocal space quite analogously to the construction of the Wigner-Seitz cell for the direct lattice.

Definition 9.2.2. Start with the reciprocal lattice point $\mathbf{G} = \mathbf{0}$. All $\mathbf{k}$ points which are closer to $\mathbf{0}$ than to any other reciprocal lattice point define the *first Brillouin zone.* Similarly all point where the point $\mathbf{0}$ is the *second* closest reciprocal lattice point to that point constitute the *second Brillouin zone*, and so forth. *Zone boundaries* are defined in terms of this definition of Brillouin zones.

As with the Wigner-Seitz cell, there is a simple algorithm to construct the Brillouin zones. Draw the perpendicular bisector between the point $\mathbf{0}$ and each of the reciprocal lattice vectors. These bisectors form the Brillouin zone boundaries. Any point that you can get to from $\mathbf{0}$ without crossing a reciprocal lattice vector is in the first Brillouin zone. If you cross only one perpendicular bisector, you are in the second Brillouin zone, and so forth.

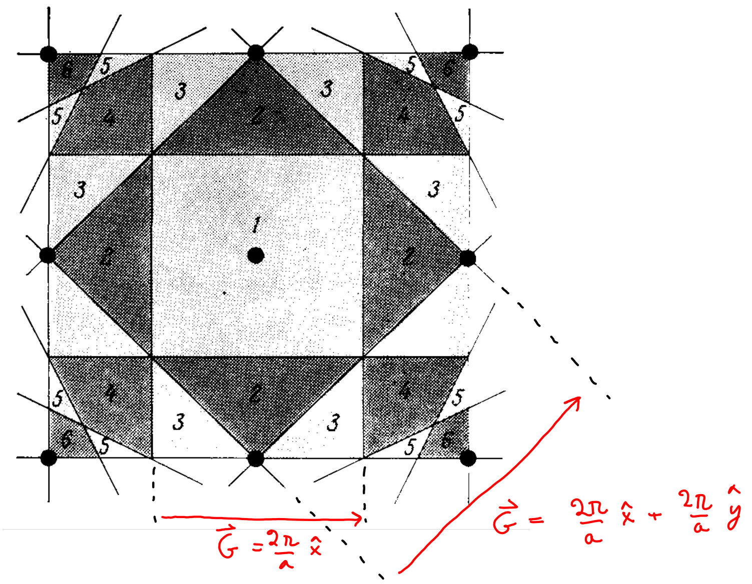


Figure 9.5: First, second, third, fourth, ... Brillioun zones of the square lattice. Note that zone boundaries occur in parallel pairs symmetric around $\mathbf{0}$ and separated by a reciprocal lattice vector.

In figure 9.5, we show the Brillouin zones of the square lattice. A few general principles to note:

1. The first Brillouin zone is necessarily connected, but the higher Brillouin zones typically are made of disconnected pieces.

2. A point on a Brillouin zone boundary lies on the perpendicular bisector between the point $\mathbf{0}$ and some reciprocal lattice point $\mathbf{G}$. Adding the vector $-\mathbf{G}$ to this point necessarily results in a point (the same distance from $\mathbf{0}$) which is on another Brillouin zone boundary (on the bisector of the segment from $\mathbf{0}$ to $-\mathbf{G}$). This means that Brillouin zone boundaries occur in parallel pairs symmetric around the point $\mathbf{0}$ which are separated by a reciprocal lattice vector (See Fig. 9.5).

3. Each Brillouin zone has exactly the same total area (or volume in 3d). This must be the case since there is a one-to-one mapping of points in each Brillouin zone to the first Brillouin zone. Finally, as in 1d, we claim that there are exactly as many k-states within the first Brillouin zone as there are unit cells in the entire system[12].

Note, that as in the case of the Wigner Seitz cell construction, the shape of the first Brillouin zone can look a bit strange, even for a relatively simple lattice (See Fig. 8.7).

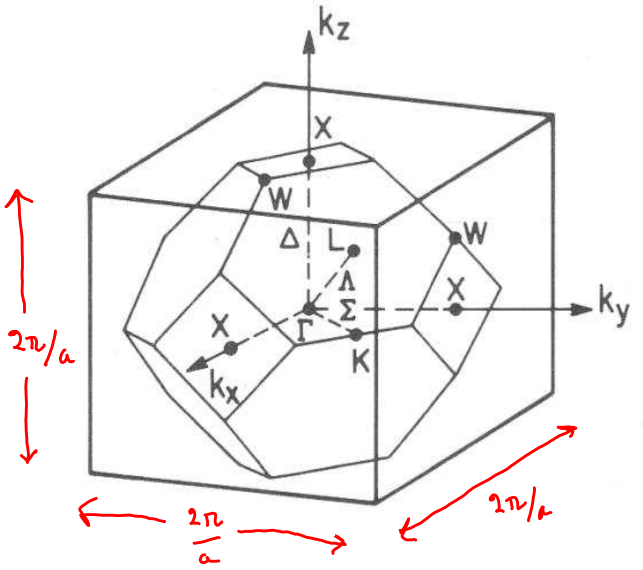


Figure 9.6: First Brillouin zone of the fcc lattice. Note that it is the same shape as the Wigner-Seitz cell of bcc lattice, see Fig. 8.13.

The construction of the Brillouin zone is analogous in 3d to the 2d case, and is again entirely analogous to the construction of the Wigner-Seitz cell in 3d. For a simple cubic lattice, the first Brillouin zone is simply a cube. For fcc and bcc lattices, however, the situation is more complicated. As we mentioned above in the Aside at the end of section 9.1.3 above, the reciprocal lattice of the fcc lattice is bcc, and vice-versa. Thus, the Brillouin zone of the fcc lattice is the same shape as the the Wigner-Seitz cell of the bcc lattice! The Brillouin zone for the fcc lattice is shown in Fig. 9.6 (compare to Fig. 8.13). Note that in Fig. 9.6, various k-points are labeled with letters. There is a complicated labeling convention that we will not discuss in this course, but it is worth knowing that it exists. For example, we can see in the figure that the point $\mathbf{k} = \mathbf{0}$ is labeled Γ, and the point $\mathbf{k} = (\pi/a)\hat{\mathbf{y}}$ is labeled X.

Given this diagram of this Brillouin zone we can finally arrive at some real physics!

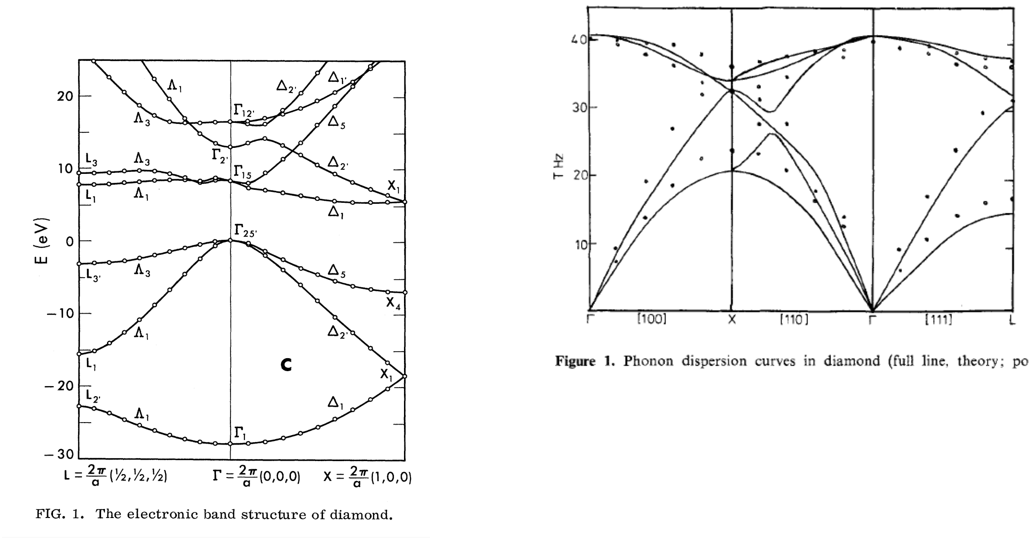


FIG. 1. The electronic band structure of diamond.

Figure 1. Phonon dispersion curves in diamond (full line, theory; points,

Figure 9.7: Left: Electronic excitation spectrum of diamond Right: Phonon spectrum of diamond (points are from experiment). The horizontal axis gives cuts through k-space as labeled in Fig. 9.6 above. (Left figure is from W. Saslow, T. K. Bergstresser, and Marvin L. Cohen Phys. Rev. Lett. 16, 354 (1966). Right figure is from R. D. Turner and J. C. Inkson, J. Phys. C: Solid State Phys., Vol. 11, 1978))

9.3 Electronic and Vibrational Waves in Crystals in 3d

In the left of Fig. 9.7 we show the electronic band-structure (i.e., dispersion relation) of diamond, which is an fcc lattice with a diatomic basis (See Fig. 8.16). As in the one-dimensional case, we can choose to work in the reduced zone scheme where we only need to consider the first Brillouin zone. Since we are trying to display a three dimensional spectrum (Energy as a function of $\mathbf{k}$) on a one dimensional diagram, what is done is to show several single-line cuts through reciprocal space[13]. Starting on the left of the diagram, we start at L-point of the Brillouin zone and show $E(\mathbf{k})$ as $\mathbf{k}$ traces a straight line to the Γ point (the center of the Brillouin zone). Then we continue to the right and $\mathbf{k}$ traces a straight line from the Γ point to the X point. Note that the lowest band is quadratic at the center of the Brillouin zone (a dispersion $\hbar^2 k^2/(2m^*)$ for some effective mass m^*).

Similarly, in the right of Fig. 9.7, we show the phonon spectrum of diamond. Several things to note about this figure. First of all, since diamond has a unit cell with 2 atoms in it (it is fcc with a basis of two atoms) there should be 6 modes of oscillation per k-points. Indeed, this is what we see in the picture, at least in the central third of the picture. In the other two parts of the picture, one sees fewer modes per k-point, but this is because, due to the symmetry of the crystal along this particular direction, several excitation modes have exactly the same energy (you can see

---

[12] Here's the proof of this statement for a square lattice. Let the system be N_x by N_y unit cells. Then, with periodic boundary conditions, the value of k_x is quantized in units of $2\pi/L_x = 2\pi/(N_x a)$ and the value of k_y is quantized in units of $2\pi/L_y = 2\pi/(N_y a)$. But the size of the Brillouin zone is $2\pi/a$ in each direction, thus there are precisely $N_x N_y$ different values of $\mathbf{k}$ in the Brillouin zone.

[13] This type of plot, because it can look like a jumble of lines, is sometimes called a "spaghetti-diagram"

for example, at the X-point, two modes come in from the right, but only one goes out to the left. This means the two modes have the same energy on the left of the X point). Secondly, we note that at the Γ-point, $\mathbf{k} = 0$ there are exactly three modes which come down linearly to zero energy. These are the three acoustic modes. The other three modes, which are finite energy at $\mathbf{k} = 0$ are the optical modes. Finally, you may note something a bit confusing about this diagram. On the far left of the diagram, we start at the Γ point, move in the (100) direction and end up at the X point. Then from the X point, we move in the (110) direction, and we end up back at the Γ point! This is because we have landed at the Γ point in a different Brillouin zone.

9.4 Summary

- Reciprocal lattice is a lattice in k-space defined by the set of points such that $e^{i\mathbf{G}\cdot\mathbf{R}} = 1$ for all $\mathbf{R}$ in direct lattice. Given this definition, the reciprocal lattice can be thought of as the Fourier transform of the direct lattice.
- A reciprocal lattice vector $\mathbf{G}$ defines a set of parallel equally spaced planes via $\mathbf{G}\cdot\mathbf{r} = 2\pi m$ such that every point of the direct lattice is included in one of the planes. The spacing between the planes is $d = 2\pi/|\mathbf{G}|$. If $\mathbf{G}$ is the smallest reciprocal lattice vector parallel to $\mathbf{G}$ then this set of planes is a family of lattice planes, meaning that all planes intersect points of the direct lattice.
- Miller Indices (h, k, l) are used to describe families of lattice planes, or reciprocal lattice vectors. For fcc and bcc lattices, one specifies the Miller indices of the associated simple cubic lattice conventional unit cell.
- General definition of Brillouin zone is any unit cell in reciprocal space. The First Brillouin zone is the Wigner-Seitz cell around the point $\mathbf{0}$ of the reciprocal lattice. Each Brillouin zone has the same volume – and contains one k-state per unit cell of the entire system. Parallel Brillouin zone boundaries are separated by reciprocal lattice vectors.

References

For reciprocal lattice, Miller indices and Brillouin zones. I recommend

- Ashcroft and Mermin, chapter 5 (again be warned of the nomenclature issue mentioned above in chapter 8, footnote 2).

Many books introduce X-ray diffraction and the reciprocal lattice at the same time. Once we have read the next chapter and we study scattering, we might go back and look at the nice introductions to reciprocal space given in the following books.

- Goodstein, section 3.4-3.5 (very brief)
- Kittel, chapter 2
- Ibach and Luth, chapter 3

Part V

Neutron and X-Ray Diffraction

Chapter 10

Wave Scattering by Crystals

In the last chapter we discussed reciprocal space, and explained that the energy dispersion of phonons and electrons is plotted within the Brillouin zone. We understand how these are similar to each other due to the wave-like nature of both the electron and the phonon. However, much of the same physics occurs when crystals scatter waves (or particles[1]) that impinge upon a crystal externally. Indeed, exposing a solid to a wave in order to probe its properties is an extremely useful thing to do. The most common probe to use are X-rays. Another common, more modern, probe to use is neutrons. In fact it can hardly be overstated how important this type of experiment is to science.

The general setup that we will examine is shown in Fig.10.1.

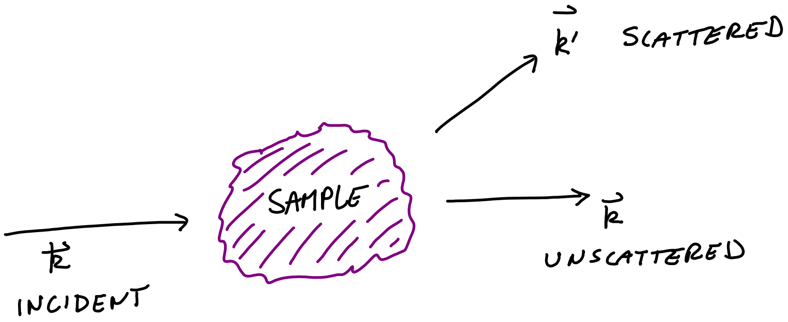


Figure 10.1: A generic scattering experiment.

---

[1] Remember, in quantum mechanics there is no real difference between particles and waves!

10.1 The Laue and Bragg Conditions

10.1.1 Fermi's Golden Rule Approach

If we think of the incoming wave as being a particle, then we should think of the sample as being some potential $V(\mathbf{r})$ that the particle experiences as it goes through the sample. According to Fermi's golden rule[2], the transition rate $\Gamma(\mathbf{k}', \mathbf{k})$ per unit time for the particle scattering from $\mathbf{k}$ to $\mathbf{k}'$ is given by

$$\Gamma(\mathbf{k}', \mathbf{k}) = \frac{2\pi}{\hbar} \left|\langle \mathbf{k}'|V|\mathbf{k}\rangle\right|^2 \delta(E_{\mathbf{k}'} - E_{\mathbf{k}})$$

The matrix element here

$$\langle \mathbf{k}'|V|\mathbf{k}\rangle = \int \mathbf{dr}\, \frac{e^{-i\mathbf{k}'\cdot\mathbf{r}}}{\sqrt{L^3}}\, V(\mathbf{r})\, \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{\sqrt{L^3}} = \frac{1}{L^3}\int \mathbf{dr}\, e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{r}}\, V(\mathbf{r})$$

is nothing more than the fourier transform of the potential (where L is the linear size of the sample, so the $\sqrt{L^3}$ terms just normalize the wavefunctions).

Note that these above expressions are true whether or not the sample is a periodic crystal. However, if the sample is periodic the matrix element is zero unless $\mathbf{k} - \mathbf{k}'$ is a reciprocal lattice vector! To see this is true, let us write positions $\mathbf{r} = \mathbf{R} + \mathbf{x}$ where $\mathbf{R}$ is a lattice vector position and $\mathbf{x}$ is a position within the unit cell

$$\langle \mathbf{k}'|V|\mathbf{k}\rangle = \frac{1}{L^3}\int \mathbf{dr}\, e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{r}}\, V(\mathbf{r}) = \frac{1}{L^3}\sum_{\mathbf{R}}\int \mathbf{dx}\, e^{i(\mathbf{k}'-\mathbf{k})\cdot(\mathbf{x}+\mathbf{R})}\, V(\mathbf{x}+\mathbf{R})$$

Now since the potential is assumed periodic, we have $V(\mathbf{x}+\mathbf{R}) = V(\mathbf{x})$, so this can be rewritten as

$$\langle \mathbf{k}'|V|\mathbf{k}\rangle = \frac{1}{L^3}\left[\sum_{\mathbf{R}} e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}}\right]\left[\int_{unit-cell} \mathbf{dx}\;\; e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{x}}\; V(\mathbf{x})\right] \tag{10.1}$$

Now, as we discussed in section 9.1.3 above, the first term in brackets must vanish unless $\mathbf{k}' - \mathbf{k}$ is a reciprocal lattice vector[3]. This condition,

$$\mathbf{k}' - \mathbf{k} = \mathbf{G} \tag{10.2}$$

is known as the *Laue equation* (or *Laue condition*)[4] This condition, is precisely the statement of *the conservation of crystal momentum.*[5] Note also that when the wave leaves the crystal, they

---

[2] Fermi's golden rule should be familiar to you from quantum mechanics. Interestingly, Fermi's golden rule was actually discovered by Dirac, giving us yet another example where something is named after Fermi when Dirac really should have credit as well, or even instead. See also footnote 11 in section 3.2.1.

[3] Also we discussed that this first term in brackets diverges if $\mathbf{k}' - \mathbf{k}$ is a reciprocal lattice vector. This divergence is not a problem here because it gives just the number of unit cells and is canceled by the $1/L^3$ normalization factor leaving a factor of the inverse volume of the unit cell.

[4] Max von Laue won the Nobel prize for his work on X-ray scattering from crystals in 1914. Although von Laue never left Germany during the second world war, he remained openly opposed to the Nazi government. During the war he hid his gold Nobel medal at the Niels Bohr Institute in Denmark to prevent the Nazi's from taking it. Had he been caught doing this, he may have been jailed or worse, since shipping gold out of Nazi Germany was considered a serious offense. After the occupation of Denmark in April 1940, George de Hevesy (a Nobel Laureate in chemistry) decided to dissolve the medal in the solvent Aqua regia to remove the evidence. He left the solution on a shelf in his lab. Although the Nazis occupied Bohr's institute and searched it very carefully, they did not find anything. After the war, the gold was recovered from solution and the Nobel Foundation presented Laue with a new medal made from the same gold.

[5] Real momentum is conserved since the crystal itself absorbs any missing momentum. In this case, the center of mass of the crystal has absorbed momentum $\hbar(\mathbf{k}' - \mathbf{k})$. See the comment in footnote 9 in section 5.2.4.

should have

$$|\mathbf{k}| = |\mathbf{k}'|$$

which is just the conservation of energy, which is enforced by the delta function in Fermi's golden rule. (In section 10.4.2 below we will consider more complicated scattering where energy is not conserved.)

10.1.2 Diffraction Approach

It turns out that this Laue condition is nothing more than the scattering condition associated with a diffraction grating. This description of the scattering from crystals is known as the Bragg Formulation of (x-ray) diffraction[6].

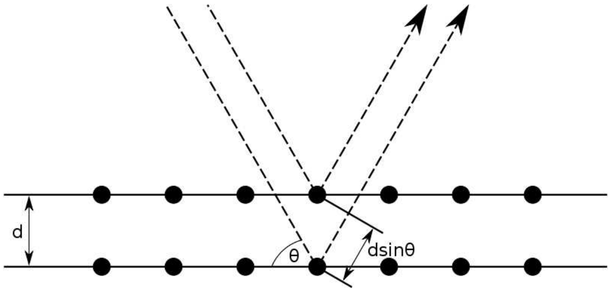


Figure 10.2: Bragg Scattering off of a plane of atoms in a crystal.

Consider the configuration shown in Fig. 10.2. An incoming wave is reflected off of two adjacent layers of atoms separated by a distance d. A few things to note about this diagram. First note that the wave has been deflected by 2θ in this diagram[7] Secondly, from simple geometry note that the additional distance traveled by the component of the wave that reflects off of the further layer of atoms is

$$\text{extra distance} = 2d \sin\theta.$$

In order to have constructive interference, this extra distance must be equal to an integer number of wavelengths. Thus we derive the Bragg condition for constructive interference, or what is known as *Bragg's law*

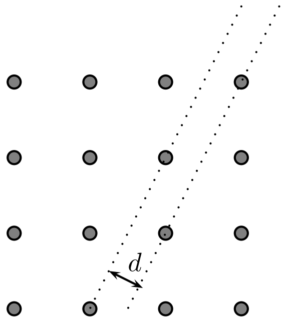
$$n\lambda = 2d \sin\theta \qquad (10.3)$$

Note that we can have diffraction from any two parallel planes of atoms such as the one shown here

---

[6] William Henry Bragg and William Lawrence Bragg were a father and son team who won the Nobel prize together in 1915 for the their work on X-ray scattering. William Lawrence Bragg was 25 years old when he won the prize, and remains the youngest Nobel laureate ever.

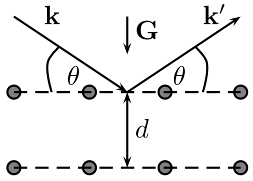
[7] This is a *very* common source of errors on exams. The *total* deflection angle is 2θ.



What we will see next is that this Bragg condition for constructive interference is precisely equivalent to the Laue condition described above.

10.1.3 Equivalence of Laue and Bragg conditions

Consider the following picture (essentially the same as Fig.10.2, I suppose). Here we have shown the reciprocal lattice vector **G** which corresponds to the family of lattice planes. As we discussed in chapter 9 the spacing between lattice planes is $d = 2\pi/|\mathbf{G}|$ (See Eqn. 9.8).



.

Just from geometry we have

$$\hat{\mathbf{k}} \cdot \hat{\mathbf{G}} = \sin\theta = -\hat{\mathbf{k}'} \cdot \hat{\mathbf{G}}$$

where the hats ˆ over vectors indicate unit vectors.

Suppose the Laue condition is satisfied. That is, $\mathbf{k} - \mathbf{k}' = \mathbf{G}$ with $|\mathbf{k}| = |\mathbf{k}'| = 2\pi/\lambda$ with λ the wavelength. We can rewrite the Laue equation as

$$\frac{2\pi}{\lambda}(\hat{\mathbf{k}} - \hat{\mathbf{k}'}) = \mathbf{G}$$

Now let us dot this equation with $\hat{\mathbf{G}}$ to give

$$\begin{aligned}\hat{\mathbf{G}}\cdot\frac{2\pi}{\lambda}(\hat{\mathbf{k}}-\hat{\mathbf{k}}') &= \hat{\mathbf{G}}\cdot\mathbf{G}\\ \frac{2\pi}{\lambda}(\sin\theta-\sin\theta') &= |\mathbf{G}|\\ \frac{2\pi}{|\mathbf{G}|}(2\sin\theta) &= \lambda\\ 2d\sin\theta &= \lambda\end{aligned}$$

which is the Bragg condition (in the last step we have used the relation between $\mathbf{G}$ and d). You may wonder why in this equation we got λ on the right hand side rather than $n\lambda$ as we had in Eq. 10.3. The point here is that there if there is a reciprocal lattice vector $\mathbf{G}$, then there is also a reciprocal lattice vector $n\mathbf{G}$, and if we did the same calculation with that lattice vector we would get $n\lambda$. In other words, in the $n\lambda$ case we are reflecting off of the spacing nd which necessarily also exists when there is a set of lattice planes with spacing d.

Thus we conclude that the Laue condition and the Bragg condition are equivalent. It is equivalent to say that interference is constructive (as Bragg indicates) or to say that crystal momentum is conserved (as Laue indicates).

10.2 Scattering Amplitudes

If the Laue condition is satisfied, we would now like to ask how much scattering we actually get. Recall in section 10.1.1 we started with Fermi's golden rule

$$\Gamma(\mathbf{k}',\mathbf{k}) = \frac{2\pi}{\hbar}\left|\langle\mathbf{k}'|V|\mathbf{k}\rangle\right|^2\delta(E_{\mathbf{k}'}-E_{\mathbf{k}})$$

and we found out that if V is a periodic function, then the matrix element is given by (See Eq. 10.1)

$$\langle\mathbf{k}'|V|\mathbf{k}\rangle = \left[\frac{1}{L^3}\sum_{\mathbf{R}} e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}}\right]\left[\int_{unit-cell}\mathbf{dx}\; e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{x}}\;V(\mathbf{x})\right] \tag{10.4}$$

The first factor in brackets gives zero unless the Laue condition is satisfied, in which case it gives a constant (due to the $1/L^3$ out front, this is now a nondivergent constant). The second term in brackets is known as the structure factor (compare to Eq. 9.6)

$$S(\mathbf{G}) = \int_{unit-cell}\mathbf{dx}\; e^{i\mathbf{G}\cdot\mathbf{x}}\;V(\mathbf{x}) \tag{10.5}$$

where we have used $\mathbf{G}$ for $(\mathbf{k}'-\mathbf{k})$ since this must be a reciprocal lattice vector or the first term in brackets vanishes.

Frequently, one writes

$$I_{(hkl)} \propto |S_{(hkl)}|^2 \tag{10.6}$$

which is shorthand for saying that $I_{(hkl)}$, the intensity of scattering off of the lattice planes defined by the reciprocal lattice vector (hkl), is proportional to the square of the structure factor at this reciprocal lattice vector. Sometimes a delta-function is also written explicitly to indicate that the wavevector difference $(\mathbf{k}'-\mathbf{k})$ must be a reciprocal lattice vector.

We now turn to examine this structure factor more closely for our main two types of scattering probes – neutrons[8] and x-rays.

Neutrons

Since neutrons are uncharged, they scatter almost exclusively from nuclei (rather than electrons) via the nuclear forces. As a result, the scattering potential is extremely short ranged, and can be approximated as a delta-function. We thus have

$$V(\mathbf{x}) = \sum_{\text{atom } j \text{ in unit cell}} f_j \; \delta(\mathbf{x} - \mathbf{x_j})$$

where $\mathbf{x_j}$ is the position of the j^{th} atom in the unit cell. Here, f_j is known as the *form factor* or *atomic form factor*, and represents the strength of scattering from that particular nucleus. In fact, for the case of neutrons this quantity is proportional to the so-called "nuclear scattering-length" b_j. Thus for neutrons we frequently write

$$V(\mathbf{x}) \sim \sum_{\text{atom } j \text{ in unit cell}} b_j \; \delta(\mathbf{x} - \mathbf{x_j})$$

Plugging this expression into Eq. 10.5 above, we obtain

$$S(\mathbf{G}) \sim \sum_{\text{atom } j \text{ in unit cell}} b_j \; e^{i(\mathbf{G})\cdot \mathbf{x_j}} \tag{10.7}$$

X-rays

X-rays scatter from the electrons in a system[9] As a result, one can take $V(\mathbf{x})$ to be proportional to the electron density. We can thus approximate

$$V(\mathbf{x}) \sim \sum_{\text{atom } j \text{ in unit cell}} Z_j \; g_j(\mathbf{x} - \mathbf{x_j})$$

where Z_j is the atomic number of atom j (i.e., its number of electrons) and g_j is a somewhat short-ranged function (i.e., it has a few angstroms range — roughly the size of an atom). Taking the Fourier transform, we obtain

$$S(\mathbf{G}) \sim \sum_{\text{atom } j \text{ in unit cell}} f_j(\mathbf{G}) \; e^{i(\mathbf{G})\cdot \mathbf{x_j}} \tag{10.8}$$

where $f_j \sim Z_j$, the form factor has some dependence on the magnitude of the reciprocal lattice vector $\mathbf{G}$. Frequently, however, we approximate f_j to be independent of $\mathbf{G}$ (which would be true if g were extremely short ranged), although this is not strictly correct.

Aside: As noted above, $f_j(\mathbf{G})$ is just the Fourier transform of the scattering potential for atom j. This scattering potential is proportional to the electron density. Taking the density to be a delta function results in

---

[8] Brockhouse and Schull were awarded the Nobel prize in 1994 for pioneering the use of neutron scattering experiments for understanding properties of crystals. Schull's development of this technique occurred around 1946 – making this the longest time-lag ever between a discovery and the awarding of the Nobel prize.

[9] The coupling of photons to matter is via the usual minimal coupling $(\mathbf{p} - e\mathbf{A})^2/(2m)$. The denominator m is why the nuclei are not important.

f_j being a constant. Taking the slightly less crude approximation that the density is constant inside a sphere of radius r_0 and zero outside of this radius will result in a Fourier transform

$$f_j(\mathbf{G}) \sim \frac{Z \sin(|\mathbf{G}| r_0)}{|\mathbf{G}| r_0} \tag{10.9}$$

(try showing this!). If the scattering angle is sufficiently small (i.e., $\mathbf{G}$ is small compared to $1/r_0$), this function is roughly just Z with no strong dependence on $\mathbf{G}$.

Comparison of Neutrons and X-rays

- For X-rays since $f_j \sim Z_j$ the x-rays scatter very strongly from heavy atoms, and hardly at all from light atoms. This makes it very difficult to "see" light atoms like hydrogen in a solid. Further it is hard to distinguish atoms that are very close to each other in their atomic number (since they scatter almost the same amount). Also f_j is slightly dependent on the scattering angle.
- In comparison the nuclear scattering length b_j varies rather erratically with atomic number (it can even be negative). In particular, hydrogen scatters fairly well, so it is easy to see. Further, one can usually distinguish atoms with similar atomic numbers rather easily.
- For neutrons, the scattering really is very short ranged, so the form factor really is proportional to the scattering length b_j independent of $\mathbf{G}$. For X-rays there is a dependence of $\mathbf{G}$ that complicates matters.
- Neutrons also have spin. Because of this they can be spin-polarized using a Stern-Gerlach-type setup[10]. One can then detect whether various electrons in the unit cell have their spins pointing up or down. We will return to this situations where the spin of the electron is spatially ordered in section 16.1.2 below.

Simple Example

Generally, as mentioned above, we write the intensity of scattering as

$$I_{(hkl)} \propto |S_{(hkl)}|^2$$

Assuming we have orthogonal primitive basis vectors, we can then generally write

$$S_{(hkl)} = \sum_{\text{atom } j \text{ in unit cell}} f_j \; e^{2\pi i (h x_j + k y_j + l z_j)} \tag{10.10}$$

where $[x_j, y_j, z_j]$ are the coordinates of atom j within the unit cell, in units of the three primitive basis vectors.

Example 1: Caesium Chloride: Let us now consider the simple example of CsCl, whose unit cell is shown in Fig. 10.3. This system can be described as simple cubic with a basis given by.[11]

---

[10] More Nobel history: Stern won the Nobel prize in 1943 without Gerlach. Stern had fled Germany before the war while Gerlach was suspected of contributing to the Nazi bomb effort.

[11] Do not make the mistake of calling this a lattice! Bcc is a lattice where all points must be the same.

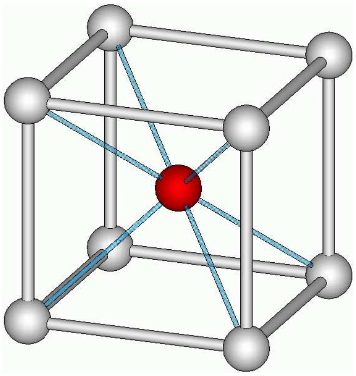


Figure 10.3: Cesium Chloride Unit Cell. Cs is white corner atoms, Cl is red central atom. This is simple cubic with a basis. Note that bcc Cs can be thought of as just replacing the Cl with another Cs atom.

Basis for CsCl		
Cs	Position=	$[0,0,0]$
Cl	Position=	$[a/2,a/2,a/2]$

Thus the structure factor is given by

$$\begin{aligned} S_{(hkl)} &= f_{Cs} + f_{Cl}\, e^{2\pi i (h,k,l)\cdot[1/2,1/2,1/2]} \\ &= f_{Cs} + f_{Cl}(-1)^{h+k+l} \end{aligned}$$

with the f's being the appropriate form factors for the corresponding atoms. Recall that the scattered wave intensity is $I_{(hkl)} \sim |S_{(hkl)}|^2$.

10.2.1 Systematic Absences and More Examples

Example 2: Caesium bcc: Let us now consider instead a pure Cs crystal. In this case the crystal is bcc. We can think of this as simply replacing the Cl in CsCl with another Cs atom. Analogously we think of the bcc lattice as a simple cubic lattice with exactly the same basis, which we now write as

Basis for Cs bcc		
Cs	Position=	$[0,0,0]$
Cs	Position=	$[a/2,a/2,a/2]$

Now the structure factor is given by

$$\begin{aligned} S_{(hkl)} &= f_{Cs} + f_{Cs}\, e^{2\pi i (h,k,l)\cdot[1/2,1/2,1/2]} \\ &= f_{Cs}\left[1 + (-1)^{h+k+l}\right] \end{aligned}$$

Crucially, note that the structure factor, and therefore the scattering intensity vanishes for $h+k+l$ being any odd integer! This phenomenon is known as a *systematic absence.*

To understand why this absence occurs, consider the simple case of the (100) family of planes (See Fig. 9.1). This is simply a family of planes along the crystal axes with spacing a. You might expect a wave of wavelength $2\pi/a$ oriented perpendicular to these planes to scatter constructively. However, if we are considering a bcc lattice, then there are additional planes of atoms half-way between the (100) planes which then cause perfect *destructive* interference. We refer back to the Important Complication mentioned in section 9.1.5. As mentioned there, the plane spacing for the bcc lattice in this case is not $2\pi/|\mathbf{G}_{(100)}|$ but is rather $2\pi/|\mathbf{G}_{(200)}|$. In fact in general, for a bcc lattice the plane spacing for any family of lattice planes is $2\pi/|\mathbf{G}_{(hkl)}|$ where $h+k+l$ is always even. This is what causes the selection rule.

Example 3: Copper fcc Quite similarly there are systematic absences in for scattering from fcc crystals as well. Recall from Eq. 8.4 that the fcc crystal can be thought of as a simple cubic lattice with a basis given by the points $[0,0,0], [1/2,1/2,0], [1/2,0,1/2]$, and $[0,1/2,1/2]$ in units of the cubic lattice constant. As a result the structure factor of fcc coppper is given by (plugging into Eq. 10.10)

$$S_{(hkl)} = f_{Cu}\left[1 + e^{i\pi(h+k)} + e^{i\pi(h+l)} + e^{i\pi(k+l)}\right] \tag{10.11}$$

It is easily shown that this expression vanishes unless h, k and l are either all odd or all even.

Summary of Systematic Absences

Systematic Absences of Scattering	
Simple Cubic	all h, k, l allowed
bcc	$h+k+l$ must be even
fcc	h, k, l must be all odd or all even

Systematic absences are sometimes known as *selection rules.*

It is very important to note that these absences, or selection rules, occur for any structure with the given Bravais lattice type. Even if a material is bcc with a basis of five different atoms per primitive unit cell, it will still show the same systematic absences as the bcc lattice we considered above with a single atom per primitive unit cell. To see why this is true we consider yet another example

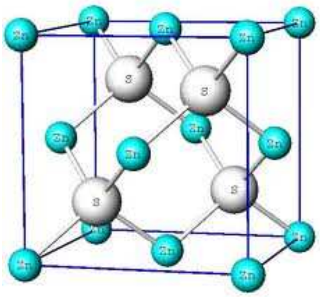


Figure 10.4: Zinc Sulfide conventional unit cell. This is fcc with a basis given by a Zn atom at $[0,0,0]$ and a S atom at $[1/4,1/4,1/4]$.

Example 4: Zinc Sulfide = fcc with a basis: As shown in Fig 10.4, the Zinc Sulfide crystal is a an fcc lattice with a basis given by a Zn atom at [0, 0, 0] and an S atom at [1/4, 1/4, 1/4] (this is known as a zincblende structure). If we consider the fcc lattice to itself be a cubic lattice with basis given by the points [0, 0, 0], [1/2, 1/2, 0], [1/2, 0, 1/2], and [0, 1/2, 1/2], we then have the 8 atoms in the conventional unit cell having positions given by the combination of the two bases, i.e.,

Basis for ZnS						
Zn	Positions=	[0, 0, 0],	[1/2, 1/2, 0],	[1/2, 0, 1/2],	and	[0, 1/2, 1/2]
S	Positions=	[1/4, 1/4, 1/4],	[3/4, 3/4, 1/4],	[3/4, 1/4, 3/4],	and	[1/4, 3/4, 3/4]

The structure factor for ZnS is thus given by

$$S_{(hkl)} = f_{Zn}\left[1 + e^{2\pi i (hkl)\cdot[1/2,1/2,1/0]} + \ldots\right] + f_S\left[e^{2\pi i (hkl)\cdot[1/4,1/4,1/4]} + e^{2\pi i (hkl)\cdot[3/4,3/4,1/4]} + \ldots\right]$$

This combination of 8 terms can be factored to give

$$S_{(hkl)} = \left[1 + e^{i\pi(h+k)} + e^{i\pi(h+l)} + e^{i\pi(k+l)}\right]\left[f_{Zn} + f_S\, e^{i(\pi/2)(h+k+l)}\right] \tag{10.12}$$

The first term in brackets is precisely the same as the term we found for the fcc crystal in Eq. 10.11. In particular it has the same systematic absences that it vanishes unless h, k and l are either all even or all odd. The second term gives additional absences associated specifically with the ZnS structure.

Since the positions of the atoms are the positions of the underlying lattice plus the vectors in the basis, it is easy to see that the structure factor of a crystal system with a basis will always factorize into a piece which comes from the underlying lattice structure times a piece corresponding to the basis. Generalizing Eq. 10.12 we can write

$$S_{(hkl)} = S^{Lattice}_{(hkl)} \times S^{basis}_{(hkl)} \tag{10.13}$$

(where to be precise the form factors only occur in the latter term).

10.3 Methods of Scattering Experiments

There are many methods of performing scattering experiments. In principle they are all similar — one sends in a probe wave of known wavelength (an X-ray, for example) and measures the angles at which it diffracts when it comes out. Then using Bragg's laws (or the Laue equation) one can deduce the spacings of the lattice planes in the system.

10.3.1 Methods that are interesting and useful that you probably won't be tested on

Laue Method

Conceptually, perhaps the simplest method is to take a large single crystal of the material in question — fire waves at it (X-rays, say) at it from one direction, and measure the direction of the outgoing waves. However, given a single direction of incoming wave, it is unlikely that you precisely achieve the diffraction condition (the Bragg condition) for any given set of lattice planes. In order to get more data, one can then vary the wavelength of the incoming wave. This allows one to achieve the Bragg condition, at least at some wavelength.

Rotating Crystal Method

A similar technique is to rotate the crystal continuously so that at some angle of the crystal with respect to the incoming waves, one achieves the Bragg condition and measures an outcoming diffracted wave.

Both of these methods are used. However, there is an important reason that they are sometimes impossible. Frequently it is not possible to obtain a single crystal of a material. Growing large crystals (such as the beautiful ones shown in Fig. 4.9) can be an enormous challenge[12] In the case of neutron scattering the problem is even more acute since one typically needs a fairly large single crystals compared to x-rays.

10.3.2 Powder Diffraction: you will almost certainly be tested on this!

Powder diffraction, or the *Debye-Scherrer method*[13] is the use of wave scattering on a sample which is not single crystalline, but is powdered. In this case, the incoming wave can scatter off of any one of many small crystallites which may be oriented in any possible direction. In spirit this technique is similar to the rotating crystal method in that there is always some angle at which a crystal can be oriented to diffract the incoming wave. A figure of the Debye-Scherrer setup is shown in Fig. 10.5. Using Bragg's law, given the wavelength of the incoming wave, we can deduce the possible spacings between lattice planes.

A Fully Worked Example. Study this!

Because this type of problem has historically ended up on exams essentially every year, and because it is hard to find references that explain how to solve these problems, I am going to work a power-diffraction problem in detail here. As far as I can tell, they will only ever ask you about cubic lattices (simple cubic, fcc, and bcc).

Before presenting the problem and solving it, however, it is useful to write down a table of possible lattice planes and the selection rules that can occur for the smallest reciprocal lattice vectors

---

[12] For example, high-temperature superconducting materials were discovered in 1986 (and resulted in a Nobel prize the next year!). Despite a concerted world-wide effort, good single crystals of these materials were not available for 5 to 10 years.

[13] Debye is the same guy from the specific heat of solids. Paul Scherrer was Swiss but worked in Germany during the second world war, where he passed information to the famous American spy (and baseball player), Moe Berg, who had been given orders to find and shoot Heisenberg if he felt that the Germans were close to developing a bomb.

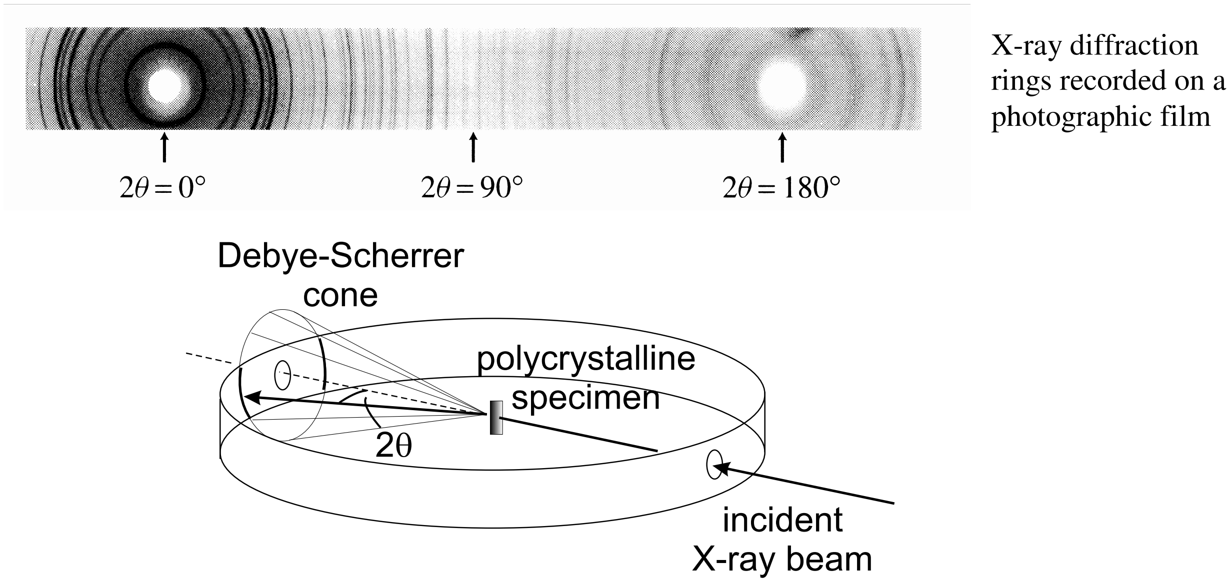


Figure 10.5: Debye-Scherrer Powder Diffraction.

Lattice Plane Selection Rules

$\{hkl\}$	$N = h^2 + k^2 + l^2$	Multiplicity	cubic	bcc	fcc
100	1	6	✓		
110	2	12	✓	✓	
111	3	8	✓		✓
200	4	6	✓	✓	✓
210	5	24	✓		
211	6	24	✓	✓	
220	8	12	✓	✓	✓
221	9	24	✓		
300	9	6	✓		
310	10	24	✓	✓	
311	11	24	✓		✓
222	12	8	✓	✓	✓
⋮	⋮				

The selection rules are exactly those listed above: simple cubic allows scattering from any plane, bcc must have $h + k + l$ be even, and fcc must have h, k, l either all odd or all even.

We have added a column N which is the square magnitude of the reciprocal lattice vector.

We have also added an addition column labeled "multiplicity". This quantity is important for figuring out the amplitude of scattering. The point here is that the (100) planes have some particular spacing but there are 5 other families of planes with the same spacing: $(010), (001), (\bar{1}00), (0\bar{1},0), (00\bar{1})$. (Because we mean all of these possible families of lattice planes, we use the notation $\{hkl\}$ introduced at the end of section 9.1.5). In the powder diffraction method, the crystal orientations are random, and here there would be 6 possible equivalent orientations of a crystal which will present the right angle for scattering off of one of these planes, so there will be scattering intensity which is 6 times as large as we would otherwise calculate — this is known as the multiplicity factor. For the case of the 111 family, we would instead find 8 possible equivalent planes: $(111), (11\bar{1}), (1\bar{1}1), (1\bar{1}\bar{1}), (\bar{1}11), (\bar{1}1\bar{1}), (\bar{1}\bar{1}1), (\bar{1}\bar{1}\bar{1})$. Thus, we should replace Eq. 10.6 with

$$I_{\{hkl\}} \propto M_{\{hkl\}} |S_{\{hkl\}}|^2 \qquad (10.14)$$

where M is the multiplicity factor.

Note: calculating the actual intensity of scattering works very well for neutron scattering, but is much harder for x-ray scattering because the form factor for X-rays depends on $\mathbf{G}$. I.e, since in Eq. 10.7 the form factor (or scattering length b_j) is a constant independent of $\mathbf{G}$, it is easy to calculate the expected amplitudes of scattering based only on these constants. For the case of X-rays you need to know the functional forms of $f_j(\mathbf{G})$. At some *very* crude level of approximation it is a constant. More precisely we see in Eq. 10.9 that it is constant for small scattering angle but can vary quite a bit for large scattering angle.

The Example

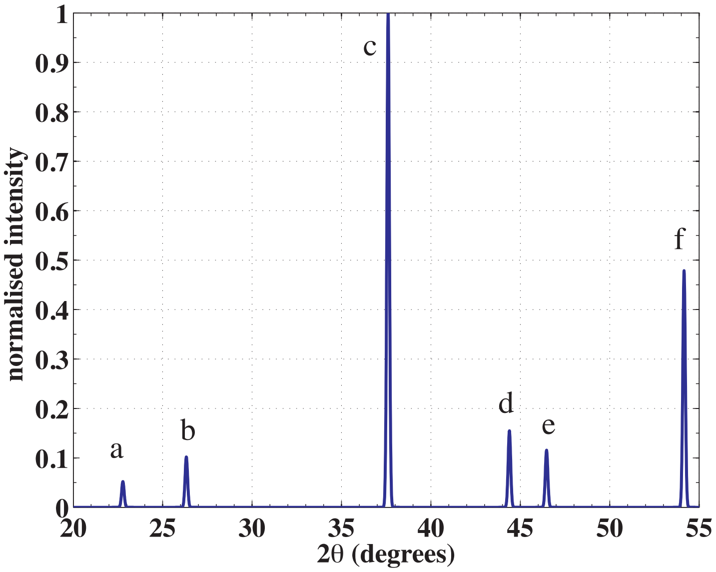


Figure 10.6: Powder Diffraction of Neutrons from PrO_2. The wavelength of the neutron beam is $\lambda = .123$ nm

Consider the powder diffraction data from PrO_2 shown in Fig. 10.6. (Exactly this data was presented in the 2009 Exam, and we were told that the lattice is some type of cubic lattice). Given the wavelength .123 nm, we first would like to figure out the type of lattice and the lattice constant.

Note that the full deflection angle is 2θ. We will want to use Bragg's law and the expression for the spacing between planes

$$d_{(hkl)} = \frac{\lambda}{2\sin\theta} = \frac{a}{\sqrt{h^2+k^2+l^2}}$$

where we have also used the expression Eq. 9.12 for the spacing between planes in a cubic lattice given the lattice constant a. Note that this then gives us

$$a^2/d^2 = h^2+k^2+l^2 = N$$

which is what we have labeled N in the above table of selection rules. We now make a table. In the first two columns we just read the angles off of the given graph *Note, you should try to make the measurements of the angle from the data as carefully as possible. It makes the analysis much easier if you measure the angles right!*

peak	2θ	$d = \lambda/(2\sin\theta)$	d_a^2/d^2	$3d_a^2/d^2$	$N = h^2+k^2+l^2$	$\{hkl\}$	$a = d\sqrt{h^2+k^2+l^2}$
a	22.7°	0.313 nm	1	3	3	111	.542 nm
b	26.3°	0.270 nm	1.33	3.99	4	200	.540 nm
c	37.7°	0.190 nm	2.69	8.07	8	220	.537 nm
d	44.3°	0.163 nm	3.67	11.01	11	311	.541 nm
e	46.2°	0.157 nm	3.97	11.91	12	222	.544 nm
f	54.2°	0.135 nm	5.35	16.05	16	400	.540 nm

In the third column of the table we calculate the distance between lattice planes for the given diffraction peak using Bragg's law. In the third fourth column we have calculated the squared ratio of the lattice spacing d for the given peak to the lattice spacing for the first peak (labeled a) as a reference. We then realize that these ratios are pretty close to whole numbers divided by three, so we try multiplying each of these quantities by 3 in the next column. If we round these numbers to integers (given in the next column), we produce precisely the values of $N = h^2+k^2+l^2$ expected for the fcc lattice (According to the above selection rules we must have h, k, l all even or all odd). The final column calculates the lattice constant from the given diffraction angle. Averaging these numbers gives us a measurement of the lattice constant $a = .541 \pm .002$ nm.

The analysis thus far is equivalent to what one would do for X-ray scattering. However, with neutrons, assuming the scattering length is independent of scattering angle (which is typically a good assumption) we can go a bit further and analyze the the intensity of the scattering peaks.

In the problem given on the 2009 Exam, it is given that the basis for this crystal is a Pr atom at position [0,0,0] and O at [1/4,1/4,1/4] and [1/4,1/4,3/4]. Thus, the Pr atoms form a fcc lattice and the O's fill in the holes as shown in Fig. 10.7.

Let us calculate the structure factor for this crystal. Using Eq. 10.13 we have

$$S_{(hkl)} = \left[1 + e^{i\pi(h+k)} + e^{i\pi(h+l)} + e^{i\pi(k+l)}\right]\left[b_{Pr} + b_O\left(e^{i(\pi/2)(h+k+l)} + e^{i(\pi/2)(h+k+3l)}\right)\right]$$

The first term in brackets is the structure factor for the fcc lattice, and it gives 4 for every allowed scattering point (when h, k, l are either all even or all odd). The second term in brackets is the structure factor for the basis.

The scattering intensity of the peaks are then given in terms of this structure factor and the peak multiplicities as shown in Eq. 10.14. We thus can write for all of our measured peaks

$$I_{\{hkl\}} = CM_{\{hkl\}}\left|b_{Pr} + b_O\left(e^{i(\pi/2)(h+k+l)} + e^{i(\pi/2)(h+k+3l)}\right)\right|^2$$

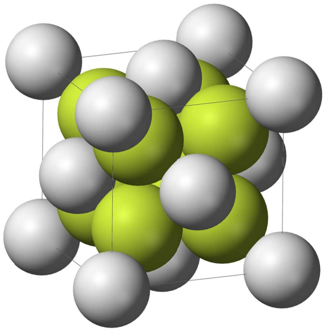


Figure 10.7: The flourite structure. This is fcc with a basis given by a white atom (Pr) atom at $[0,0,0]$ and yellow atoms at $[1/4,1/4,1/4]$ and $[1/4,1/4,3/4]$.

where the constant C contains other constant factors (including the factor of 4^2 from the fcc structure factor). Note: We have to be a bit careful here to make sure that the bracketed factor gives the same result for all possible (hkl) included in $\{hkl\}$, but in fact it does. Thus we can compile another table showing the predicted relative intensities of the peaks.

Scattering Intensity			
peak	$\{hkl\}$	$I_{\{hkl\}}/C \propto M\|S\|^2$	Measured Intensity
a	111	$8\, b_{Pr}^2$	0.05
b	200	$6\, [b_{Pr} - 2b_O]^2$	0.1
c	220	$12\, [b_{Pr} + 2b_O]^2$	1.0
d	311	$24\, b_{Pr}^2$	0.15
e	222	$8\, [b_{Pr} - 2b_O]^2$	0.1
f	400	$6\, [b_{Pr} + 2b_O]^2$	0.5

where the final column are the intensities measured from the data in Fig. 10.6.

From the analytic expressions in the third column we can immediately predict that we should have

$$I_d = 3I_a \qquad I_c = 2I_f \qquad I_e = \frac{4}{3} I_b$$

Examining the fourth column of this table, it is clear that the first two of these equations are properly satisfied. However the final equation does not appear to be correct. This points to some error in constructing the plot. Thus we suspect some problem in either I_e or I_b. Either I_e is too small or I_b is too large[14].

To further home in on this problem with the data, we can look at the ratio I_c/I_a which in the measured data has a value of about 20. Thus we have

$$\frac{I_c}{I_a} = \frac{12[b_{Pr} + 2b_O]^2}{8\, b_{Pr}^2} = 20$$

---

[14] Another possibility is that the form factor is not precisely independent of scattering angle, as is the case for X-ray scattering. However, the fact that all the peaks are consistent but for this one peak suggests a transcription error.

with some algebra this can be reduced to a quadratic equation with two roots, resulting in

$$b_{Pr} = -.43 b_O \qquad \text{or } .75 b_O \tag{10.15}$$

Let us suppose now that our measurement of I_b is correct. In this case we have

$$\frac{I_b}{I_a} = \frac{6[b_{Pr} - 2b_O]^2}{8\, b_{Pr}^2} = 2$$

which we can solve to give

$$b_{Pr} = .76 b_O \qquad \text{or } \; -3.1 b_O$$

The former solution being reasonably consistent with the above. However, were we to assume I_e were correct, we would have instead

$$\frac{I_e}{I_a} = \frac{8[b_{Pr} - 2b_O]^2}{8\, b_{Pr}^2} = 2$$

we would obtain

$$b_{Pr} = .83 b_O \qquad \text{or } \; -4.8 b_O$$

which appears inconsistent with Eq. 10.15. We thus conclude that the measured intensity of I_e given in Fig. 10.6 is actually incorrect, and should be larger by about a factor of 4/3.

10.4 Still more about scattering

Scattering experiments such as those discussed here are *the* method for determining the microscopic structures of materials. One can use these methods (and extensions thereof) to sort out even very complicated atomic structures such as those of biological molecules.

Aside: In addition to the obvious work of von Laue and Bragg that initiated the field of X-ray diffraction (and Brockhouse and Schull for neutrons) there have been about half a dozen Nobel prizes that have relied on, or further developed these techniques. In 1962 a chemistry Nobel prize was awarded to Perutz and Kendrew for using X-rays to determine the structure of the biological proteins hemoglobin and myoglobin. The same year, Watson and Crick were awarded the prize in Biology for determining the structure of DNA — which they did with the help of X-ray diffraction data taken by Rosalind Franklin[15] Two years later in 1964, Dorothy Hodgkin[16] won the prize for determination of the structure of penicillin and other biological molecules. Further Nobels were given in chemistry for determining the structure of Boranes (Lipscomb, 1976) and for the structure of photosynthetic proteins (Deisenhofer, Huber, Michel 1988).

10.4.1 Variant: Scattering in Liquids and Amorphous Solids

A material need not be crystalline to scatter waves. However, for amorphous solids or liquids, instead of having delta-function peaks in the structure factor at reciprocal lattice vectors (as in Fig. 10.6), the structure factors (which is again defined as the fourier transform of the density) will have smooth behavior — with incipient peaks corresponding to $2\pi/d$ where d is roughly the

---

[15] There remains quite a controversy over the fact that Watson and Crick, at a critical juncture, were shown Franklin's data without her knowledge! Franklin may have won the prize in addition to Watson and Crick and thereby received a bit more of the appropriate credit, but she tragically died of cancer at age 37 in 1958, two years before the prize was awarded.

[16] Dorothy Hodgkin was a student and later a fellow at **Somerville College, Oxford. Yay!**

typical distance between atoms. An example of a measured structure factor in liquid Al is shown in Fig. 10.8. As the material gets close to its freezing point, the peaks in the structure factor will get more pronounced, becoming more like the structure of a solid where the peaks are delta functions.

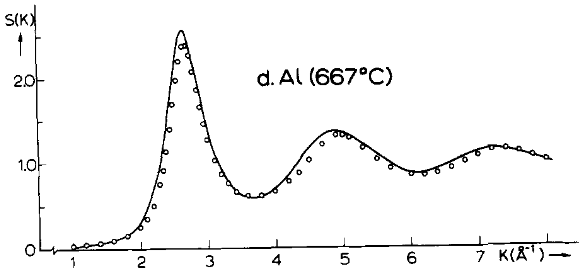


Figure 10.8: The structure factor of liquid Aluminum

10.4.2 Variant: Inelastic Scattering

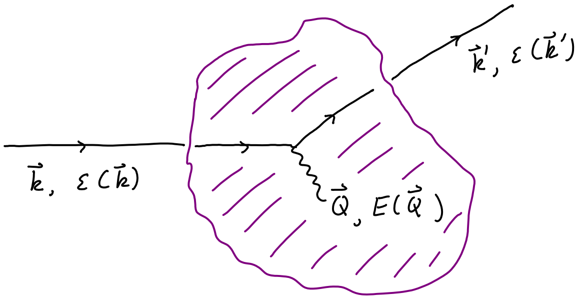


Figure 10.9: Inelastic scattering. Energy and crystal momentum must be conserved.

It is also possible to perform scattering experiments which are inelastic. Here, "inelastic" means that energy of the incoming wave is left behind in the sample, and the energy of the outgoing wave is lower. The general process is shown in Fig. 10.9. A wave is incident on the crystal with momentum $\mathbf{k}$ and energy $\epsilon(\mathbf{k})$ (For neutrons the energy would be $\hbar^2\mathbf{k}^2/(2m)$ whereas for photons the energy would be $\hbar c|\mathbf{k}|$). This wave transfers some of its energy and momentum to some internal excitation mode of the material — such as a phonon, or a spin or electronic excitation quanta. One then measures the outgoing energy and momentum of the wave. Since energy and crystal momentum must be conserved, one has

$$\begin{aligned}\mathbf{Q} &= \mathbf{k}' - \mathbf{k} + \mathbf{G} \\ E(\mathbf{Q}) &= \epsilon(\mathbf{k}') - \epsilon(\mathbf{k})\end{aligned}$$

thus allowing one to determine the dispersion relation of the internal excitation (i.e., the relationship between $\mathbf{Q}$ and $E(\mathbf{Q})$). This technique is extremely useful for determining phonon dispersions experimentally. In practice, the technique is much more useful with neutrons than with X-rays. The reason for this is, because the speed of light is so large, (and $E = \hbar c|\mathbf{k}|$) the energy differences that one obtains are enormous except for a tiny small range of $\mathbf{k}'$ for each $\mathbf{k}$. Since there is a maximum energy for a phonon, the X-rays therefore have a tiny total cross section for exciting phonons. A second reason that this technique is difficult for X-rays is because it is much harder to build an X-ray detector that determines the X-ray energy than for neutrons.

10.4.3 Experimental Apparatus

Perhaps the most interesting piece of this kind of experiments is the question of how one actually produces and measures the waves in questions.

Since at the end of the day, one ends up counting photons or neutrons, brighter sources (higher flux of probe particles) are always better —- as it allows one to do experiments quicker, and allows one to reduce noise (since counting error on N counts is proportional to $\sqrt{N}$). Further, with a brighter source, one can examine smaller samples more easily.

X-rays: Even small laboratories can have X-ray sources that can do very useful crystallography. A typical source accelerates electrons electrically (with 10s of keV) and smashes them into a metal target. X-rays with a discrete spectrum of energies are produced when an electron is knocked out of a low atomic orbital and an electron in a higher orbital drops down to re-fill the hole (this is known as X-ray flourescence). Also a continuous Bremsstrahlung spectrum is produced by electrons coming near the charged nuclei, but for monochromatic diffraction experiments, this is less useful. (One wavelength from a spectrum can be selected — using diffraction from a known crystal!).

Much higher brightness X-ray sources are provided by huge (and hugely expensive) facilities known as synchrotron light sources – where particles (usually electrons) are accelerated around enormous loops (at energies in the GeV range). Then using magnets these electrons are rapidly accelerated around corners which makes them emit X-rays extremely brightly and in a highly columnated fashion.

Detection of X-rays can be done with photographic films (the old style) but is now more frequently done with more sensitive semiconductors detectors.

Neutrons: Although it is possible to generate neutrons in a small lab, the flux of these devices is extremely small and neutron scattering experiments are always done in large neutron source facilities. Although the first neutron sources simply used the byproduct neutrons from nuclear reactors, more modern facilities now use a technique called *spallation* where protons are accelerated into a target and neutrons are emitted. As with X-rays, neutrons can be monochromated (made into a single wavelength), by diffracting them from a known crystal. Another technique is to use time-of-flight. Since more energetic neutrons move faster, one can send a pulse of poly-chromatic neutrons and select only those that arrive at a certain time in order to obtain mono-chromatic neutrons. On the detection side, one can again select for energy very easily. I won't say too much about neutron detection as there are many methods. Needless to say, they all involve interaction with nuclei.

Figure 10.10: The Rutherford-Appleton Lab in Oxfordshire, UK. On the right, the large circular building is the DIAMOND synchrotron light source. The building on the left is the ISIS spallation neutron facility. This was the world's brightest neutron source on earth until August 2007 when it was surpassed by one in Oak Ridge, US.

10.5 Summary of Diffraction

- Diffraction of waves from crystals in Laue and Bragg formulations (equivalent to each other).

- The structure factor (the fourier transform of the scattering potential) in a periodic crystal has sharp peaks at allowed reciprocal lattice vectors for scattering. The scattering intensity is the square of the structure factor.
- There are systematic absences of diffraction peaks depending on the crystal structure (fcc, bcc). Know how to figure these out.
- Know how to analyze a powder diffraction pattern (very common exam question!)

References

It is hard to find references that give enough information about diffraction to suit the Oxford course. These are not bad.

- Kittel, chapter 2
- Ashcroft and Mermin, chapter 6
- Dove, chapter 6 (most detailed, with perhaps a bit too much information in places)

In addition, the following have nice, but incomplete discussions.

- Rosenberg, chapter 2.
- Ibach and Luth, chapter 3.
- Burns, chapter 4.

Part VI

Electrons in Solids

Chapter 11

Electrons in a Periodic Potential

In chapters 5 and 6 we discussed the wave nature of phonons in solids, and how crystal momentum is conserved (i.e., momentum is conserved up to reciprocal lattice vector). Further we found that we could describe the entire excitation spectrum within a single Brillouin zone in a reduced zone scheme). We also found in chapter 10 that X-rays and neutrons similarly scatter from solids by conserving crystal momentum. In this chapter we will consider the nature of electron waves in solids and we will find that similarly crystal momentum is conserved and the entire excitation spectrum can be described within a single Brillouin zone using a reduced zone scheme.

We have seen a detailed preview of properties of electrons in periodic systems when we considered the one dimensional tight binding model in chapter 7, so the results of this section will be hardly surprising. However, in the current chapter we will approach the problem from a very different (and complimentary) starting point. Here, we will consider electrons as free-electron waves that are very only very weakly perturbed by the periodic arrangement of atoms in the solid. The tight binding model is exactly the opposite limit where we consider electrons bound strongly to the atoms, and they only weakly hop from one atom to the next.

11.1 Nearly Free Electron Model

We start with completely free electrons whose hamiltonian is

$$H_0 = \frac{\mathbf{p}^2}{2m}$$

with corresponding energy eigenstates, the plane waves $|\mathbf{k}\rangle$, with eigenenergies

$$\epsilon_0(\mathbf{k}) = \frac{\hbar^2|\mathbf{k}|^2}{2m}$$

We now consider a weak periodic potential perturbation to this Hamiltonian

$$H = H_0 + V(\mathbf{r})$$

with

$$V(\mathbf{r}) = V(\mathbf{r} + \mathbf{R})$$

where **R** is any lattice vector. The matrix elements of this potential are then just the Fourier components

$$\langle \mathbf{k}'|V|\mathbf{k}\rangle = \frac{1}{L^3}\int \mathbf{dr}\, e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{r}}\; V(\mathbf{r}) \equiv V_{\mathbf{k}'-\mathbf{k}} \tag{11.1}$$

which is zero unless $\mathbf{k}' - \mathbf{k}$ is a reciprocal lattice vector (See Eq. 10.1). Thus, any plane wave state $\mathbf{k}$ can scatter into another plane wave state $\mathbf{k}'$ only if these two plane waves are separated by a reciprocal lattice vector.

We now apply the rules of perturbation theory. At first order in the perturbation V, we have

$$\epsilon(\mathbf{k}) = \epsilon_0(\mathbf{k}) + \langle \mathbf{k}|V|\mathbf{k}\rangle = \epsilon_0(\mathbf{k}) + V_0$$

which is just an uninteresting constant energy shift to all of the eigenstates.

At second order in perturbation theory we have

$$\epsilon(\mathbf{k}) = \epsilon_0(\mathbf{k}) + V_0 + {\sum_{\mathbf{k}'=\mathbf{k}+\mathbf{G}}}' \frac{|\langle \mathbf{k}'|V|\mathbf{k}\rangle|^2}{\epsilon_0(\mathbf{k}) - \epsilon_0(\mathbf{k}')} \tag{11.2}$$

where the $'$ means that the sum is restricted to have $\mathbf{G} \neq \mathbf{0}$. In this sum, however, we have to be careful. It is possible that $\epsilon_0(\mathbf{k})$ is very close to $\epsilon_0(\mathbf{k}')$ or even that they are equal, in which case the term of the sum diverges and the perturbation expansion makes no sense. This is what we call a *degenerate* situation and it needs to be handled with degenerate perturbation theory, which we shall consider below.

To see when this degenerate situation happens, we look for solutions of

$$\begin{aligned} \epsilon_0(\mathbf{k}) &= \epsilon_0(\mathbf{k}') \quad (11.3)\\ \mathbf{k}' &= \mathbf{k} + \mathbf{G} \quad (11.4) \end{aligned}$$

First, let us consider the one-dimensional case. Since $\epsilon(k) \sim k^2$, the only possible solutions of

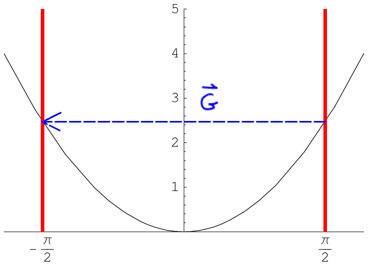


Figure 11.1: Scattering from Brillouin zone boundary to Brillouin zone boundary (separated by a reciprocal lattice vector **G**) at the same energy. This situation leads to a divergence in perturbation theory, Eq. 11.2 because when the two energies match, the denominator is zero.

Eq. 11.3 is $k' = -k$. This means the two equations are only satisfied for

$$k' = -k = \pi n/a$$

or precisely on the Brillouin zone boundaries (See Fig. 11.1).

In fact, this is quite general even in higher dimensions: given a point $\mathbf{k}$ on a Brillouin zone boundary, there is another point $\mathbf{k}'$ (also on a Brillouin zone) boundary such that Eqs. 11.3 and 11.4 are satisfied (See in particular Fig. 9.5 for example)[1].

Since Eq. 11.2 is divergent, we need to handle this situation with *degenerate perturbation theory*[2]. In this approach, one diagonalizes the Hamiltonian within the degenerate space first (and other perturbations can be treated after this). In other words, we take states of the same energy that are connected by the matrix element and treat their mixing exactly.

11.1.1 Degenerate Perturbation Theory

If two plane wave states $|\mathbf{k}\rangle$ and $|\mathbf{k}'\rangle = |\mathbf{k}+\mathbf{G}\rangle$ are of approximately the same energy (meaning that $\mathbf{k}$ and $\mathbf{k}'$ are close to zone boundaries), then we must diagonalize the matrix elements of these states first. We have

$$\begin{array}{rcl} \langle\mathbf{k}|\, H \,|\mathbf{k}\rangle & = & \epsilon_0(\mathbf{k}) \\ \langle\mathbf{k}'|\, H \,|\mathbf{k}'\rangle & = & \epsilon_0(\mathbf{k}') = \epsilon_0(\mathbf{k}+\mathbf{G}) \\ \langle\mathbf{k}|\, H \,|\mathbf{k}'\rangle & = & V_{\mathbf{k}'-\mathbf{k}} = V_{\mathbf{G}} \\ \langle\mathbf{k}'|\, H \,|\mathbf{k}\rangle & = & V_{\mathbf{k}-\mathbf{k}'} = V_{\mathbf{G}}^* \end{array} \tag{11.5}$$

where we have used the definition of $V_{\mathbf{G}}$ from Eq. 11.1, and the fact that $V_{-\mathbf{G}} = V_{\mathbf{G}}^*$ is guaranteed by the fact that $V(\mathbf{r})$ is real.

Now, within this two dimensional space we can write any wavefunction as

$$|\Psi\rangle = \alpha|\mathbf{k}\rangle + \beta|\mathbf{k}'\rangle = \alpha|\mathbf{k}\rangle + \beta|\mathbf{k}+\mathbf{G}\rangle \tag{11.6}$$

Using the variational principle to minimize the energy is equivalent to solving the effective Schroedinger equation

$$\begin{pmatrix} \epsilon_0(\mathbf{k}) & V_{\mathbf{G}} \\ V_{\mathbf{G}}^* & \epsilon_0(\mathbf{k}+\mathbf{G}) \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = E \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{11.7}$$

The secular equation determining E is then

$$\left(\epsilon_0(\mathbf{k}) - E\right)\left(\epsilon_0(\mathbf{k}+\mathbf{G}) - E\right) - |V_{\mathbf{G}}|^2 = 0 \tag{11.8}$$

(Note that once this degenerate space is diagonalized, one can go back and treat further, nondegenerate, scattering processes, in perturbation theory.)

Simple Case: k exactly at the zone boundary

The simplest case we can consider is when $\mathbf{k}$ is precisely on a zone boundary (and therefore $\mathbf{k}' = \mathbf{k}+\mathbf{G}$ is also precisely on a zone boundary). In this case $\epsilon_0(\mathbf{k}) = \epsilon_0(\mathbf{k}+\mathbf{G})$ and our secular equation simplifies to

$$\left(\epsilon_0(\mathbf{k}) - E\right)^2 = |V_{\mathbf{G}}|^2$$

---

[1] To see this generally, recall that a Brillouin zone boundary is a perpendicular bisector of the segment between $\mathbf{0}$ and some $\mathbf{G}$. We can write the given point $\mathbf{k} = \mathbf{G}/2 + \mathbf{k}_\perp$ where $\mathbf{k}_\perp \cdot \mathbf{G} = 0$. Then if we construct the point $\mathbf{k}' = -\mathbf{G}/\mathbf{2} + \mathbf{k}_\perp$, then clearly 11.4 is satisfied, $\mathbf{k}'$ is a perpendicular bisector of the segment between $\mathbf{0}$ and $-\mathbf{G}$ and therefore is on a zone boundary, and $|\mathbf{k}| = |\mathbf{k}'|$ which implies that Eq. 11.3 is satisfied.

[2] Hopefully you have learned this in your quantum mechanics courses already!

or equivalently

$$E_\pm = \epsilon_0(\mathbf{k}) \pm |V_\mathbf{G}|$$

Thus we see that a gap opens up at the zone boundary. Whereas both $\mathbf{k}$ and $\mathbf{k}'$ had energy $\epsilon_0(\mathbf{k})$ in the absence of the added potential $V_\mathbf{G}$, when the potential is added, the two eigenstates form two linear combinations with energies split by $\pm|V_\mathbf{G}|$.

In one dimension

In order to understand this better, let us focus on the one dimensional case. Let us assume we have a potential $V(x) = \tilde{V}\cos(2\pi x/a)$ with $\tilde{V} > 0$. The Brillouin zone boundaries are at $k = \pi/a$ and $k' = -k = -\pi/a$ so that $k' - k = G = -2\pi/a$ and $\epsilon_0(k) = \epsilon_0(k')$.

Examining Eq. 11.7, we discover that the solutions (when $\epsilon_0(k) = \epsilon_0(k')$) are given by $\alpha = \pm\beta$ thus giving the eigenstates

$$|\psi_\pm\rangle = \frac{1}{\sqrt{2}}\left(|k\rangle \pm |-k\rangle\right) \tag{11.9}$$

corresponding to $E_\pm$ respectively. Since we can write the real space version of these $|k\rangle$ wavefunctions as[3]

$$\begin{aligned} |k\rangle &\rightarrow e^{ikx} = e^{ix\pi/a} \\ |-k\rangle &\rightarrow e^{-ikx} = e^{-ix\pi/a} \end{aligned}$$

we discover that the two eigenstates are given by

$$\begin{aligned} \psi_+ &\sim e^{ix\pi/a} + e^{-ix\pi/a} &\propto \cos(x\pi/a) \\ \psi_- &\sim e^{ix\pi/a} - e^{-ix\pi/a} &\propto \sin(x\pi/a) \end{aligned}$$

If we then look at the densities $|\psi_\pm|^2$ associated with these two wavefunctions (See fig. 11.2) we see that the higher energy eigenstate $\psi+$ has its density concentrated mainly at the maxima of the potential V whereas the lower energy eigenstate ψ_- has it density concentrated mainly at the minima of the potential.

So the general principle is that the periodic potential scatters between the two plane waves $\mathbf{k}$ and $\mathbf{k}+\mathbf{G}$. If the energy of these two plane waves are the same, the mixing between them is strong, and the two plane waves can combine to form one state with higher energy (concentrated on the potential maxima) and one state with lower energy (concentrated on the potential minima).

k not quite on a zone boundary (and still in 1d)

It is not too hard to extend this calculation to the case where $\mathbf{k}$ is not quite on a zone boundary. For simplicity though we will stick to the one dimensional situation[4]. We need only solve the secular equation 11.8 for more general k. To do this, we expand around the zone boundaries.

Let us consider the states at the zone boundary $k = \pm n\pi/a$ which are separated by the reciprocal lattice vectors $G = \pm 2\pi n/a$. As noted above, the gap that opens up precisely at the zone boundary will be $\pm|V_G|$. Now let us consider a plane wave near this zone boundary $k = n\pi/a + \delta$

---

[3] Formally what we mean here is $\langle x|k\rangle = e^{ikx}/\sqrt{L}$.

[4] If you are very brave and good with geometry, you can try working out the 3d case.

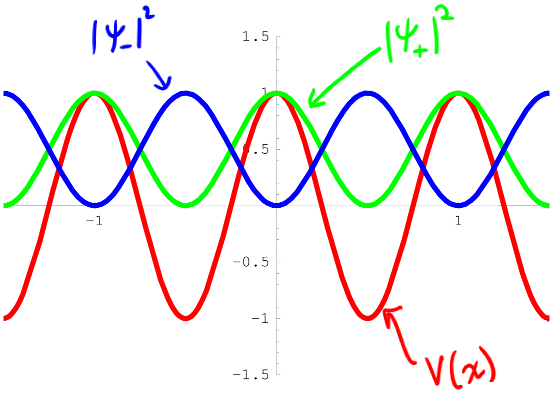


Figure 11.2: The higher energy eigenstate ψ_+ has its density concentrated near the maxima of the potential V whereas the lower energy eigenstate has its density concentrated near the minima.

with δ being very small (and n an integer). This wavevector can scatter into $k' = -n\pi/a + \delta$ due to the periodic potential. We then have

$$\begin{aligned}\epsilon_0(n\pi/a+\delta) &= \frac{\hbar^2}{2m}\left[(n\pi/a)^2 + 2n\pi\delta/a + \delta^2\right] \\ \epsilon_0(-n\pi/a+\delta) &= \frac{\hbar^2}{2m}\left[(n\pi/a)^2 - 2n\pi\delta/a + \delta^2\right]\end{aligned}$$

The secular equation (Eq. 11.8)) is then

$$\left(\frac{\hbar^2}{2m}\left[(n\pi/a)^2+\delta^2\right] - E + \frac{\hbar^2}{2m}2n\pi\delta/a\right)\left(\frac{\hbar^2}{2m}\left[(n\pi/a)^2+\delta^2\right] - E - \frac{\hbar^2}{2m}2n\pi\delta/a\right) - |V_{\mathbf{G}}|^2 = 0$$

which simplifies to

$$\left(\frac{\hbar^2}{2m}\left[(n\pi/a)^2+\delta^2\right] - E\right)^2 = \left(\frac{\hbar^2}{2m}2n\pi\delta/a\right)^2 + |V_{\mathbf{G}}|^2$$

or

$$E_\pm = \frac{\hbar^2}{2m}\left[(n\pi/a)^2+\delta^2\right] \pm \sqrt{\left(\frac{\hbar^2}{2m}2n\pi\delta/a\right)^2 + |V_{\mathbf{G}}|^2}$$

Expanding the square root for small δ we obtain

$$E_\pm = \frac{\hbar^2(n\pi/a)^2}{2m} \pm |V_G| + \frac{\hbar^2\delta^2}{2m}\left[1 \pm \frac{\hbar^2(n\pi/a)^2}{m}\frac{1}{|V_G|}\right]$$

Thus we see that near the band gap at the Brillouin zone boundary, the dispersion is quadratic as shown in the left of Fig. 11.3. In the right of this figure, we see (using the *repeated zone scheme*) that small gaps open at the Brillouin zone boundaries in what is otherwise a parabolic spectrum. (This plotting scheme is equivalent to the reduced zone scheme if restricted to a single zone).

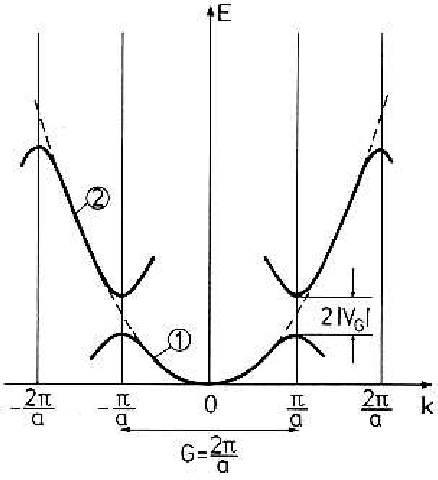


Figure 11.3: In the nearly free electron model, gaps open up at the Brillouin zone boundaries in an otherwise parabolic spectrum. Compare this to what we found for the tight binding model in Fig 7.5.

The general structure we find is thus very much like what we expected from the tight binding model we considered previously in chapter 7 above. As in the tight binding picture there are *energy bands* where there are energy eigenstates, and there are gaps between bands, where there are no energy eigenstates. As in the tight binding model, the spectrum is periodic in the Brillouin zone (See Fig 11.4).

Nearly free electrons in two (and higher) dimensions

The principles of the nearly free electron model are quite similar in 2d. In short, near the Brillouin zone boundary, a gap opens up due to scattering by a reciprocal lattice vector. States of energy slightly higher than the zone boundary intersection point are pushed up in energy, whereas states of energy slightly lower than the zone boundary intersection point are pushed down in energy. We will return to the detailed geometry of this situation in section 12.2.

There is one more key difference between 1d and higher dimensions. In 1d, we found that if $\mathbf{k}$ is on a zone boundary, then there will be exactly one other $\mathbf{k}'$ such that $\mathbf{k} - \mathbf{k}' = \mathbf{G}$ is a reciprocal lattice vector and such that $\epsilon(\mathbf{k}') = \epsilon(\mathbf{k})$. (I.e., Eqs. 11.3 and 11.4 are satisfied). As described above, these two plane wave states mix with each other (See Eq. 11.6) and open up a gap. However, in higher dimensions it may occur that given $\mathbf{k}$ there may be *several* different $\mathbf{k}'$ which will satisfy these equations — i.e., many $\mathbf{k}'$ which differ from $\mathbf{k}$ by a reciprocal lattice vector and which all have the same unperturbed energy. In this case, we need to mix together all of the possible plane waves in order to discover the true eigenstates. One example of when this occurs is the two dimensional square lattice, where the four points $(\pm\pi/a, \pm\pi/a)$ all have the same unperturbed energy and are all separated from each other by reciprocal lattice vectors.

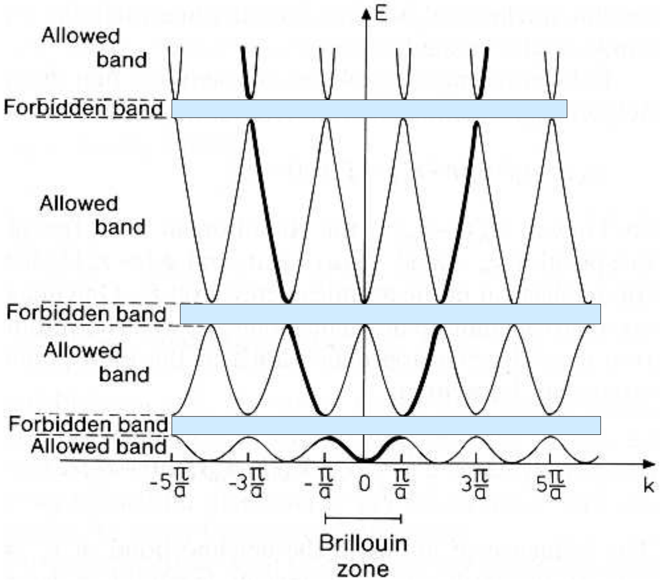


Figure 11.4: Same as Fig. 11.3 above, but plotted in repeated zone scheme. This is equivalent to the reduced zone scheme but the equivalent zones are repeated. Forbidden bands are marked where there are no eigenstates. The similarity to the free electron parabolic spectrum is emphasized.

11.2 Bloch's Theorem

In the above, "nearly free electron" approach, we started from the perspective of plane waves that are weakly perturbed by a periodic potential. But in real materials, the scattering from atoms can be very strong so that perturbation theory may not be valid (or may not converge until very high order). How do we know that we can still describe electrons with anything remotely similar to plane waves?

In fact, by this time, after our previous experience with waves, we should know the answer in advance: the plane wave momentum is not a conserved quantity, but the crystal momentum is. No matter how strong the periodic potential, so long as it is periodic, crystal momentum is conserved. This important fact was first discovered by Felix Bloch[5] in 1928, very shortly after the discovery of the Schroedinger equation, in what has become known as Bloch's theorem[6]

> **Bloch's Theorem:** An electron in a periodic potential has eigenstates of the form
>
> $$\Psi^{\alpha}_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u^{\alpha}_{\mathbf{k}}(\mathbf{r})$$
>
> where $u^{\alpha}_{\mathbf{k}}$ is periodic in the unit cell and $\mathbf{k}$ (the crystal momentum) can be chosen within the first Brillouin zone.

---

[5] Felix Bloch later won a Nobel prize for inventing Nuclear Magnetic Resonance. NMR was then renamed MRI when people decided the word "Nuclear" sounds too much like it must be related to some sort of bomb.

[6] Bloch's theorem was actually discovered by a Mathematician Gaston Floquet in 1883, and rediscovered later by Bloch in the context of solids. This is an example of what is known as Stigler's Law of Eponomy: "Most things are not named after the person who first discovers them". In fact, Stigler's law was discovered by Merton.

In reduced zone scheme there may be many states at each $\mathbf{k}$ and these are indexed by α. The periodic function u, usually known as a *Bloch function*, and Ψ is sometimes known as a *modified plane-wave*. Because u is periodic, it can be rewritten as a sum over reciprocal lattice vectors

$$u^{\alpha}_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} \tilde{u}^{\alpha}_{\mathbf{G},\mathbf{k}} \; e^{i\mathbf{G}\cdot\mathbf{r}}$$

This form guarantees[7] that $u^{\alpha}_{\mathbf{k}}(\mathbf{r}) = u^{\alpha}_{\mathbf{k}}(\mathbf{r}+\mathbf{R})$ for any lattice vector $\mathbf{R}$. Therefore the full wavefunction is expressed as

$$\Psi^{\alpha}_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} \tilde{u}^{\alpha}_{\mathbf{G},\mathbf{k}} \; e^{i(\mathbf{G}+\mathbf{k})\cdot\mathbf{r}} \tag{11.10}$$

Thus an equivalent statement of Bloch's theorem is that we can write each eigenstate as being made up of a sum of plane wave states $\mathbf{k}$ which differ by reciprocal lattice vectors $\mathbf{G}$.

Given this equivalent statement of Bloch's theorem, we now understand that the reason for Bloch's theorem is that the scattering matrix elements $\langle \mathbf{k}'|V|\mathbf{k}\rangle$ are zero unless $\mathbf{k}'$ and $\mathbf{k}$ differ by a reciprocal lattice vector. As a result, the Schroedinger equation is "block diagonal"[8] in the space of $\mathbf{k}$ and in any given wavefunction only plane waves $\mathbf{k}$ that differ by some $\mathbf{G}$ can be mixed together. One way to see this more clearly is to is to take the Schroedinger equation

$$\left[\frac{\mathbf{p}^2}{2m} + V(\mathbf{r})\right] \Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

and Fourier transform it to obtain

$$\sum_{\mathbf{G}} V_{\mathbf{G}} \Psi_{\mathbf{k}-\mathbf{G}} = \left[E - \frac{\hbar^2|\mathbf{k}|^2}{2m}\right] \Psi_{\mathbf{k}}$$

where we have used the fact that $V_{\mathbf{k}-\mathbf{k}'}$ is only nonzero if $\mathbf{k}-\mathbf{k}' = \mathbf{G}$. It is then clear that for each $\mathbf{k}$ we have a Schroedinger equation for the set of $\Psi_{\mathbf{k}-\mathbf{G}}$'s and we must obtain solutions of the form of Eq. 11.10.

Although by this time it may not be surprising that electrons in a periodic potential have eigenstates labeled by crystal momenta, we should not overlook how important Bloch's theorem is. This theorem tells us that even though the potential that the electron feels from each atom is extremely strong, the electrons still behave *almost* as if they do not see the atoms at all! They still *almost* form plane wave eigenstates, with the only modification being the periodic Bloch function u and the fact that momentum is now crystal momentum.

A quote from Felix Bloch:

> 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. By straight Fourier analysis I found to my delight that the wave differed from the plane wave of free electrons only by a periodic modulation.

11.3 Summary of Electrons in a Periodic Potential

- When electrons are exposed to a periodic potential, gaps arise in their dispersion relation at the Brillouin zone boundary. (The dispersion is quadratic approaching a zone boundary).

---

[7] In fact, the function u is periodic in the unit cell if and only if it can be written as a sum over reciprocal lattice vectors in this way.

[8] No pun intended.

- Thus the electronic spectrum breaks into bands, with forbidden energy gaps between the bands. In the nearly free electron model, the gaps are proportional to the periodic potential $|V_{\mathbf{G}}|$.
- Bloch's theorem guarantees us that all eigenstates are some periodic function times a plane wave. In repeated zone scheme the wavevector (the *crystal momentum*) can always be taken in the first Brillouin zone.

References

- Goodstein, Section 3.6a.
- Burns, section 10.1-10.6
- Kittel, chapter 7 (Skip Kronig-Penny model)
- Hook and Hall, section 4.1
- Ashcroft and Mermin, chapters 8-9 (not my favorite)
- Ibach and Luth, sections 7.1-7.2
- Singleton, chapter 2-3

Chapter 12

Insulator, Semiconductor, or Metal

In chapter 7, when we discussed the tight-binding model in 1d, we introduced some of the basic ideas of band structure. In chapter 11 we found that an electron in a periodic potential shows exactly the same type of band-structure as we found for the tight-binding model: In both cases, we found that the spectrum is periodic in momentum (so all momenta can be taken to be in the first Brillouin zone, in reduced zone scheme) and we find that gaps open at Brillouin zone boundaries. These principles, the idea of bands and band structure form the fundamental underpinning of our understanding of electrons in solids. In this chapter (and the next) we explore these ideas in further depth.

12.1 Energy Bands in One Dimension: Mostly Review

As we pointed out in chapter 9 the number of k-states in a single Brillouin zone is equal to the number of unit cells in the entire system. Thus, if each atom has exactly one electron (i.e., is valence 1) there would be exactly enough electrons to fill the band *if there were only one spin state of the electron.* Being that there are two spin states of the electron, when each atom has only one (valence) electron, then the band is precisely half full. This is shown in the left of Fig. 12.1. Here, there is a Fermi surface where the unfilled states meet the filled states. (In the figure, the Fermi energy is shown as green). When a band is partially filled, the electrons can repopulate when a small electric field is applied, allowing current to flow. Thus, the partially filled band is a *metal.*

On the other hand, if there are two electrons per atom, then we have precisely enough electrons to fill one band. One possibility is shown on the left of Fig. 12.2 — the entire lower band is filled and the upper band is empty, and there is a *band gap* between the two bands (note that the chemical potential is between the bands). When this is the situation, the lower (filled) band is known as the *valence band* and the upper (empty) band is known as the *conduction band.* In this situation the minimum energy excitation is moving an electron from the valence to the conduction band, which is nonzero energy. Because of this, at zero temperature, a sufficiently small electric perturbation will not create any excitations at — the system does not respond at all to electric field. Thus, systems of this type are known as (electrical) *insulators* (or more specifically *band insulators*). If the band gap is below about 4 eV, then these type of insulators are called

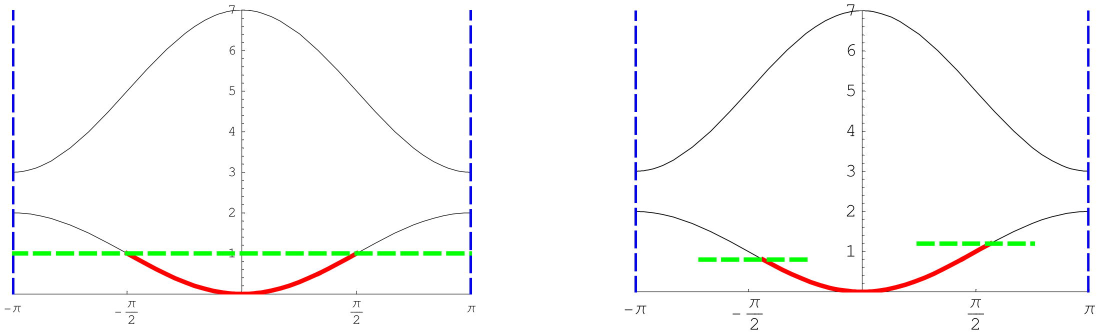


Figure 12.1: Left: A band diagram with two bands shown where each atom has one electron so that the lowest band is exactly half filled, and is therefore a metal. The filled states are colored red, the chemical potential is the green line. Right: When electric field is applied, electrons accelerate, filling some of the k states to the right and emptying k-states to the left (in 1 d this can be thought of as having a different chemical potential on the left versus the right). Since there are an unequal number of left-moving versus right-moving electrons, the situation on the right represents net current flow.

semiconductors since at finite temperature electrons can be thermally excited into the conduction band, and these electrons then can move around freely, carrying some amount of current.

One might want to be aware that in the language of chemists, a band insulator is a situation where all of the electrons are tied up in bonds. For example, in diamond, carbon has valence four — meaning there are four electrons per atom in the outer-most shell. In the diamond lattice, each carbon atom is covalently bonded to each of its four nearest neighbors – and each covalent bond requires two electrons. One electron is donated to each bond from each of the two atoms on either end of the bond – this completely accounts for all of the four electrons in each atom. Thus all of the electrons are tied up in bonds. This turns out to be equivalent to the statement that certain bonding bands are completely filled, and there is no mobility of electrons in any partially filled bands (See the left of Fig. 13.3).

When there are two electrons per atom, one frequently obtains a band insulator. However, another possibility is that the band energies overlap, as shown in the right of Fig. 12.2. In this case, although one has precisely the right number of electrons to fill a single band, instead one has two partially filled bands. As in Fig. 12.1 there are low energy excitations available, and the system is metallic.

In seeking to determine whether a system is of the insulating type of the left of Fig. 12.2 or of the metallic type on the right of Fig. 12.2, one must look at the details of the band structure. However, it is certainly true that if the gap between the bands is sufficiently large, then there is no chance that the higher band will dip down far enough in energy to overlap with the lower band. Recalling from our experience with the nearly-free electron model that the gap between bands formed at the Brillouin zone boundary is proportional to $|V_{\mathbf{G}}|$, we realize that a stronger periodic potential separates the bands and helps form insulators rather than metallic overlapping bands. The limit of strong periodic potential where there is a large spacing between the bands compared to the width of the bands is analogous to the "atomic" limit in the tight binding model discussed in chapter 7.

Back in Section. 4.5 we learned that one property of metals is that they are typically soft

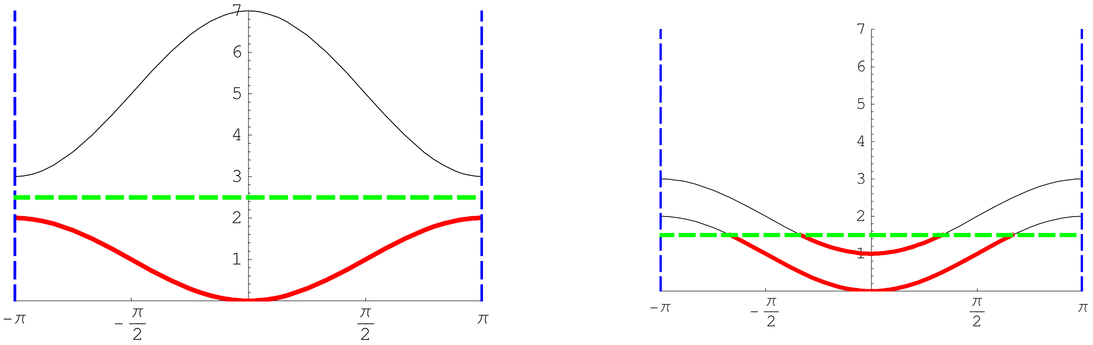


Figure 12.2: When there are two electrons per atom, then there are exactly enough electrons to fill the lowest band. In both pictures the chemical potential is drawn in green. Left: one possibility is that the lowest band (the valence band) is completely filled and there is a gap to the next band (the conduction band) in which case we get an insulator. This is a *direct* band gap as the valence band maximum and the conduction band minimum are both at the same crystal momentum (the zone boundary). Right: Another possibility is that the band energies overlap, in which case there are two bands, each of which is partially filled, giving a metal. If the bands were separated by more (imagine just increasing the vertical spacing between bands) we would have an insulator again, this time with an *indirect* band gap, since the valence band maximum is at the zone boundary while the conduction band minimum is at the zone center.

and malleable. This is a result of the fact that the electrons that hold the atoms together are mobile – in essence, because they are mobile, they can readjust their positions somewhat as the crystal is deformed.

12.2 Energy bands in two (or more) dimensions

It is useful to try to understand how the nearly-free electron model results in band structure in two dimensions. Let us consider a monatomic atom in a square lattice. The Brillouin zone is correspondingly square, and since there is one electron per atom, there should be enough electrons to half fill a single Brillouin zone. In absence of a periodic potential, the Fermi sea forms a circle as shown in the left of Fig. 12.3. The area of this circle is precisely half the area of the zone. Now when a periodic potential is added, gaps open up at the zone boundaries. This means that states close to the zone boundary get moved down in energy – and the closer they are to the boundary, the more they get moved down. As a result, states close to the boundary get filled up preferentially at the expense of states further from the boundary. This deforms the Fermi surface[1] roughly as shown in the right of Fig. 12.3. In either case, there are low energy excitations possible and therefore the system is a metal.

If the periodic potential is strong enough the Fermi surface may even touch[2] the Brillouin

---

[1] Recall that the Fermi surface is the locus of points at the Fermi energy (so all states at the Fermi surface have the same energy), separating the filled from unfilled states. Keep in mind that the area inside the Fermi surface is fixed by the total number of electrons in the system.

[2] Note that whenever a Fermi surface touches the Brillouin zone boundary, it must do so perpendicularly. This is due to the fact that the group velocity is zero at the zone boundary – i.e., the energy is quadratic as one approaches the zone boundary. Since the energy is essentially not changing in the direction perpendicular to the zone boundary,

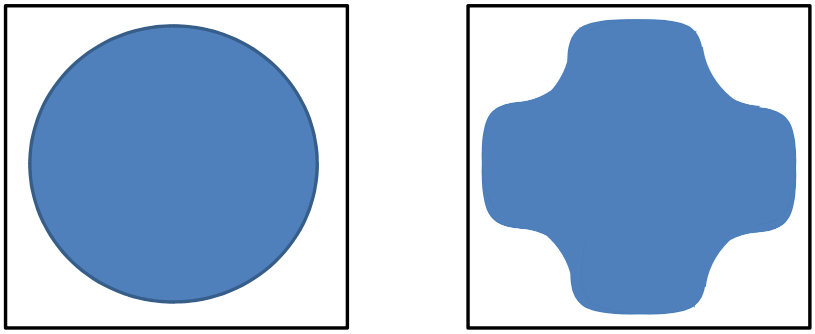


Figure 12.3: Fermi sea of a monatomic square lattice solid in two dimensions. Left: In the absence of a periodic potential, the Fermi sea forms a circle whose area is precisely half that of the Brillouin zone (the black square). Right: when a periodic potential is added, states closer to the zone boundary are pushed down in energy deforming the Fermi sea. Note that the area of the Fermi sea remains fixed.

zone boundary as shown in the left of Fig. 12.4. This is not uncommon in real materials. On the right of Fig. 12.4 the Fermi surface of copper is shown, which similarly touches the zone boundary.

Let us now consider the case of a two-dimensional square lattice of divalent atoms. In this case the number of electrons is precisely enough to fill a single zone. In the absence of a periodic potential, the Fermi surface is still circular, although it now crosses into the second Brillouin zone, as shown in the left of Fig. 12.5. Again, when a periodic potential is added a gap opens at the zone boundary — this gap opening pushes down the energy of all states within the first zone and pushes up energy of all states in the second zone. If the periodic potential is *sufficiently strong*[3], then the states in the first zone are all lower in energy than states in the second zone. As a result, the Fermi sea will look like the right of Fig. 12.5. I..e, the entire lower band is filled, and the upper band is empty. Since there is a gap at the zone boundary, there are no low energy excitations possible, and this system is an insulator.

It is worth considering what happens for intermediate strength of the periodic potential. Again, states outside of the first Brillouin zone are raised in energy and states inside the first

---

the shape of the Fermi surface must be unchanging as the zone boundary is approached.

[3] We can estimate how strong the potential needs to be. We need to have the highest energy state in the first Brillouin zone be lower energy than the lowest energy state in the 2nd zone. The highest energy state in the first zone, in the absence of periodic potential, is in the zone corner and therefore has energy $\epsilon_{state-1} = 2(\pi/2)^2/(2m)$. The lowest energy state in the second zone is it the middle of the zone boundary edge and in the absence of periodic potential has energy $\epsilon_{state-2} = (\pi/2)^2/(2m)$. Thus we need to open up a gap at the zone boundary which is sufficiently large that state 2 becomes higher in energy than state 1. This requires roughly that $2|V_G| = \epsilon_{state-1} - \epsilon_{state-2}$.

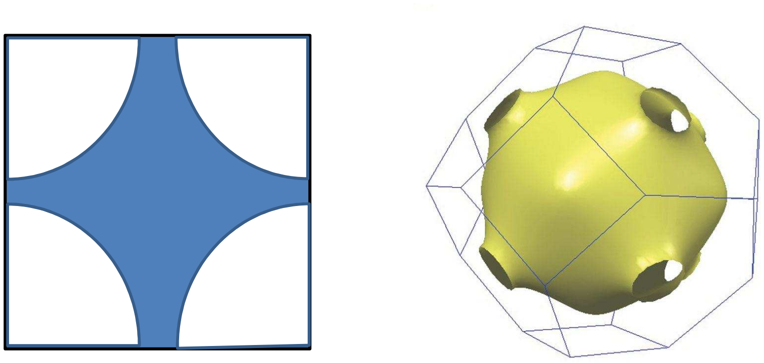


Figure 12.4: Left: Fermi sea of a monatomic square lattice solid in two dimensions with strong periodic potential. The fermi surface touches the Brillouin zone boundary. Right: The fermi surface of copper, which is monovalent (the lattice structure is fcc, which determines the shape of the Brillouin zone, see Fig. 9.6).

Brillouin zone are lowered in energy. Therefore fewer states will be occupied in the second zone and more states occupied in the first zone. However, for intermediate strength of potential, there will remain some states occupied in the second zone and some states empty within the first zone. This is precisely analogous to what happens in the right half of Fig. 12.2. Analogously, there will still be some low energy excitations available, and the system remains a metal.

12.3 Failures of the Band-Structure Picture of Metals and Insulators

The picture we have developed is that the band structure, and the filling of bands, determines whether a material is a metal or insulator, (or semiconductor, meaning an insulator with a small band gap). One thing we might conclude at this point is that any system where the unit cell has a single valence electron (so the first Brillouin zone is half-full) must be a metal. However, it turns out that this is not always true! The problem is that we have left out a very important effect – Coulomb interaction between electrons. We have so far completely ignored the Coulomb repulsion between electrons. Is this neglect justified at all? If we try to estimate how strong the Coulomb interaction is between electrons, ($e^2/(4\pi\epsilon_0 r)$ where r is the typical distance between two electrons — i.e., the lattice constant a) we find numbers on the order of several eV. This number can be larger, or even far larger, than the Fermi energy (which is already a very large number, on the order of 10,000 K). Given this, it is hard to explain why it is at all justified to have thrown out such an important contribution. In fact, one might expect that neglecting this term would give complete nonsense! Fortunately, it turns out that in many cases it *is* OK to assume noninteracting electrons.

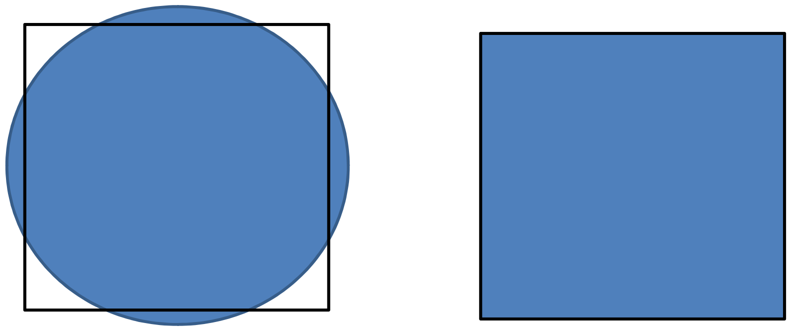


Figure 12.5: Fermi sea of a diatomic square lattice solid in two dimensions. Left: In the absence of a periodic potential, the Fermi sea forms a circle whose area is precisely that of the Brillouin zone (the black square). Right: when a *sufficiently strong* periodic potential is added, states inside the zone boundary are pushed down in energy so that all of these states are filled and no states outside of the first Brillouin zone are filled. Since there is a gap at the zone boundary, this situation is an insulator. (Note that the area of the Fermi sea remains fixed).

The reason this works is actually quite subtle and was not understood until the 1950s due to the work of Lev Landau. This (rather deep) explanation, however, is beyond the scope of this course so we will not discuss it. Nonetheless, with this in mind it is perhaps not too surprising that there are cases where the noninteracting electron picture, and hence our view of band-structure, fails.

Magnets

A case where the band picture of electrons fails is when the system is ferromagnetic[4]. We will discuss ferromagnetism in detail in chapters 16-19 below, but in short this is where, due to interaction effects, the electron spins spontaneously align. From a kinetic energy point of view this seems unfavorable, since filling the lower energy states with two spins can lower the Fermi energy. However, it turns out that aligning all of the spins can lower the Coulomb energy between the electrons, and thus our rules of non-interacting electron band theory no longer hold.

Mott Insulators

Another case where interaction physics is important is the so-called Mott insulator[5]. Consider a monovalent material. From band theory one might expect a half-filled lowest band, therefore a

---

[4] Or antiferromagnetic or ferrimagnetic, for that matter. See chapter 16 below for definitions of these terms.

[5] Named after the English Nobel Laureate, Nevill Mott. Classic examples of Mott insulators include NiO and CoO.

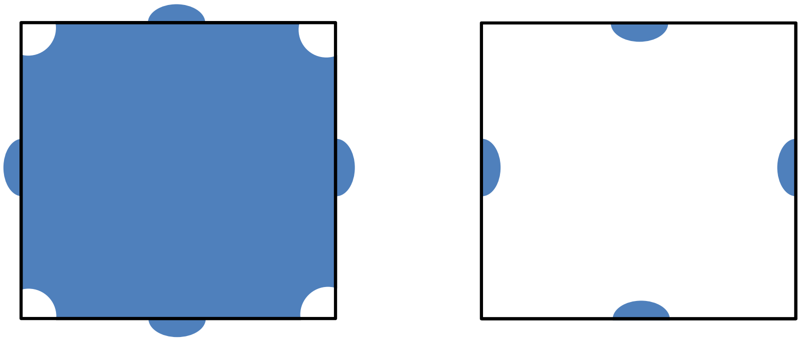


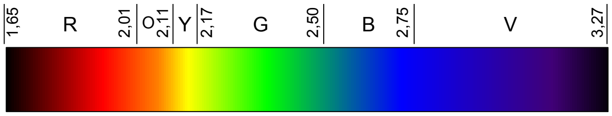
Figure 12.6: Fermi sea of a diatomic square lattice solid in two dimensions. Left: For intermediately strong periodic potential, there are still some states filled in the second zone, and some states empty in the first zone, thus the system is still a metal. Right: The states in the second zone can be moved into the first zone by translation by a reciprocal lattice vector. This is the reduced zone scheme representation of the occupancy of the second Brillouin zone.

metal. But if one considers the limit where the electron-electron interaction is extremely strong, this is not what you get. Instead, since the electron-electron interaction is very strong, there is a huge penalty for two electrons to be on the same atom (even with opposite spins). As a result, the ground state is just one electron sitting on each atom. Since each atom has exactly one electron, no electron can move from its atom – since that would result in a double occupancy of the atom it lands on. As a result, this type of ground state is insulating. In some sense this type of insulator — which can be thought of as more-or-less a traffic jam of electrons — is actually simpler to visualize than a band insulator! We will also discuss Mott insulators further in sections 15.1 and particularly 19.2 below.

12.4 Band Structure and Optical Properties

To the extent that electronic band structure is a good description of the properties of materials (and usually it is), one can attribute many of the optical properties of materials to this band structure. First one needs to know a few simple facts about light shown here in this table:

Color	$\hbar\omega$
Infrared	< 1.65 eV
Red	~ 1.8 eV
Orange	~ 2.05 eV
Yellow	~ 2.15 eV
Green	~ 2.3 eV
Blue	~ 2.7 eV
Violet	~ 3.1 eV
Ultraviolet	> 3.2 eV



12.4.1 Optical Properties of Insulators and Semiconductors

With this table in mind we see that if an insulator (or wide-bandgap semiconductor) has a bandgap of greater than 3.2 eV, then it appears transparent. The reason for this is that a single photon of visible light cannot excite an electron in the valence band into the conduction band. Since the valence band is completely filled, the minimum energy excitation is of the band gap energy – so the photon creates no excitations at all. As a result, the visible optical photons do not scatter from this material at all and they simply pass right through the material[6]. Materials such as quartz, diamond, aluminum-oxide, and so forth are insulators of this type.

Semiconductors with somewhat smaller band gaps will absorb photons with energies above the band gap (exciting electrons from the valence to the conduction band), but will be transparent to photons below this band gap. For example, cadmium-sulfide (CdS) is a semiconductor with band gap of roughly 2.6 eV, so that violet and blue light are absorbed but red and green light are transmitted. As a result this material looks reddish. (See Fig. 12.7).

Figure 12.7: Orange crystals of CdS. This particular crystal is the naturally occurring mineral called "Greenockite" which is CdS with trace amounts of impurity which can change its color somewhat.

---

[6] Very weak scattering processes can occur where, say, two photons together can excite an electron, or a photon excites a phonon

12.4.2 Direct and Indirect Transitions

While the band gap determines the minimum energy excitation that can be made in an insulator (or semiconductor), this is not the complete story in determining whether or not a photon can be absorbed by a material. It turns out to matter quite a bit at which values of $\mathbf{k}$ the maximum of the valence band and the minimum of the conduction band lies. If the value of $\mathbf{k}$ for the valence band maximum is the same as the value of $\mathbf{k}$ for the conduction band minimum, then we say that it is a *direct* band gap. If the values of $\mathbf{k}$ differ, then we say that it is an *indirect* band gap. For example, the system shown on the left of Fig. 12.2 is a direct band gap, where both the valence band maximum and the conduction band minimum are at the zone boundary. In comparison, if the band shapes were as in the right of Fig. 12.2, but the band gap were large enough such that it would be an insulator (just imagine the bands separated by more), this would be an indirect band gap since the valence band maximum is at the zone boundary, but the conduction band minimum is at $\mathbf{k} = \mathbf{0}$.

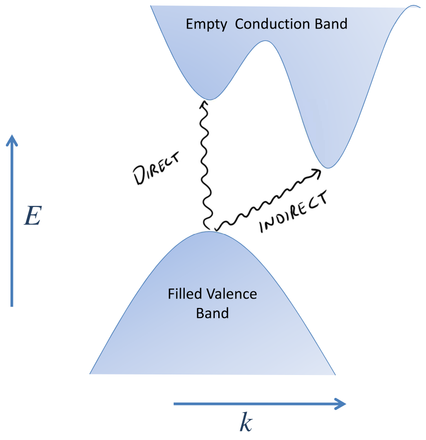


Figure 12.8: Direct and Indirect transitions. While the indirect transition is lower energy, it is hard for a photon to excite an electron across an indirect band gap because photons carry very little momentum (since the speed of light, c, is large).

One can also have both indirect and direct band gaps in the same material, as shown in Fig. 12.8. The minimum energy excitation is the *indirect transition* – meaning excitation of an electron across an indirect band gap, or equivalently a transition of nonzero crystal momentum[7] where the electron is excited from the top of the valence band to the bottom of the lower conduction band at a very different $\mathbf{k}$. While this may be the lowest energy excitation that can occur, it is very hard for this type of excitation to result from exposure of the system to light – the reason for this is energy-momentum conservation. If a photon is absorbed, the system absorbs both the energy and the momentum of the photon. But given an energy E in the eV range, the momentum of the photon $|k| = E/c$ is extremely small, because c is so large. Thus the system cannot conserve momentum while exciting an electron across an indirect band gap. Nonetheless, typically if a system like this is exposed to photons with energy greater than the indirect band gap a small number of electrons will manage to get excited – usually by some complicated process including absorbtion of a photon exciting an electron with simultaneous emission of a phonon[8] to arrange the conservation of energy and momentum. In comparison, if a system has a direct band gap, and

---

[7] By "nonzero" we mean, substantially nonzero – like a fraction of the Brillouin zone.

[8] Another way to satisfy the conservation of momentum is via a "disorder assisted" process. You recall that the

is exposed to photons of energy matching this direct band gap, then it strongly absorbs this energy of photon while exciting electrons from the valence band to the conduction band.

12.4.3 Optical Properties of Metals

The optical properties of metals, however, are a bit more complicated. Since these materials are very conductive, photons (which are electromagnetic) excite the electrons[9], which then re-emit light. This re-emission (or reflection) of light is why metals look shiny. Noble metals (gold, silver, platinum) look particularly shiny because their surfaces do not form insulating oxides when exposed to air, which many metals (such as sodium) do within seconds.

Even amongst metals (ignoring possible oxide surfaces), colors vary. For example, Silver looks brighter than gold and copper, which look yellow or orange-ish. This again is a result of the band structure of these materials. Both of these materials have valence one meaning that a band should be half-filled. However, the total energy width of this band is greater for silver than it is for gold or copper (In tight-binding language t is larger for silver, see chapter 7). This means that higher energy electronic transitions within the band are much more possible for silver than they are for gold and copper. For copper and gold, photons with blue and violet colors are not well absorbed and re-emitted, leaving these material looking a bit more yellow and orange. For silver on the other hand, all visible colors are re-emitted well, resulting in a more perfect (or "white") mirror. While this discussion of the optical properties of metals is highly over-simplified[10], it captures the correct essence — that the details of the band structure determine which color photons are easily absorbed and/or reflected, and this in turn determines the apparent color of the material.

12.4.4 Optical Effects of Impurities

It turns out that small levels of impurities put into periodic crystals (particularly into semiconductors and insulators) can have dramatic effects on many of their optical (as well as electrical!) properties. For example, one nitrogen impurity per million carbon atoms in a diamond crystal gives the crystal a yellow-ish color. One boron atom per million carbon atoms give the diamond a blue-ish color[11]. We will discuss the physics that causes this in section 13.2.1 below.

---

reason we conserve crystal momentum is because the system is perfectly periodic. If the system has some disorder, and is therefore not perfectly periodic, then crystal momentum is not perfectly conserved. Thus the greater the disorder level, the less crystal momentum needs to be conserved and the easier it is to make a transition across an indirect band gap.

[9] Note the contrast with insulators — when an electron is excited above the band gap, since the conductivity is somewhat low, the electron does not re-emit quickly, and the material mostly just absorbs the given wavelength.

[10] Really there are many bands overlapping in these materials and the full story addresses inter and intra-band transitions.

[11] Natural blue diamonds are extremely highly prized and are very expensive. Possibly the world's most famous diamond, the Hope Diamond, is of this type (it is also supposed to be cursed, but that is another story). With modern crystal growth techniques, in fact it is possible to produce man-made diamonds of "quality" better than those that are mined. Impurities can be placed in as desired to give the diamond any color you like. Due to the powerful lobby of the diamond industry, all synthetic diamonds are labeled as such — so although you might feel cheap wearing a synthetic, in fact, you probably own a better product than those that have come out of the earth! (Also you can rest with a clean conscience that the production of this diamond did not finance any wars in Africa).

12.5 Summary

- A material is a metal if it has low energy excitations. This happens when at least one band is partially full. (Band) Insulators and semiconductors have only filled bands and empty bands and have a gap for excitations.
- A semiconductor is a (band) insulator with a small band gap.
- The valence of a material determines the number of carriers being put into the band – and hence can determine if one has a metal or insulator/semiconductor. However, if bands overlap (and frequently they do) one might not be able to fill the bands to a point where there is a gap.
- The gap between bands is determined by the strength of the periodic potential. If the periodic potential is strong enough (the atomic limit in tight binding language), bands will not overlap.
- The band picture of materials fails to account for electron-electron interaction. It cannot describe (at least without modification) interaction driven physics such as magnetism and Mott insulators.
- Optical properties of solids depend crucially on the possible energies of electronic transitions. Photons easily create transitions with low momentum, but cannot create transitions with larger momentum easily. Optical excitations over an indirect (finite momentum) gap are therefore weak.

References

- Goodstein section 3.6c
- Kittel, chapter 7; first section of chapter 8; first section of chapter 9.
- Burns, section 10.7, 10.10
- Hook and Hall, section 4.2,4.3, section 5.4
- Rosenberg section 8.9-8.19

Chapter 13

Semiconductor Physics

13.1 Electrons and Holes

Suppose we start with an insulator or semiconductor and we excite one electron from the valence band to the conduction band, as shown in the left of Fig. 13.1. This excitation may be due to absorbing a photon, or it might be a thermal excitation. (For simplicity in the figure we have shown a direct band gap. For generality we have not assumed that the curvature of the two bands are the same). When the electron has been moved up to the conduction band, there is an absence of an electron in the valence band known as a *hole*. Since a completely filled band is inert, it is very convenient to only keep track of the few holes in the valence band (assuming there are only a few) and to treat these holes as individual elementary particles. Note that the electron can fall back into the empty state that is the hole, emitting energy (a photon, say) and "annihilating" both the electron from the conduction band and the hole from the valence band[1].

Effective Mass of Electrons

As mentioned in section 7.2, it is useful to describe the curvature at the bottom of a band in terms of an effective mass. Let us assume[2] that near the bottom of the conduction band (assumed to be at $\mathbf{k} = \mathbf{k_{min}}$) the energy is given by[3,4]

$$E = E_{min} + \alpha|\mathbf{k} - \mathbf{k_{min}}|^{\mathbf{2}} + \ldots$$

---

[1] This is equivalent to pair annihilation of an electron with a positron. In fact, the analogy between electron-hole and electron-positron is very precise. As soon as Dirac constructed his equation (in 1928) describing the relativistic motion of electrons, and predicting positrons, it was understood that the positron could be thought of as an absence of an electron in an filled sea of states. The filled sea of electron states, with a gap to exciting electron-positron pairs is the inert vacuum, which is analogous to an inert filled valence band.

[2] It is an important principle that near a minimum or a maximum one can always expand and get something quadratic plus higher order corrections.

[3] For simplicity we have assumed the system to be isotropic. In the more general case we would have

$$E = E_{min} + \alpha_x(k_x - k_x^{min})^2 + \alpha_y(k_y - k_y^{min})^2 + \alpha_z(k_z - k_z^{min})^2 + \ldots$$

for some orthogonal set of axes (the "principle axes") x, y, z. In this case we would have an effective mass which can be different in the three different principle directions.

[4] For simplicity we neglect the spin of the electron here. In general, spin-orbit coupling can change the effective g-factor of the electron, for example.

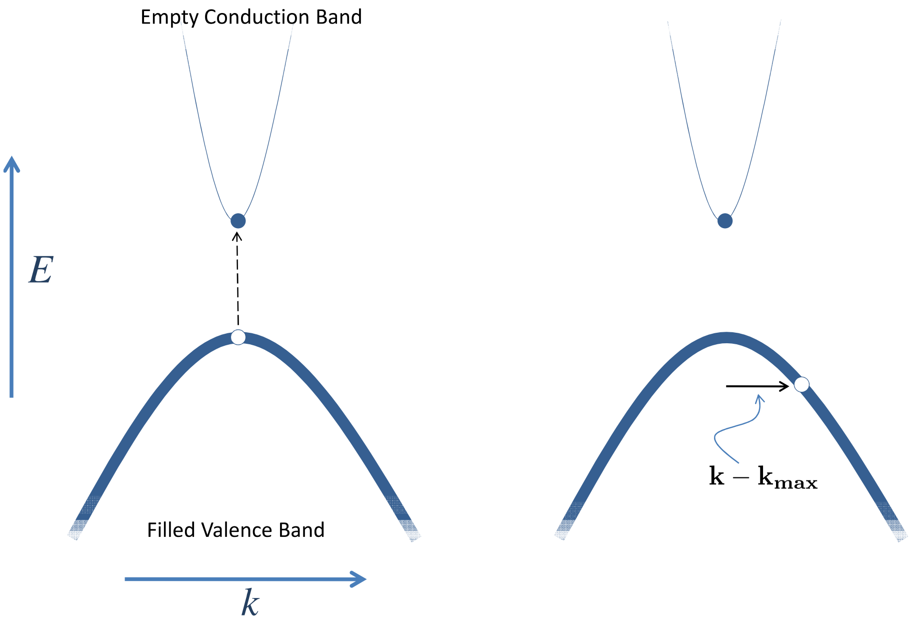


Figure 13.1: Left: A single hole in the valence band and a single electron in the conduction band. Right: Moving the hole to a momentum away from the top of the valence band costs *positive* energy — like pushing a balloon under water. As such, the effective mass of the hole is defined to be positive. The energy of the configuration on the right is greater than that on the left by $E = \hbar^2(k-k_{max})^2/(2m^*)$

where the dots mean higher order term in the deviation from $\mathbf{k_{min}}$. We then *define* the *effective mass* to be given by

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

at the bottom of the band (with the derivative being taken in any direction for an isotropic system). Correspondingly, the (group) velocity is given by

$$\mathbf{v} = \nabla_{\mathbf{k}} E/\hbar = \hbar(\mathbf{k} - \mathbf{k_{min}})/m^*$$

This definition is chosen to be in analogy with the free electron behavior $E = \frac{\hbar^2|\mathbf{k}|^2}{2m}$ with corresponding velocity $\mathbf{v} = \nabla_{\mathbf{k}} E/\hbar = \hbar\mathbf{k}/m$.

Effective Mass of Holes

Analogously we can define the effective mass for holes. Here things get a bit more complicated[5]. For the top of the valence band, the energy dispersion for electrons would be

$$E = E_{max} - \alpha|\mathbf{k} - \mathbf{k_{max}}|^{\mathbf{2}} + \ldots$$

---

[5] Some people find the concept of effective mass for holes to be a bit difficult to digest. I recommend chapter 12 of Ashcroft and Mermin to explain this in more detail (in particular see page 225 and thereafter).

The modern convention is to *define the effective mass for holes to be always positive*[6]

$$\frac{\hbar^2}{m^*} = \left|\frac{\partial^2 E}{\partial k^2}\right| = 2\alpha$$

The convention of the effective mass being positive makes sense because the energy to boost the hole from zero velocity ($\mathbf{k} = \mathbf{k_{max}}$ at the top of the valence band) to finite velocity is positive. This energy is naturally given by

$$E_{hole} = \frac{\hbar^2 |\mathbf{k} - \mathbf{k_{max}}|^\mathbf{2}}{2m^*}$$

The fact that boosting the hole away from the top of the valence band is positive energy may seem a bit counter-intuitive being that the dispersion of the hole band is an upside-down parabola. However, one should think of this like pushing a balloon under water. The lowest energy configuration is with the *electrons* at the lowest energy possible and the hole at the highest energy possible. So pushing the hole under the electrons costs positive energy. (This is depicted in the right hand side of Fig. 13.1.)

13.1.1 Drude Transport: Redux

Back in section 3.1 we studied Drude theory — a simple kinetic theory of electron motion. The main failure of Drude theory was that it did not properly treat the Pauli exclusion principle properly, and neglected the fact that in metals the high density of electrons makes the Fermi energy extremely high. However, in semiconductors or band insulators, when only a few electrons are in the conduction band and/or only a few holes are in the valence band, then we can consider this to be a low density situation, and to a very good approximation, we can ignore Fermi statistics. (For example, if only a single electron is excited into the conduction band, then we can completely ignore the Pauli principle, since it is the only electron around — there is no chance that any state it wants to sit in will already be filled!). As a result, when there is a low density of conduction electrons or valence holes, it turns out that Drude theory works extremely well! We will come back to this issue later in section 13.3 and make this statement much more precise.

At any rate, in the semiclassical picture, we can write a simple Drude transport equation (really Newton's equations!) for electrons in the conduction band

$$m_e^* d\mathbf{v}/dt = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) - m_e^* \mathbf{v}/\tau$$

with m_e^* the electron effective mass. Here the first term on the right hand side is the force on the electron, and the second term is a drag force with an appropriate scattering time τ. The scattering time determines the so-called *mobility* μ which measures the ease with which the particle moves[7]

$$\mu = |\mathbf{E}|/|\mathbf{v}| = |e\tau/m^*|$$

Similarly we can write equations of motion for holes in the valence band

$$m_h^* d\mathbf{v}/dt = e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) - m_h^* \mathbf{v}/\tau$$

---

[6] Be warned: a few books define the mass of holes to be negative. This is extremely annoying but not inconsistent as long as the negative sign shows up somewhere else!

[7] Mobility is defined to be positive for both electrons and holes.

Note that here the charge on the hole is *positive*. This should make sense — the electric field pulls on the electrons in a direction opposite than it pulls on the absence of an electron!

If we think back all the way to chapter 3, one of the physical puzzles that we could not understand is why the Hall coefficient sometimes changes sign (See the table in section 3.1.2). In some cases it looked as if the charge carrier had positive charge. Now we understand why this is true. In some materials the main charge carrier is the hole!

13.2 Adding electrons or holes with impurities: Doping

In a pure band insulator or semiconductor, if we excite electrons from the valence to the conduction band (either with photons or thermally) we can be assured that the number of electrons in the conduction band (typically called n) is precisely equal to the number of holes left behind in the valence band (typically called p). However, in an impure semiconductor or band insulator this is not the case.

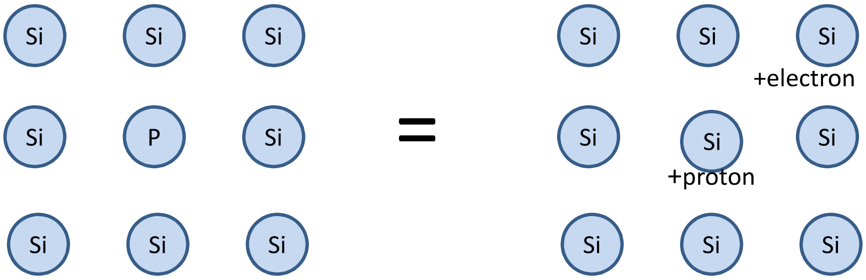


Figure 13.2: Doping Si with P adds one free electron

Consider for example, Silicon (Si), which is a semiconductor with a band gap of about 1.1 eV. Without impurities, a semiconductor is known as *intrinsic*. Now imagine that a Phosphorus (P) atom replaces one of the Si atoms in the lattice as shown on the left of Fig. 13.2. This P atom, being directly to the right of Si on the periodic table, can be thought of as nothing more than a Si atom plus an extra proton and an extra electron[8] as shown on the right of Fig. 13.2. Since the valence band is already filled this additional electron must go into the conduction band. The P atom is known as a *donor* (or *electron donor*) in silicon since it donates an electron to the

---

[8] There is an extra neutron as well, but it doesn't do much in this context.

conduction band. It is also sometimes known as an n-dopant, since n is the term for the density of electrons in the conduction band.

Analogously, we can consider Aluminum, the element directly to the left of Si on the periodic table. In this case, the Aluminum dopant provides one fewer electron than Si, so there will be one missing electron from the valence band. In this case Al is known as an electron *acceptor*, or equivalently as a p-dopant (since p is the symbol for the density of holes)[9].

In a more chemistry oriented language, we can depict the donors and acceptors as shown in Fig. 13.3. Note that in the intrinsic case, all of the electrons are tied up in covalent bonds of two electrons. With the n-dopant, there is an extra unbound electron, whereas with the p-dopant there is an extra unbound hole (one electron too few).

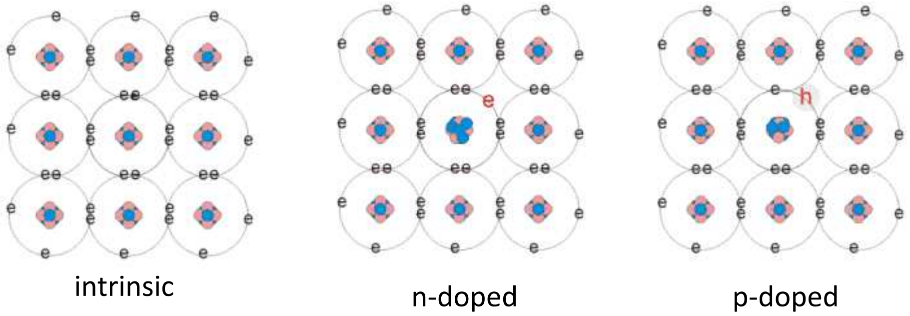


Figure 13.3: n and p doping: In the intrinsic case, all of the electrons are tied up in covalent bonds of two electrons. In the n-dopant case, there is an extra unbound electron, whereas with the p-dopant there is an extra hole.

13.2.1 Impurity States

Let us consider even more carefully what happens when we add dopants. For definiteness let us consider adding an n-dopant such as P to a semiconductor such as Si. Compared to the intrinsic case adding the n-dopant we get a single electron above the gap in the conduction band. This electron behaves like a free particle with mass m_e^*. However, in addition, we have an single extra positive charge $+e$ at some point in the crystal due to the P nucleus. The free electron is attracted back to this positive charge and forms a bound state that is just like a hydrogen atom. There

---

[9] Yes, it is annoying that Phosphorus, the most common n-dopant, has the chemical symbol P

are two main differences between a real hydrogen atom and this bound state of an electron in the conduction band and the impurity nucleus. First of all, the electron has effective mass m_e^* which can be very different from the real (bare) mass of the electron (and is typically smaller than the bare mass of the electron). Secondly, instead of the two charges attracting each other with a potential $V = \frac{e^2}{4\pi\epsilon_0 r}$ they attract each other with a potential $V = \frac{e^2}{4\pi\epsilon_r\epsilon_0 r}$ where ϵ_r is the relative permittivity (or relative dielectric constant) of the material. With these two small differences, the calculation of the hydrogen bound states proceeds exactly as we do for genuine hydrogen in our quantum mechanics courses.

We recall the energy eigenstates of the hydrogen atom are given by $E_n^{H-atom} = -\text{Ry}/n^2$ where Ry is the Rydberg given by

$$\text{Ry} = \frac{me^2}{8\epsilon_0^2 h^2} \approx 13.6\text{eV}$$

with m the electron mass. The corresponding radius of this wavefunction is $r_n \approx n^2 a_0$ with the Bohr radius given by

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{me^2} \approx .51 \times 10^{-10}\text{m}$$

The analogous calculation for a hydrogenic impurity state in a semiconductor gives precisely the same expression, only ϵ_0 is replaced by $\epsilon_0\epsilon_r$ and m is replaced by m^*. One obtains

$$\text{Ry}^{\text{eff}} = \text{Ry}\left(\frac{m^*}{m}\frac{1}{\epsilon_r^2}\right)$$

and

$$a_0^{\text{eff}} = a_0\left(\epsilon_r \frac{m}{m^*}\right)$$

Because the dielectric constant of semiconductors is typically high (roughly 10 for most common semiconductors) and because the effective mass is frequently low (a third of m or even smaller), the effective Rydberg Ry^{eff} can be tiny compared to the real Rydberg, and the effective Bohr radius a_0^{eff} can be huge compared to the real Bohr radius[10]. For example, in Silicon[11], Ry^{eff} is much less than .1 eV and a_0^{eff} is above 30 angstroms! Thus this donor impurity form an energy eigenstate just below the conduction band. At zero temperature this eigenstate will be filled, but it takes only a small temperature to excite some of the bound electrons out of the hydrogenic orbital and into the conduction band.

A depiction of this physics is given in Fig. 13.4 where we have plotted an energy diagram for a semiconductor with donor or acceptor impurities. Here the energies eigenstates are plotted as a function of position. Between the valence and conduction band (which are uniform in position), there are many localized hydrogen-atom-like eigenstates. The energies of these states are in a range of energies but are not all exactly the same since each impurity atom is perturbed by other impurity atoms in its environment. If the density of impurities is high enough, electrons (or holes) can hop from one impurity to the next, forming an *impurity band.*

---

[10] Note that the large Bohr Radius justifies *post-facto* our use of a continuum approximation for the dielectric constant ϵ_r. On small length scales, the electric field is extremely inhomogeneous due to the microscopic structure of the atoms, but on large enough length scales we can use classical electromagnetism and simply model the material as a medium with a dielectric constant.

[11] Because Silicon has an anisotropic band, and therefore an anisotropic mass, the actual formula is more complicated.

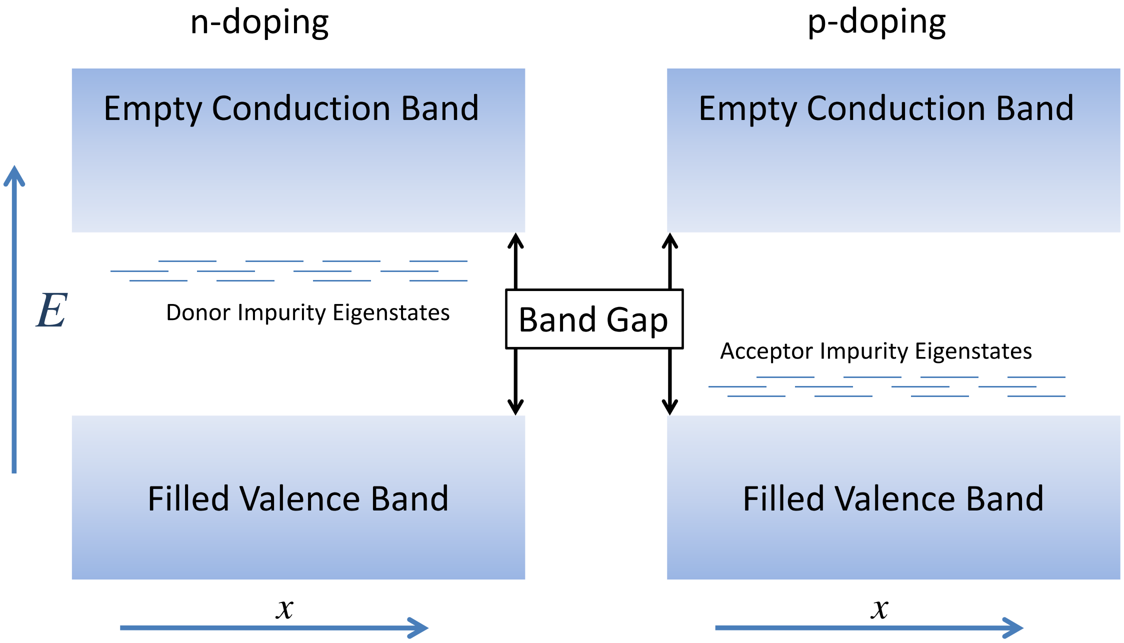


Figure 13.4: An energy diagram of a semiconductor with donor impurities (left) and acceptor impurities (right). The energy eigenstates of the hydrogenic orbitals tiesd to the impurities are not all the same because each impurity is perturbed by neighbor impurities. At low temperature, the donor impurity eigenstates are filled and the acceptor eigenstates are empty. But with increasing temperature, the electrons in the donor eigenstates are excited into the conduction band and similarly the holes in the acceptor eigenstates are excited into the valence band.

Note that because the effective Rydberg is very small, the impurity eigenstates are only slightly below the conduction band or above the valence band respectively. With a small temperature, these donors or acceptors can be thermally excited into the band. Thus, except at low enough temperature that the impurities can bind the carrier, we can think of the impurities as simply adding carriers to the band. So the donor impurities donate free electrons to the conduction band, whereas the acceptor impurities give free holes to the valence band. However, at very low temperature, these carriers get bound back to their respective nuclei so that they can no longer carry electricity, a phenomenon known as *carrier freeze out.*

Note that in the absence of impurities, the Fermi energy (the chemical potential at zero temperature) is in the middle of the band gap. When donor impurities are added, at zero temperature, these states are near the top of the band gap, and are filled. Thus the fermi energy is moved up to the top of the band gap. On the other hand, when acceptors are added, the acceptor states near the bottom of the band gap are empty. (Remember it is a bound state of a hole to a nucleus!). Thus, the fermi energy is moved down to the bottom of the band gap.

Optical Effects of Impurities

As mentioned previously, the presence of impurities in a material can have dramatic effects on its optical properties. There are two main optical effects of impurities. The first effect is that the impurities add charge carriers to an otherwise insulating material – turning an insulator into something that conducts at least somewhat. This obviously can have some important effects on the interaction with light. The second important effect is the introduction of new energy levels within the gap. Whereas before the introduction of impurities, the lowest energy transition that can be made is of the full energy of the gap, now one can have optical transitions between impurity states, or from the bands to the impurity states.

13.3 Statistical Mechanics of Semiconductors

We now use our knowledge of statistical physics to analyze the occupation of the bands at finite temperature.

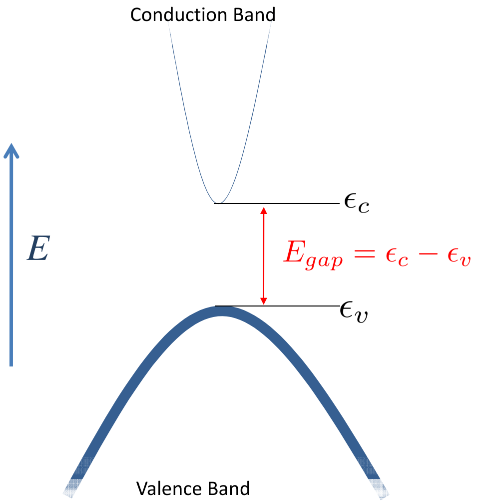


Figure 13.5: A band diagram.

Imagine a band structure as shown in Fig. 13.5. The minimum energy of the conduction band is defined to be ϵ_c and the maximum energy of the valence band is defined to be ϵ_v. The band gap is correspondingly $E_{gap} = \epsilon_c - \epsilon_v$.

Recall from way back in Eq. 3.15 that the density of states per unit volume for free electrons (in three dimensions with two spin states) is given by

$$g(\epsilon \geqslant 0) = \frac{(2m)^{3/2}}{2\pi^2\hbar^3}\sqrt{\epsilon}$$

The electrons in our conduction band are exactly like these free electrons, except that (a) the bottom of the band is at energy ϵ_c and (b) they have an effective mass m_e^*. Thus the density of states for these electrons near the bottom of the conduction band is given by

$$g_e(\epsilon \geqslant \epsilon_c) = \frac{(2m_e^*)^{3/2}}{2\pi^2\hbar^3}\sqrt{\epsilon - \epsilon_c}$$

Similarly the density of states for holes near the top of the valence band is given by

$$g_e(\epsilon \leqslant \epsilon_v) = \frac{(2m_h^*)^{3/2}}{2\pi^2\hbar^3}\sqrt{\epsilon_v - \epsilon}$$

At fixed chemical potential μ the total number of electrons n in the conduction band, as a function of temperature T is thus given by

$$n(T) = \int_{\epsilon_c}^{\infty} d\epsilon \, \frac{g_e(\epsilon)}{e^{\beta(\epsilon-\mu)} + 1}$$

(with $\beta = 1/(k_b T)$ as usual). If the chemical potential is "well below" the conduction band (i.e., if $\beta(\epsilon - \mu) \gg 1$), then we can approximate

$$\frac{1}{e^{\beta(\epsilon-\mu)} + 1} \approx e^{-\beta(\epsilon-\mu)}$$

In other words, Fermi statistics can be replaced by Boltzmann statistics when the temperature is low enough that the density of electrons in the band is very low. (We have already run into this principle in section 13.1.1 when we discussed that Drude theory, a classical approach that neglects Fermi statistics actually works very well for electrons above the band gap in semiconductors!). We thus obtain

$$\begin{aligned} n(T) &\approx \int_{\epsilon_c}^{\infty} d\epsilon g_e(\epsilon) e^{-\beta(\epsilon-\mu)} = \frac{(2m_e^*)^{3/2}}{2\pi^2\hbar^3} \int_{\epsilon_c}^{\infty} d\epsilon \, (\epsilon - \epsilon_c)^{1/2} e^{-\beta(\epsilon-\mu)} \\ &= \frac{(2m_e^*)^{3/2}}{2\pi^2\hbar^3} e^{\beta(\mu-\epsilon_c)} \int_{\epsilon_c}^{\infty} d\epsilon \, (\epsilon - \epsilon_c)^{1/2} e^{-\beta(\epsilon-\epsilon_c)} \end{aligned}$$

The last integral is (using $y^2 = x$).

$$\int_0^{\infty} dx \, x^{1/2} e^{-\beta x} = 2\int_0^{\infty} dy \, y^2 e^{-\beta y^2} = 2\frac{d}{d\beta}\int_0^{\infty} e^{-\beta y^2} = \frac{d}{d\beta}\sqrt{\frac{\pi}{\beta}} = \frac{1}{2}\beta^{-3/2}\sqrt{\pi}$$

Thus we obtain the standard expression for the density of electrons in the conduction band

$$n(T) = \frac{1}{4}\left(\frac{2m_e^* k_b T}{\pi\hbar^2}\right)^{3/2} e^{-\beta(\epsilon_c-\mu)} \tag{13.1}$$

Note that this is mainly just exponential activation from the chemical potential to the bottom of the conduction band, with a prefactor which doesn't change too quickly as a function of temperature (obviously the exponential changes very quickly with temperature!).

Quite similarly, we can write the number of holes in the valence band p as a function of temperature

$$p(T) = \int_{-\infty}^{\epsilon_v} d\epsilon\, g_h(\epsilon) \left[1 - \frac{1}{e^{\beta(\epsilon-\mu)}+1}\right] = \int_{-\infty}^{\epsilon_v} d\epsilon\, \frac{g_h(\epsilon) e^{\beta(\epsilon-\mu)}}{e^{-\beta(\epsilon-\mu)}+1}$$

Again, if μ is substantially above the top of the valence band, we have $e^{\beta(\epsilon-\mu)} \ll 1$ so we can replace this by

$$p(T) = \int_{-\infty}^{\epsilon_v} d\epsilon\, g_h(\epsilon) e^{\beta(\epsilon-\mu)}$$

and the same type of calculation then gives

$$p(T) = \frac{1}{4}\left(\frac{2 m_h^* k_b T}{\pi \hbar^2}\right)^{3/2} e^{-\beta(\mu-\epsilon_v)} \tag{13.2}$$

again showing that the holes are activated from the chemical potential *down* into the valence band. (Recall that pushing a hole down into the valence band costs energy!).

Law of Mass Action

A rather crucial relation is formed by combining Eq. 13.1 with 13.2.

$$\begin{aligned} n(T)p(T) &= \frac{1}{2}\left(\frac{k_b T}{\pi\hbar^2}\right)^3 (m_e^* m_h^*)^{3/2} e^{-\beta(\epsilon_c-\epsilon_v)} \\ &= \frac{1}{2}\left(\frac{k_b T}{\pi\hbar^2}\right)^3 (m_e^* m_h^*)^{3/2} e^{-\beta E_{gap}} \end{aligned} \tag{13.3}$$

where we have used the fact that the gap energy $E_{gap} = \epsilon_c - \epsilon_v$. Eq. 13.3 is sometimes known as the *law of mass action*[12], and it is true independent of doping of the material.

Intrinsic Semiconductors

For an intrinsic (i.e., undoped) semiconductor the number of electrons excited into the conduction band must be equal to the number of holes left behind in the valence band so $p = n$. We can then divide Eq. 13.1 by 13.2 to get

$$1 = \left(\frac{m_e^*}{m_h^*}\right)^{3/2} e^{-\beta(\epsilon_v+\epsilon_c-2\mu)}$$

---

[12] The nomenclature here "law of mass action" is a reference to an analogue in chemistry. In chemical reactions we may have an equilibrium between two objects A and B and their compound AB. This is frequently expressed as

$$A + B \rightleftharpoons AB$$

There is some chemical equilibrium constant K which gives the ratio of concentrations

$$K = \frac{[A][B]}{[AB]}$$

where $[X]$ is the concentration of species X. This constant K remains fixed independent of the individual concentrations. In semiconductor physics it is quite similar, only the "reaction" is

$$e + h \rightleftharpoons 0$$

the annihilation of an electron and a hole. So that the product of $[e] = n$ and $[h] = p$ is fixed

Taking log of both sides gives the useful relation

$$\mu = \frac{1}{2}(\epsilon_c + \epsilon_v) + \frac{3}{4}(k_b T)\log(m_h^*/m_e^*) \tag{13.4}$$

Note that at zero temperature, the chemical potential is precisely mid-gap.

Using either this expression, or by using the law of mass action along with the constraint $n = p$, we can obtain an expression for the *intrinsic* density of carriers in the semiconductor

$$n_{intrinsic} = p_{intrinsic} = \sqrt{np} = \frac{1}{\sqrt{2}}\left(\frac{k_b T}{\pi\hbar^2}\right)^{3/2}(m_e^*\, m_h^*)^{3/4}\, e^{-\beta E_{gap}/2}$$

Doped Semiconductors

For doped semiconductors, the law of mass action still holds. If we further assume that the temperature is high enough so that there is no carrier freeze out (i.e, carriers are not bound to impurities) then we have

$$n - p = (\text{density of donors}) - (\text{density of acceptors})$$

This, along with the law of mass action gives us two equations in two unknowns which can be solved[13] . In short, the result is that if we are at a temperature where the undoped intrinsic carrier density is much greater than the dopant density, then the dopants do not matter much, and the chemical potential is roughly midgap as in Eq. 13.4. On the other hand, if we are at a temperature where the intrinsic undoped density is much smaller than the dopant density, then the temperature does not matter much and we can think of this as a zero temperature situation. In the n-doped case, the bottom of the conduction band gets filled with the density of electrons from the donors, and the chemical potential gets shifted up into the conduction band. Correspondingly, in the p-doped case, holes fill the top of the valence band, and the chemical potential gets shifted down into the valence band. (Note that in the latter case of strong doping, the majority carrier concentration is obtained just from the doping, whereas the minority carrier concentration — which might be very small — is obtained via law of mass action).

---

[13] Here is how to solve these two equations. Let

$$D = \text{doping} = n - p = (\text{density of donors}) - (\text{density of acceptors})$$

Let us further assume that $n > p$ so $D > 0$ (we can do the calculation again making the opposite assumption, at the end). Also let

$$I = n_{intrinsic} = p_{intrinsic}$$

so that

$$I^2 = \frac{1}{2}\left(\frac{k_b T}{\pi\hbar^2}\right)^3 (m_e^*\, m_h^*)^{3/2}\, e^{-\beta E_{gap}}$$

from the law of mass action. We can then construct

$$D^2 + 4I^2 = (n-p)^2 + 4np = (n+p)^2$$

So we obtain

$$\begin{aligned} n &= \frac{1}{2}\left(\sqrt{D^2+4I^2} + D\right) \\ p &= \frac{1}{2}\left(\sqrt{D^2+4I^2} - D\right) \end{aligned}$$

As stated in the main text, if $I \gg D$ then the doping D is not important. On the other hand, if $I \ll D$ then the majority carrier density is determined by the doping only and the thermal factor I is unimportant.

13.4 Summary of Statistical Mechanics of Semiconductors

- Holes are the absence of an electron in the valence band. These have positive charge (electrons have negative charge), and positive effective mass. Energy of a hole gets larger at larger momentum (away from the maximum of the band) as they get pushed down into the valence band. The positive charge of the hole as a charge carrier explains the puzzle of the sign of the Hall coefficient.
- Effective mass of electrons determined by the curvature at the bottom of the conduction band. Effective mass of holes determined by curvature at top of conduction band.
- Mobility of a carrier is $\mu = |e\tau/m^*|$
- When very few electrons are excited into the conduction band, or very few holes into the valence band, Boltzmann statistics is a good approximation for Fermi statistics and Drude theory is accurate.
- Electrons or Holes can be excited thermally, or can be added to a system by *doping*. The law of mass action assures that the product np is fixed independent of the amount of doping (only depends on the temperature, the effective masses, and the band gap).
- At very low temperature carriers may freeze-out, binding to the impurity atoms that they came from. However, because the effective Rydberg is very small, carriers are easily ionized into the bands.

References

- Ashcroft and Mermin, chapter 28. A very good discussion of holes and their effective mass is given in chapter 12.
- Rosenberg, chapter 9
- Hook and Hall, 5.1-5.5
- Kittel, chapter 8
- Burns, chapter 10 not including 10.17 and after.
- Singleton chapter 5-6

Chapter 14

Semiconductor Devices

The development of semiconductor devices, such as the transistor no doubt changed the world. Every iPad, iPod, iPhone, and iBook literally contains billions of semiconductor transistors. Simple devices, like alarm clocks, TVs, radios, or cars, contain many thousands or even millions of them. It is hard to overstate how much we take these things for granted these days.

This chapter discusses the physics behind some of devices you can make with semiconductors.

14.1 Band Structure Engineering

To make a semiconductor device one must have control over the detailed properties of materials (band gap, doping, etc) and one must be able to assemble together semiconductors with differing such properties.

14.1.1 Designing Band Gaps

A simple example of engineering a device is given by Aluminum-Galium-Arsenide. GaAs is a semiconductor (zincblende structure as in Fig. 10.4) with a direct band gap about $E_{gap,k=0}$(GaAs) = 1.4 eV. AlAs is the same structure except that the Ga has been replaced by Al and the gap[1] at $k = 0$ is about 2.7 eV. One can also produce alloys (mixtures) where some fraction (x) of the Ga has been replaced by Al which we notate as $\mathrm{Al_xGa_{1-x}As}$. To a fairly good approximation the direct band gap just interpolates between the direct band gaps of the pure GaAs and the pure AlAs. Thus we get roughly (for $x < .4$)

$$E_{gap}(x) = (1-x)\,1.4\,\mathrm{eV} + x\,2.7\,\mathrm{eV}$$

By producing this type of alloyed structure allows one to obtain any desired band gap in this type of material[2].

---

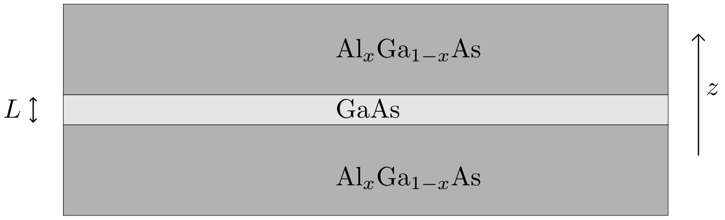
[1] AlAs is actually an indirect band gap semiconductor, but for $x < .4$ or so $\mathrm{Al_xGa_{1-x}As}$ is direct band gap as well.

[2] By alloying the material with arbitrary x, one must accept that the system can no longer be precisely periodic but instead will be some random mixture. It turns out that as long as we are concerned with long wavelength electron waves (i.e, states near the bottom of the conduction band or the top of the valence band) this randomness is very effectively averaged out and we can roughly view the system as being a periodic crystal of As and some average of a $\mathrm{Al_xGa_{1-x}}$ atom.. This is known as a "virtual crystal" approximation.

In the context of device physics one might want to build, for example, a laser out of a semiconductor. The lowest energy transition which recombines a hole with an electron is the gap energy (this is the "lasing" energy typically). By tuning the composition of the semiconductor, one can tune the energy of the gap and therefore the optical frequency of the laser.

14.1.2 Non-Homogeneous Band Gaps

By constructing structures where the materials (or the alloying of a material) is a function of position, one can design more complex environments for electrons or holes in a system. Consider for example, the structure shown in the following figure:



Here a layer of GaAs with lower band gap is inserted between two layers of AlGaAs which has a higher band gap. This structure is known as a "quantum well". In general a semiconductor made of several varieties of semiconductors is known as a *semiconductor heterostructure.* A band diagram of the quantum well structure as a function of the vertical position z is given in Fig. 14.1. The band gap is lower in the GaAs region than in the AlGaAs region. The changes in band energy can be thought of as a potential that an electron(or hole) would feel. For example, an electron in the valence band can have a lower energy if it is in the quantum well region (the GaAs region) than it can have in the AlGaAs region. An electron in the valence band with low energy will be trapped in this region. Just like a particle in a box, there will be discrete eigenstates of the electron's motion in the z direction, as shown in the Figure. The situation is similar for holes in the valence band (recall that it requires energy to push a hole down into the valence band), so there will similarly be confined particle in a box states for holes in the quantum well.

14.1.3 Summary of the Examinable Material

- One can tune band gaps by forming an alloy
- Band gaps act as a potential for electrons and holes.

References on Inhomogeneous Semiconductors

There are many good references on semiconductors (See also the material listed below). Almost all of them discuss the p-n junction first (which is nonexaminable for us). I recommend Hook and Hall section 6.6 on the quantum well to cover the above material.

The rest of the material in this chapter is **NOT EXAMINABLE**. But since semiconductors really did change the world, you might be interested in learning it anyway!

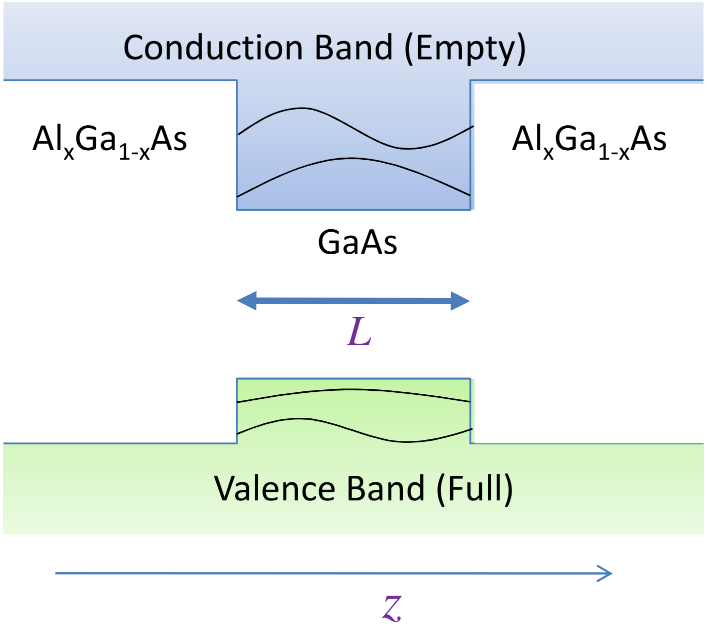


Figure 14.1: Band diagram of a quantum well. A single electron in the conduction band can be trapped in the particle-in-a-box states in the quantum well. Similarly, a hole in the valence band can be trapped in the quantum well.

14.2 p-n Junction

The p-n junction is a junction in a semiconductor between a region of p-doping and a region of n doping. This type of junction has the remarkable property of *rectification*: it will allow current to flow through the junction easily in one direction, but not easily (with very high resistance) in the other direction[3].

Consider

OK, I haven't finished this chapter. Cut me some slack, typing these notes was a load of work! Anyway, it is nonexaminable material, so don't worry about it too much.

---

[3] The phenomenon of rectification in semiconductors was discovered by Karl Ferdinand Braun way back in 1874, but was not understood in detail until the middle of the next century. This discovery was fundamental to the development of radio technology. Braun was awarded the Nobel Prize in 1909 with Guglielmo Marconi for contributions to wireless telegraphy. Perhaps as important to modern communication, Braun also invented the cathode ray tube (CRT) which formed the display for televisions for many years until the LCD display arrived very recently. (The CRT is known as a "Braun Tube" in many countries).

Part VII

Magnetism and Mean Field Theories

Chapter 15

Paramagnets and Diamagnets

The first question one might ask is why we are interested in magnets. While the phenomenon of magnetism was known to the ancients[1], it has only been since the discovery of quantum mechanics that we have come to any understanding of what causes this effect[2]. It may seem like this is a relatively small corner of physics for us to focus so much attention (indeed, several chapters), but we will see that magnetism is a particularly good place to observe interesting effects of both statistical physics and quantum physics[3]. As we mentioned in section 12.3, one place where the band theory of electrons fails is in trying to describe magnets. Indeed, this is precisely what makes magnets interesting! In fact, magnetism remains an extremely active area of research in physics (with many many hard and unanswered questions remaining). Much of condensed matter physics continues to use magnetism as a testing ground for understanding complex and interesting physics both theoretically and in the laboratory.

Let us first make some definitions. Recall that for a small magnetic field, the magnetization of a system $\mathbf{M}$ (moment per unit volume) is often related linearly to the applied[4] magnetic field $\mathbf{H}$ by a (magnetic) susceptibility χ. We write for small fields $\mathbf{H}$,

$$\mathbf{M} = \chi \mathbf{H} \tag{15.1}$$

Note that χ is dimensionless. For small susceptibilities (which is almost always true, except in ferromagnets) there is little difference between $\mu_0 \mathbf{H}$ and $\mathbf{B}$ (with μ_0 the permeability of free space), so we can also write

$$\mathbf{M} = \chi \mathbf{B}/\mu_{\mathbf{0}} \tag{15.2}$$

A *paramagnet* is a material where $\chi > 0$ (i.e., the resulting magnetization is in the same direction as the applied field).

---

[1] Both the Chinese and the Greeks probably knew about magnetic properties of magnetite (Fe_3O_4) (also known as loadstone when magnetized) possibly as far back as several thousands of years BCE (with written records existing as far back as 600 years BCE). One legend has it that a shepherd named Magnes, in the provence of Magnesia, had the nails in his shoes stuck to a large metallic rock and the scientific phenomenon became named after him.

[2] Animal magnetism not withstanding... (that was a joke).

[3] In fact there is a theorem by Niels Bohr and Hendrika van Leeuwen which shows that any treatment of statistics mechanics without quantum mechanics (i.e., classical statistical mechanics) can never produce a nonzero magnetization.

[4] The susceptibility is defined in terms of H. With a long rod-shaped sample oriented parallel to the applied field, H is the same outside and inside the sample, and is thus directly controlled by the experimentalist. The susceptibility is defined in terms of this standard configuration. However, more generally, one needs to take care that the internal field $\mathbf{B}$ that any electrons in the sample respond to is related to $\mathbf{H}$ via $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$.

A *diamagnet* is a material where $\chi < 0$ (i.e., the resulting magnetization is in the opposite direction from the applied field).

What we usually think of as a "magnet" (the thing that holds notes to the fridge) is a ferromagnet (or ferrimagnet). We will discuss these in the next chapter. In this chapter we focus on para- and dia- magnetism. For the sake of completeness though, let us also define

A *ferromagnet* is a material where $\mathbf{M}$ can be nonzero, even in the absence of any applied magnetic field.

Review: Pauli paramagnetism

In section 3.2.3 we calculated the susceptibility of the free Fermi gas. We found a susceptibility of

$$\chi_{Pauli} = \mu_0 \mu_B^2 g(E_F) \tag{15.3}$$

with $g(E_F)$ the density of states at the Fermi surface. The same expression holds for metals with nontrivial band structure — only the density of states may be modified. We will not say much more about this type of paramagnetism since it was discussed before.

15.1 Free Spin (Curie) Paramagnetism

Another example that we are probably familiar with from your statistical physics course is the case of a free spin-1/2. The Hamiltonian, you recall, of a single spin-1/2 is given by[5]

$$\mathcal{H} = -g\mu_B \mathbf{B} \cdot \boldsymbol{\sigma} \tag{15.4}$$

with g the g-factor of the spin which we set to be 2, and $\mu_B = e\hbar/(2m_e)$ is the Bohr magneton. The eigenstates of $\mathbf{B} \cdot \boldsymbol{\sigma}$ are $\pm B/2$ so we have a partition function

$$Z = e^{-\beta\mu_B B} + e^{\beta\mu_B B} \tag{15.5}$$

and a corresponding free energy $F = -T \log Z$ giving us a magnetic moment (per spin) of

$$m = -\frac{dF}{dB} = \mu_B \tanh(\beta\mu_B B) \tag{15.6}$$

With M being the moment per volume, at small field this results in a susceptibility of[6]

$$\chi = \lim_{H\to 0} \frac{dM}{dH} = \frac{\rho\mu_0\mu_B^2}{k_b T} \tag{15.7}$$

where ρ is the number of spins per unit volume (and we have used $B \approx \mu_0 H$ with μ_0 the permeability of free space). Expression 15.7 is known as the "Curie Law"[7] susceptibility (Actually any susceptibility of the form $\chi \sim C/(k_b T)$ for any constant C is known as Curie law).

---

[5] Again we have switched to using $\mathcal{H}$ for the Hamiltonian since H can be confused with magnetic field.

[6] Note that the susceptibility, per spin, of a Fermi gas, given in Eq. 15.3 is smaller than this susceptibility of a free spin by roughly a factor of T/E_F. We should be familiar this idea, that due to the Pauli exclusion principle, only the small fraction of spins near the fermi surface can be flipped over, therefore giving a small susceptibility.

[7] Named after Pierre Curie. This was well before Pierre married his mega-brilliant wife Marie Sklodowska Curie. She won one physics Nobel with him, and then another one in chemistry after he died. Half-way between the two prizes, Pierre was killed when he was run over by a horse-drawn vehicle while crossing the street (Be careful!).

Aside: From Eq. 15.5 we notice that the partition function of a free spin is only a function of the dimensionless ratio $\mu_B B/(k_b T)$. From this we can derive that the entropy S is also a function only of the same dimensionless ratio. Let us imagine now we have a system of free spins at magnetic field B and temperature T, and we thermally isolate it from the environment. If we adiabatically reduce B, then since S must stay fixed, the temperature must drop proportionally to the reduction in B. This is the principle of the adiabatic demagnetization refrigerator.

We might ask how a "free spin" can occur in a solid, making the above calculation of free spin paramagnetism valid. Our understanding of electrons in solids so far describes electrons as being either in full bands, in which case they cannot be flipped over, at all; or in partially full bands, in which case the calculation of the Pauli susceptibility in section 3.2.3 is valid — albeit possibly with a modified density of states at the Fermi surface to reflect the details of the band structure. So how is it that we can have a free spin?

Let us think back to the description of Mott insulators in section 12.3. In these materials, the Coulomb interaction between electrons is strong enough that no two electrons can double occupy the same site of the lattice. As a result, having one electron per site results in a "traffic jam" of electrons where no one can hop to any other site. When this sort of Mott insulator forms, there is exactly one electron per site, which can be either spin-up or spin-down. Thus we have a free spin on each site exactly as we considered in the previous section!

In fact, the situation can be much more complex than this. There may be several valence electrons per atom, and these electrons have both orbital angular momentum **L** as well as spin angular momentum **S**, and these angular momenta for the several valence electrons may or may not align with each other.

15.1.1 Isolated Atom Limit: Hund's Rules

To understand how the electrons on the atoms behave, it is worth first considering the behavior of electrons on isolated atoms (the atomic limit). Here, there are a set of rules, known as "Hund's Rules"[8] which determine how the electrons do, or do not, align (this should be familiar from your atomic physics course). Perhaps the simplest way to describe these rules is to consider an explicit example.

Here we consider the atom praseodymium (Pr) which is a rare earth element. This element has two electrons in its outer-most shell, which is an f-shell[9] meaning it has angular momentum $L = 3$, and therefore 7 possible values of L_z, and of course 2 possible values of the spin for each electron. So where in these possible state do we put the two electrons?

Hund's First Rule (paraphrased): Electrons try to align their spins.

Given this rule, we know that the two valence electrons in Pr will have their spins point in the same direction, thus giving us a total spin-angular momentum $S = 1$ from the two $S = 1/2$ spins. So locally (meaning on the same atom), the two electron spins behave ferromagnetically — they align. The reason for this we will discuss below in section 15.3, but in short, it is a result of the Coulomb interaction between electrons (and between the electrons and the nucleus) – the Coulomb energy is lower when the two electron spins align.

---

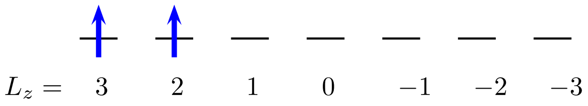
[8] Friedrich Hermann Hund was an important physicist and chemist. He died in 1997 at the age of 101. He wrote down his rules in 1925.

[9] Recall the nomenclature s,p,d,f to mean $L = 0, 1, 2, 3$

We now have to decide which orbital states to put the electrons in.

> **Hund's Second Rule**(paraphrased): Electrons try to maximize their total orbital angular momentum, consistent with Hund's first rule.

For the case of Pr, we fill the $L_z = 3$ and $L_z = 2$ states to make the maximum possible $L_z = 5$ (this gives $L = 5$, and by rotational invariance we can point $\mathbf{L}$ in any direction equally well). Thus, we have a picture as follows



We have put the spins as far as possible to the left to maximize L_z, but note that we could not have put both electrons in the same orbital, since they have to be spin-aligned and we must obey the Pauli principle. Again the rule of maximizing orbital angular momentum is driven by the physics of Coulomb interaction.

At this point we have $S = 1$ and $L = 5$, but we still need to think about whether the orbital and spin angular momenta actually align with each other

> **Hund's Third Rule**(paraphrased): Given Hund's first and second rules, the orbital and spin angular momentum either align or antialign, so that the total angular momentum is $J = |L \pm S|$ with the sign being determined by whether the shell of orbitals is more than half filled (+) or less than half filled (−).

The reason for this rule is not interaction physics, but is spin-orbit coupling. The Hamiltonian will typically have a spin-orbit term $\alpha \mathbf{L} \cdot \mathbf{S}$, and the sign of α determines how the spin and orbit align to minimize the energy (we will not discuss why this sign obeys the half-filled shell rule above).

Thus for the case of Pr, where $L = 5$ and $S = 1$ and the shell is less than half filled, we have total angular momentum $J = L - S = 4$.

Free Spin of Spin J

Thus we expect our Hamiltonian to be of the form[10]

$$\mathcal{H} = -\tilde{g}\mu_B \mathbf{B} \cdot \mathbf{J} \tag{15.8}$$

with $\mathbf{J}$ being of spin J as dictated by Hund's rule (compare to Eq. 15.4). It is easy enough to construct the partition function

$$Z = \sum_{J_z=-J}^{J} e^{\beta \tilde{g} \mu_B B J_z}$$

---

[10] Actually, the Hamiltonian should be of the form

$$\mathcal{H} = -\mu_B \mathbf{B} \cdot (\mathbf{L} + g\mathbf{S})$$

This can be converted to the form of Eq. 15.8 using Wigner-Eckart theorem where the effective g-factor is

$$\tilde{g} = \frac{3}{2} + \frac{1}{2}\left[\frac{S(S+1) - L(L+1)}{J(J+1)}\right]$$

See for example, Ashcroft and Mermin chapter 31 and appendix P.

and from it determine the magnetization and the susceptibility (this is assigned as an "Additional Problem" for those who are interested). The result, of the Curie form, is that the susceptibility per unit volume is given by

$$\chi = \frac{\rho \mu_0 (\tilde{g}\mu_B)^2}{3} \frac{J(J+1)}{k_B T}$$

(Compare Eq. 15.7)

15.1.2 Atoms in Solids

While the above discussion of paramagnetism works extremely well for isolated atoms, when atoms are in solids, the above description may or may not work. It turns out that for rare earth atoms (such as Pr) the above discussion is extremely good. However, in other cases, the atoms may be highly influenced by their environment, modifying (or completely invalidating) the above discussion. In particular, the atoms are no longer in a rotationally symmetric environment, they see the potential due to neighboring atoms — the so-called "crystal field". In this case orbital angular momentum is not conserved and the degeneracy of states all having the same L^2 is broken, a phenomenon known as *crystal field splitting*.

As a (very) cartoon picture of this physics, we can imagine a crystal which is highly tetragonal (see Fig. 8.11) where the lattice constant in one direction is quite different from the constant in the other two. We might imagine that an atom that is living inside such an elongated box would have a lower energy if its orbital angular momentum pointed along the long axes (say, the z-axis), rather than in some other direction. In this case, we might imagine that $L_z = +L$ and $L_z = -L$ might be lower energy than any of the other possible values of L.

Another thing that may happen due to crystal field splitting is that the orbital angular momentum may be pinned to have zero expectation (for example, if the ground state is a superposition of $L_z = +L$ and $L_z = -L$). In this case, the orbital angular momentum decouples from the problem completely (a phenomenon known as *quenching* of the orbital angular momentum), and the only magnetically active degrees of freedom are the spins. This is precisely what happens for most transition metals.

The most important moral to take home from this section is that paramagnets can have many different effective values of J, and one needs to know the microscopic details of the system before deciding which spin and orbital degrees of freedom are active.

15.2 Larmor Diamagnetism

Diamagnetism is when a material develops a magnetization that *opposes* the applied magnetic field. This is similar in spirit to Lenz's law (part of Faraday's law) that an induced current always opposes the change causing it. However, the analogy is not precise. If a magnetic field is applied to a loop of wire, current will flow to create a magnetization in the opposite direction. However, in any (nonsuperconducting) loop of wire, the current will eventually decay back to zero and there will be no magnetization remaining. In a diamagnet, in contrast, the magnetization remains so long as the applied magnetic field remains.

So where might we find diamagnetism? First of all, we must not have free spins, because these are paramagnetic, as we just saw. Secondly, we should not have a Fermi surface, because then we have Pauli paramagnetism as we saw in section 3.2.3[11]. So we must have a non-metal system without free spins. I.e., where

$$\langle \mathbf{L} \rangle = \langle \mathbf{S} \rangle = 0$$

A classic situation in which this occurs is atoms with filled shell configurations, like the Noble gases – these form solids at very low temperature (with the exception of Helium which forms a superfluid instead). These solids are insulators, and have overall zero spin and zero orbital angular momentum on each atom. Nonetheless, the electrons can still respond to the external magnetic field, as we shall see now.

Let us recall how electrons are usually coupled to a magnetic field. Recall that the Hamiltonian for a charged particle in a magnetic field $\mathbf{B}$ takes the minimal coupling form[12]

$$\mathcal{H} = \frac{(\mathbf{p} + e\mathbf{A})^2}{2m}$$

where $-e$ is the charge of the particle, and $\mathbf{A}$ is the vector potential such that

$$\nabla \times \mathbf{A} = \mathbf{B}$$

For a uniform magnetic field, we may take $\mathbf{A} = \frac{1}{2}\mathbf{r} \times \mathbf{B}$ we then have[13]

$$\mathcal{H} = \frac{\mathbf{p}^2}{2m} + \frac{e}{2m}\mathbf{p} \cdot (\mathbf{r} \times \mathbf{B}) + \frac{e^2}{2m}\frac{1}{4}|\mathbf{B} \times \mathbf{r}|^2 \tag{15.9}$$

The first term in this equation is the usual kinetic energy. The second term can be rewritten as

$$\frac{e}{2m}\mathbf{p} \cdot (\mathbf{r} \times \mathbf{B}) = -\frac{e}{2m}\mathbf{B} \cdot (\mathbf{r} \times \mathbf{p}) = -\frac{e\hbar}{2m}\mathbf{B} \cdot \mathbf{L} = -\mu_B \mathbf{B} \cdot \mathbf{L} \tag{15.10}$$

which is exactly the form of the Hamiltonian that we used in Eq. 15.8 above! We have already decided that in this section we are considering atoms where $\langle \mathbf{L} \rangle = 0$, so in this particular case this second term is not important at lowest order[14].

The third term of Eq. 15.9, which is known as the *diamagnetic term* of the Hamiltonian, is the one in which we are interested. If we imagine that $\mathbf{B}$ is applied in the $\hat{z}$ direction, the expectation of this term can be written as

$$\delta E = \frac{e^2}{8m}\langle |\mathbf{B} \times \mathbf{r}|^2 \rangle = \frac{e^2 B^2}{8m}\langle x^2 + y^2 \rangle$$

---

[11] See footnote 13 in section 3.2.3: free electrons have a diamagnetic term which cancels some of the Pauli paramagnetism, but is smaller than the Pauli term so overall the free electrons are paramagnetic

[12] Recall that minimal coupling requires $\mathbf{p} \to \mathbf{p} - q\mathbf{A}$ where q is the charge of the particle. Here our particle has charge $q = -e$.

[13] Note that while p_i does not commute with r_i, it does commute with r_j for $j \neq i$, so there is no ordering problem between $\mathbf{p}$ and $\mathbf{r} \times \mathbf{B}$

[14] To be more precise, if we include the coupling to the particle spin, $\mathbf{L}$ is replaced here by $\mathbf{J} = \mathbf{L} + g\mathbf{S}$. Now, even though $\langle \mathbf{J} \rangle$ may be zero, the term in Eq. 15.10 may be important in second order perturbation theory. At second order, the energy of the system will be corrected by a term proportional to

$$\delta E_0 \sim + \sum_{n>0} \frac{|\langle n|\mathbf{B} \cdot \mathbf{J}|0\rangle|^2}{E_0 - E_n}$$

This contribution need not vanish. It is largest when there is a low energy excitation E_n so the denominator can be small. Since this energy *decreases* with increasing H, this term is paramagnetic. At any rate, this contribution can be important for certain elements where $J = 0$ but L and S are individually nonzero — as this usually implies there is a low energy excitation that can occur by misorienting $\mathbf{L}$ and $\mathbf{S}$ with respect to each other. (However, for atoms like Noble gases, where $\mathbf{L}$ and $\mathbf{S}$ are individually zero, then there are no low energy excitations and this contribution is negligible.) This type of paramagnetism is known as *Van Vleck paramagnetism* after the Nobel Laureate J. H. Van Vleck who was a professor at Baliol in 1961-1962 but spent most of his later professional life at Harvard.

Using the fact that the atom is rotationally symmetric, we can write

$$\langle x^2 + y^2 \rangle = \frac{2}{3}\langle x^2 + y^2 + z^2 \rangle = \frac{2}{3}\langle r^2 \rangle$$

Thus we have

$$\delta E = \frac{e^2 B^2}{12m}\langle r^2 \rangle$$

Thus the magnetic moment per electron is

$$m = -\frac{dE}{dB} = \left[\frac{e^2}{6m}\langle r^2 \rangle\right] B$$

or a susceptibility of

$$\chi = -\frac{\rho e^2 \mu_0 \langle r^2 \rangle}{6m} \tag{15.11}$$

where ρ is the density of electrons per unit volume. Eq. 15.11 is known as *Larmor Diamagnetism.*[15] For most atoms, $\langle r^2 \rangle$ is on the order of a few Bohr radii squared. Note that to determine the density of electrons, one uses the density of atoms in the solid multiplied by the number of electrons per atom. This expression works extremely well for calculating the diamagnetic susceptibility of atoms such as Noble gases. (The same expression can sometimes work for large conductive molecules if the electrons can freely travel the length of the molecule – by taking $\langle r^2 \rangle$ to be the radius squared of the molecule instead of that of the atom).

15.3 Why Spins Align

We now return, as promised above, to discuss roughly why Hund's rules work – in particular we want to know why spins (real spin moments or orbital moments) like to align. This section will be only qualitative, but should give at least a rough idea of the right physics.

First of all, the tendency for spins to align has nothing to do with magnetic dipole interactions. While two magnetic dipoles do interact with each other, when dipole moments are on the order of the Bohr magneton, this energy scale becomes tiny — way too small to matter for anything interesting.

Instead, the tendency for spins to align comes from the Coulomb interaction energy. To see how this works, let us consider a wavefunction for two electrons on an atom.

Naive Argument

The overall wavefunction must be antisymmetric by Pauli's exclusion principle. We can generally write

$$\Psi(\mathbf{r_1}, \sigma_\mathbf{1}; \mathbf{r_2}, \sigma_\mathbf{2}) = \psi_{orbital}(\mathbf{r_1}, \mathbf{r_2})\ \chi_{spin}(\sigma_1, \sigma_2)$$

where $\mathbf{r_i}$ are the particles' positions and σ_i are their spin. Now, if the two spins are aligned, say both are spin-up (i.e., $\chi_{spin}(\uparrow, \uparrow) = 1$ and $\chi_{spin} = 0$ for other spin configurations) then the spin

---

[15] Larmor was a rather important physicist in the late 1800s. Among other things, he published the Lorentz transformations for time dilation and length contraction two years before Lorentz, and seven years before Einstein. However, he insisted on the aether, and rejected relativity at least until 1927 (maybe longer).

wavefunction is symmetric and the spatial wavefunction $\psi_{orbital}$ must be antisymetric. As a result we have

$$\lim_{\mathbf{r_1} \to \mathbf{r_2}} \psi_{orbital}(\mathbf{r_1}, \mathbf{r_2}) \to 0$$

So electrons with aligned spins cannot get close to each other, thus reducing the Coulomb energy of the system.

The argument we have just given is frequently stated in textbooks. Unfortunately, it is not the whole story.

More Correct

In fact it turns out that the crucial Coulomb interaction is that between the electron and the nucleus. Consider the case where there are two electrons and a nucleus as shown in Fig. 15.1. What we see from this figure is that positive charge of the nucleus seen by one electron is screened by the negative charge of the other electron. This screening reduces the binding energy of the electrons to the nucleus. However, when the two spins are aligned, the electrons repel each other and therefore screen the nucleus less effectively. In this case, the electrons see the full charge of the nucleus and bind more strongly, thus lowering their energies.

Another way of understanding this is to realize that when the spins are not aligned, sometimes one electron gets between the other electron and the nucleus – thereby reducing the effective charge seen by the outer electron, reducing the binding energy, and increasing the total energy of the atom. However, when the electrons are spin aligned, the Pauli principle largely prevents this configuration from occurring, thereby lowering the total energy of the system.

One must be somewhat careful with this argument however. In the case of a diatomic molecule (say H_2), we have two electrons and two nuclei. While the screening effect discussed above still occurs, and tries to align the electrons, it is somewhat less effective than for two electrons on a single atom — since most of the time the two electrons are near opposite nuclei anyway. Furthermore, there is a competing effect that tends to make the electrons want to anti-align. As we discussed in section 4.3.1 when we discussed covalent bonding, we can think of the two nuclei as being a square well (See Fig. 4.4) – and the bonding is really a particle-in-a-box problem. There is some lowest energy (symmetric) wavefunction in this large two-atom box, and the lowest energy state of two electrons would be to have the two spins anti-aligned so that both electrons can go in the same low energy spatial wavefunction. It can thus be quite difficult to determine whether electrons on neighboring atoms want to be aligned or anti-aligned. Generally either behavior is possible. We will discuss this much further below. The energy difference between having the spins on two atoms aligned versus anti-aligned is usually known as the *exchange interaction.*

15.4 Summary

- Susceptibility $\chi = dM/dH$ is positive for paramagnets and negative for diamagnets.
- Sources of paramagnetism: (a) Pauli paramagnetism of free electron gas (See section 3.2.3) (b) Free spin paramagnetism – know how to do the simple statmech exercise of calculating the paramagnetism of a free spin.
- The magnitude of the free spin determined by Hund's rules. The environment of this atom, or crystal field, can modify this result. ixtHund's Rules

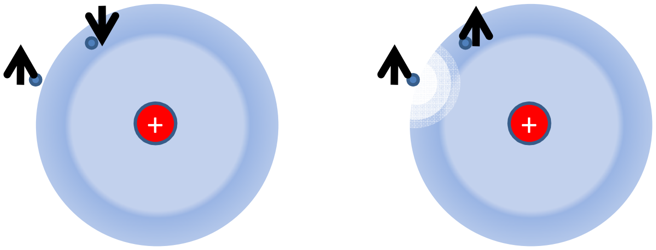


Figure 15.1: Why aligned spins have lower energy (Hund's rule). In this figure, the wavefunction is depicted for one of the electrons whereas the other electron (the one further left) is depicted as having fixed position. When the two electrons have opposite spin, the effective charge of the nucleus seen by the fixed electron is reduced by the screening provided by the other electron (left figure) . However, when the spins are aligned, the two electrons cannot come close to each other (right figure) and the fixed electron sees the full charge of the nucleus. As such the binding of the fixed electron to the nucleus is stronger in the case where the two electrons are spin aligned, therefore it is a lower energy configuration.

- Larmor diamagnetism can occur when atoms have $\mathbf{J} = \mathbf{0}$, therefore not having strong paramagnetism. This comes from the magnetic term of the Hamiltonian in first order perturbation theory. The diamagnetism per electron is proportional to the radius of the orbital.

References

- Ibach and Luth, section 8.1
- Hook and Hall, chapter 7
- Ashcroft and Mermin, chapter 31
- Kittel, chapter 11
- Blundell, chapter 2
- Burns, chapter 15A
- Goodstein, section 5.4a-c (doesn't cover diamagnetism)
- Rosenberg, chapter 11 (doesn't cover diamagnetism)

Chapter 16

Antiferro-, Ferri-, and Ferro-magnetism.

We concluded the last chapter by noting that spins on neighboring atoms could favor either having their spins aligned or could favor having their spins anti-aligned. In this chapter we explore the possibility this interaction between neighboring spin aligns spins on a macroscopic scale.

We first assume that we have an insulator, i.e., electrons do not hop from site to site[1]. We then write a model Hamiltonian as

$$\mathcal{H} = -\frac{1}{2}\sum_{i,j} J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j - \sum_i g\mu_B \mathbf{B} \cdot \mathbf{S}_i \tag{16.1}$$

Here $\mathbf{S}_i$ is the electron spin on site i and $\mathbf{B}$ is the magnetic field experienced by the spins. Note that in Eq. 16.1, $J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j$ is the interaction energy[2] between spin i and spin j. Note that we have included a factor of 1/2 out front to avoid overcounting, since the sum actually counts both J_{ij} and J_{ji} (which are equal).

If $J_{ij} > 0$ then it is lower energy when spins i and j are aligned, whereas if $J_{ij} < 0$ then it is lower energy when the spins are anti-aligned.

The coupling between spins drops rapidly as the distance between the spins increases. A good model to use is one where only nearest neighbor spins interact with each other. Frequently one writes (neglecting the magnetic field $\mathbf{B}$)

$$H = -\frac{1}{2}\sum_{i,j\; neighbors} J_{ij}\,\mathbf{S}_i \cdot \mathbf{S}_j$$

or as a shorthand,

$$H = -\frac{1}{2}\sum_{\langle i,j\rangle} J_{ij}\,\mathbf{S}_i \cdot \mathbf{S}_j$$

---

[1] This might be the situation if we have a Mott insulator, as described in section 12.3 above where strong interaction prevents electron hopping.

[2] WARNING: Many references use Heisenberg's original convention that the interaction energy should be defined as $2J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j$ rather than $J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j$. However, more modern researchers use the latter, as we have here. This matches up with the convention used for the Ising model below Eq. 16.5, where the convention $2J$ is never used. At any rate, if someone on the street tells you J, you should ask whether the intend a factor of 2 or not.

In a uniform system where each spin is coupled to its neighbors with the same strength, we can drop the indices from J_{ij} (since they all have the same value) and we obtain the so-called *Heisenberg Hamiltonian*

$$H = -\frac{1}{2} \sum_{\langle i,j \rangle} J \, \mathbf{S}_i \cdot \mathbf{S}_j \tag{16.2}$$

16.1 (Spontaneous) Magnetic Order

As in the case of a ferromagnet, it is possible that even in the absence of any applied magnetic field, magnetism — or ordering of magnetic moments — may occur. This type of phenomenon is known as *spontaneous* magnetic order (since it occurs without application of any field). It is a subtle statistical mechanical question as to when magnetic interaction in a Hamiltonian actually results in spontaneous magnetic order. At our level of analysis we will assume that systems can always find ground states which "satisfy" the magnetic Hamiltonian. In chapter 18 we will consider how temperature might destroy this magnetic ordering.

16.1.1 Ferromagnets

As mentioned above, if $J > 0$ then neighboring spins want to be aligned. In this case the ground state is when all spins align together developing a macroscopic magnetic moment — this is what we call a *ferromagnet*, and is depicted on the left of Fig. 16.1. We will return to study these further in sections 17.1 and 18 below.

16.1.2 Antiferromagnets

On the other hand, if $J < 0$, neighboring spins want to point in opposite directions, and the most natural ordered arrangement is a periodic situation where alternating spins point in opposite directions, as shown on the right of Fig. 16.1 – this is known as an *antiferromagnet*. This antiferromagnet has zero net magnetization but yet is magnetically ordered. This type of antiperiodic ground state is sometimes known as a *Néel state* after Louis Néel who first proposed that these states exist[3]. We should be cautioned that our picture of spins pointing in directions is a classical picture, and is not quite right quantum mechanically. Particularly when the spin is small (like spin-1/2) the effects of quantum mechanics are strong and classical intuition can fail us. We will have a homework problem that shows that this classical picture of the antiferromagnet is not quite right, although it is fairly good when the spin on each site is larger than 1/2.

Detecting Antiferromagnetism with Diffraction

Being that antiferromagnets have zero net magnetization, how do we know they exist? What is their signature in the macroscopic world? For homework we will explore a very nice method of determining that something is an antiferromagnet by examining its susceptibility as a function of temperature (in fact it was this type of experiment that Néel was analyzing when he realized that antiferromagnets exist). However, this method is somewhat indirect. A more direct approach is

---

[3] Néel won a Nobel prize for this work in 1970.

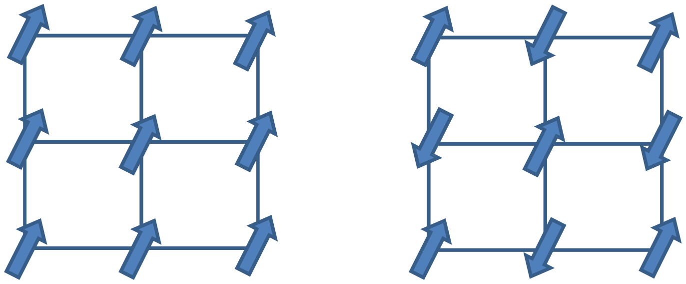


Figure 16.1: Left: Ferromagnet – all spins aligned (at least over some macroscopic regions) giving finite magnetization. Middle: Antiferromagnet — Neighboring spins antialigned, but periodic. this so-called Néel state has zero net magentization.

to examine the spin configuration using diffraction of neutrons. As mentioned in section 10.2, one can prepare spin-polarized neutrons. With spin polarized neutrons, the scattering from the two different spin states of the the atom differ. The neutrons then see that the unit cell in this antiferromagnet is actually of size $2a$ where a is the distance between atoms (i..e, the distance between two atoms with the same spin is $2a$). Thus when the spins align antiferromagnetically, the neutrons will develop scattering peaks at reciprocal wavevectors $G = 2\pi/(2a)$ which would not exist if all the atoms were aligned the same way. This type of neutron diffraction experiment are definitive in showing that antiferromagnetic order exists[4].

Frustrated Antiferromagnets

On certain lattices, for certain interactions, there is no ground state that fully "satisfies" the interaction for all spins. For example, on a triangular lattice if there is an antiferromagnetic interaction between bonds, there is no way that all the spins can point in the opposite direction from their neighbors. For example, as shown in the left of Fig. 16.2 on a triangle, once two of the of the spins are aligned opposite each other, independent of which direction the spin on the last site points, it not be antiparallel to one of its neighbors. It turns out that (assuming the spins are classical variables) the ground state of the antiferromagnetic Heisenberg hamiltonian on a triangle is the configuration shown on the right of Fig. 16.2. While each bond is not quite optimally anti-aligned, the overall energy is optimal for this Hamiltonian[5]

---

[4] These are the experiments that won the Nobel prize for Clifford Schull.

[5] Try showing this!

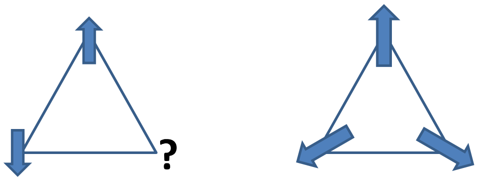


Figure 16.2: Left: An antiferromagnetic interaction on a triangular lattice is frustrated – not all spins can be antialigned with all of their neighbors. Right: The ground state of antiferromagnetic interaction on a triangle for classical spins (large S) is the state on the right, where spins are at 120° to their neighbor.

16.1.3 Ferrimagnetism

Once one starts to look for magnetic structure in materials, one can find many other interesting possibilities. One very common possibility is where you have a unit cell with more than one variety of atom, where the atoms have differing moments, and although the ordering is antiferromagnetic (neighboring spins point in opposite direction) there is still a net magnetic moment. An example of this is shown in Fig. 16.3. Here, the red atoms have a smaller moment than the green atoms and point opposite the green atoms. This type of configuration, where one has antiferromagnetic order, yet a net magnetization due to differing spin species, is known as *ferrimagnetism*. In fact, many of the most common magnets, such as magnetite (Fe_3O_4) are ferrimagnetic.

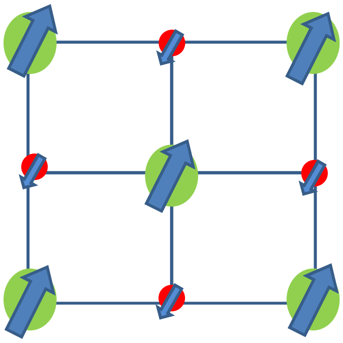


Figure 16.3: A Ferrimagnet: Ordering is antiferromagnetic, but because the differnt spin species have different moments, there is a net magneization

16.2 Breaking Symmetry

In any of these ordered states, we have not yet addressed the question of which direction the spins will actually point. Strictly speaking, the Hamiltonian Eq. 16.2 is rotationally symmetric — if it is a ferromagnet, the magnetization can point in any direction and the energy will be the same! In a real system, however, this is rarely the case: due to the asymmetric environment the atom feels within the lattice, there will be some directions that the spins would rather point than others (This physics was also discussed above in section 15.1.2). Thus to be more accurate we might need to add an additional term to the Heisenberg Hamiltonian. One possible example of the type of term that might be added is[6]

$$H = -\frac{1}{2}\sum_{\langle i,j\rangle} J\mathbf{S}_i \cdot \mathbf{S}_j - \kappa \sum_i (S_i^z)^2 \tag{16.3}$$

(again dropping any external magnetic field). The κ term here favors the spin to be pointing in the $+\hat{z}$ direction or the $-\hat{z}$ direction, but not in any other direction. (You could imagine this being appropriate for a tetragonal crystal elongated in the $\hat{z}$ direction). This energy from the κ term is sometimes known as the *anisotropy energy* since it favors certain directions over others. Another possible Hamiltonian is

$$H = -\frac{1}{2}\sum_{\langle i,j\rangle} J\mathbf{S}_i \cdot \mathbf{S}_j - \tilde{\kappa} \sum_i [(S_i^x)^4 + (S_i^y)^4 + (S_i^z)^4] \tag{16.4}$$

which favors the spin pointing along any of the orthogonal axis directions – but not towards any in-between angle.

In some cases (as we discussed in Section 15.1.2) the coefficient κ may be substantial. In other cases it may be very small. However, since the pure Heisenberg Hamiltonian Eq. 16.2 does not prefer any particular direction, even if the anisotropy (κ) term is extremely small, it will determine the direction of the magnetization in the ground state. We say that this term "breaks the symmetry." Of course, there may be some symmetry remaining. For example, in Eq. 16.3, if the interaction is ferromagnetic, the ground state magnetization may be all spins pointing in the $+\hat{z}$ direction, or equally favorably, all spins pointing in the $-\hat{z}$ direction.

16.2.1 Ising Model

If the anisotropy (κ) term is extremely large, then this term can fundamentally change the Hamiltonian. For example, let us take a spin-S Heisenberg model. Adding the κ term in 16.3 with a large coefficient, forces the spin to be either $S_z = +S$ or $S_z = -S$ with all other values of S_z having a much larger energy. In this case, a new effective model may be written

$$\mathcal{H} = -\frac{1}{2}\sum_{\langle i,j\rangle} J\sigma_i\sigma_j - g\mu_B B \sum_i \sigma_i \tag{16.5}$$

---

[6] For small values of the spin quantum number, these added interactions may be trivial. For example, for spin 1/2, we have $(S_z)^2 = (S_y)^2 = (S_z)^2 = 1$. However, as S becomes larger, the spin becomes more like a classical vector and these such κ terms will favor the spin pointing in the corresponding directions.

where $\sigma_i = \pm S$ only (and we have re-introduced the magnetic field B). This model is known as the *Ising model*[7] and is an extremely important model in statistical physics[8].

16.3 Summary of Magnetic Orders

- Ferromagnets: spins align. Antiferromagnets: spins antialign with neighbors so no net magnetization. Ferrimagnets: spins antialign with neighbors, *but* alternating spins are different magnitude so there is a net magnitization anyway. Microscopic spins stuctures of this sort can be observed with spin-polarized neutrons.
- Useful model Hamiltonians include Heisenberg ($-J\mathbf{S_i} \cdot \mathbf{S_j}$) for isotropic spins, and Ising $-JS_i^z S_j^z$ for spins that prefer to align along only one axis.
- Spins generally do not equally favor all directions (as the Heisenberg model suggest). Terms that favor spins along particular axes may be weak or strong. Even if they are weak, they will pick a direction among otherwise equally likely directions.

References

- Blundell, section 5.1-5.3 (Very nice discussion, but covers mean field theory at the same time which we will cover in chapter 18 below.)
- Burns, section 15.4-15.8 (same comment).

---

[7] "Ising" is properly pronounced "Ee-sing" or "Ee-zing". In the United States it is habitually mispronounced "Eye-sing". The Ising model was actually invented by Wilhelm Lenz (another example of Stigler's law, see footnote 6 in section 11.2). Ising was the graduate student who worked on this model for his graduate thesis.

[8] The Ising model is frequently referred to as the "hydrogen atom" of statistical mechanics since it is extremely simple, yet it shows many of the most important features of complex statistical mechanical systems. The one dimensional version of the model was solved in by Ising in 1925, and the the dimensional version of the model was solved by Onsager in 1944 (a chemistry nobel Laureate, who was amusingly fired by my alma-mater, Brown University, in 1933). Onsager's achievement was viewed as so important that Wolfgang Pauli wrote after world war two that "nothing much of interest has happened [in physics during the war] except for Onsagers exact solution of the two-dimensional Ising model. If you are very brave, you might try calculating the free energy of the one-dimensional Ising model at finite temperature.

Chapter 17

Domains and Hysteresis

17.1 Macroscopic Effects in Ferromagnets: Domains

We might think that in a ferromagnet, all the spins in the system will align as described above in the Heisenberg (or Ising) models. However, in real magnets, this is frequently not the case. To understand why this is we imagine splitting our sample into two halfs as shown in Fig. 17.1. Once we have two dipoles it is clear that they would be lower energy if one of them flipped over as shown it at the far right of Fig. 17.1. (The two north faces of these magnets repel each other[1]). This energy, the long range dipolar force of a magnet, is not described in the Heisenberg or Ising model at all. In those models we have only included nearest neighbor interaction between spins. As we mentioned above, the actual dipolar force between electronic spins (or orbital spins) is tiny compared to the interaction "exchange" force between neighboring spins. But when you put together a whole lot of atoms (like 10^{23} of them!) to make a macroscopic magnet, the summed effect of their dipole moment can be substantial.

Of course, in an actual ferromagnet (or ferrimagnet), the material does not really break apart, but nonetheless different regions will have magnetization in different directions in order to minimize the dipolar energy. A region where the moments all point in one given direction is known as a *domain* or a *Weiss domain*.[2] The boundary of a domain, where the magnetization switches direction is known as a *domain wall*[3]. Some possible examples of domain structures are sketched in Fig. 17.2. In the left two frames we imagine an Ising-like ferromagnet where the moment can only point up or down. The left most frame shows a magnet with net zero magnetization. Along the domain walls, the ferromagnetic Hamiltonian is "unsatisfied". In other words, spin-up atoms on one side of the domain wall have spin-down neighbors — where the Hamiltonian says that they should want to have spin up neighbors only . What is happening is that the system is paying an

---

[1] Another way to understand the dipolar force is to realize that the magnetic field far away from the magnets will be much lower if the two magnets are antialigned with each other. Since the electromagnetic field carries energy $\int dV |B|^2/\mu_0$, minimizing this magnetic field lowers the energy of the two dipoles.

[2] After Pierre-Ernest Weiss, one of the fathers of the study of magnets from the early 1900s.

[3] Domain walls can also occur in antiferromagnets. Instead of the magnetization switching directions we imagine a situation where to the left of the wall, the up-spins are on the even sites, and the down-spins are on the odd-sites, whereas on the right of the domain wall the up-spins are on the odd sites and the down spins are on the even sites. At the domain wall, two neighboring sites will be aligned rather than anti-aligned. Since antiferromagnets have no net magnetization, the argument that domain walls should exist in ferromagnets is not valid for antiferromagnets. In fact, it is always energetically unfavorable for domain walls to exist in antiferromagnets, although they can occur at finite temperature.

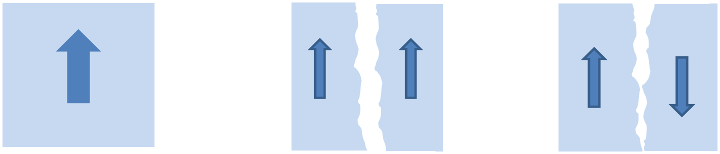


Figure 17.1: Left: The original ferromagnet. Middle: The original ferromagnet broken into two halfs. Right: Because two dipoles next to each other are lower energy if their moments are anti-aligned, the two broken halves would rather line up in opposing directions to lower their energies (the piece on the right hand side has been flipped over here). This suggests that in large ferromagnets, domains may form.

energy cost along the domain wall in order that the *global* energy associated with the long range dipolar forces is minimized.

If one applies a small up-pointing external field to this system, one will obtain the middle picture where the up pointing domains grow at the expense of the down pointing domains to give an overall magnetization of the sample. In the rightmost frame of Fig. 17.2 we imagine a sample where the moment can point along any of the crystal axis directions[4]. Again in this picture the total magnetization is zero but it has rather complicated domain structure.

17.1.1 Disorder and Domain Walls

The detailed geometry of domains in a ferromagnet depends on a number of factors. First of all, it depends on the overall geometry of the sample. (For example, if the sample is a very long thin rod and the system is magnetized along the long axis, it may gain very little energy by forming domains). It also depends on the relative energies of the neighbor interaction versus the long range dipolar interaction: increasing the strength of the long range dipolar forces with respect to the neighbor interaction will obviously decrease the size of domains (having no long range dipolar forces, will result in domains of infinite size). Finally, the detailed disorder in a sample can effect the shape and size of magnetic domains. For example, if the sample is polycrystaline, each domain could be a single crystallite (a single microscopic crystal).

---

[4] See for example the Hamiltonian, Eq. 16.4, which would have moments pointing only along the coordinate axes – although that particular Hamiltonian does now have the long range magnetic dipolar interaction written, so it would not form domains.

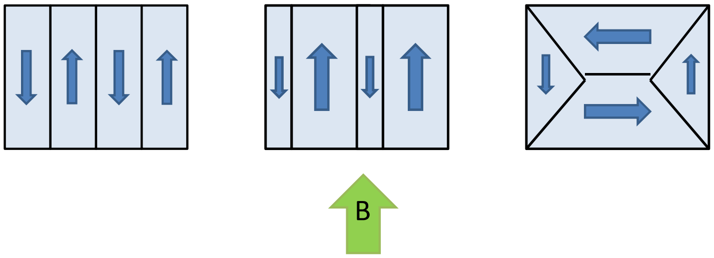


Figure 17.2: A number of possible domain structures for a ferromagnet. Left: An Ising-like ferromagnet where in each domain the moment can only point either up or down. Middle: When an external magnetic field pointing upwards is applied to this ferromagnet, it will develop a net moment by having the down-domains shrink and the up-domains expand (The local moment per atom remains constant — only the size of the domains change). Right: In this ferromagnet, the moment can point in any of the crystal axis directions.

17.1.2 Disorder Pinning

Even for single-crystal samples, disorder can play an extremely important role in the physics of domains. For example, a domain-wall can have lower energy if it passes over a defect in the crystal. To see how this occurs let us look at a domain wall in an Ising ferromagnet as shown in Fig. 17.3. All bonds are marked red where spins are antialigned rather than aligned. In both figures the domain wall starts and ends at the same points, but on the right it follows a path through a defect in the crystal — in this case a site that is missing an atom. When it intersects the location of the missing atom, the number of antialigned bonds (marked) is lower, and therefore the energy is lower. Since this lower energy makes the domain wall stick to the missing site, we say the domain wall is *pinned* to the disorder.

17.1.3 The Bloch/Néel Wall

Our discussion of domain walls so far has assumed that the spins can only point up or down — that is, the κ term in Eq. 16.3 is extremely strong. However, it often happens that this is not true – the spins would prefer to point either up or down, but there is not a huge energy penalty for pointing in other directions instead. In this case the domain wall might instead be more of a smooth rotation from spins pointing up to spins pointing down as shown on the right of Fig. 17.4. This type of smooth domain wall is known as a *Bloch wall* or *Néel wall*[5] depending on which direction the spin rotates in with respect to the direction of the domain wall itself (a somewhat subtle difference, which we will not discuss further here). The length of the domain wall (L in the

---

[5] We have already met our heros of magnetism – Felix Bloch and Louis Néel.

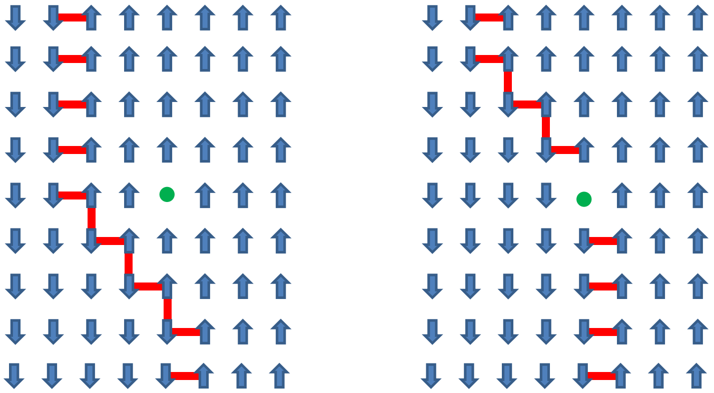


Figure 17.3: The energy of a domain wall is lower if the domain wall goes through the position of a defect in the crystal. Here, the green dot is supposed to represent a missing spin. The red bonds, where spins are anti-aligned each cost energy. When the domain wall intersects the location of the missing spin, there are fewer red bonds, therefore it is a lower energy configuration. (There are 12 red bonds on the left, but only 10 on the right).

figure, i.e., how many spins are pointing neither up nor down) is clearly dependent on a balance between the J term of Eq. 16.3 (known as the *spin stiffness*) and the κ term, the anisotropy. As mentioned above, if κ/J is very large, then the spins must point either up or down only. In this case, the domain wall is very sharp, as depicted on the left of Fig. 17.4. On the other hand, if κ/J is small, then it costs little to point the spins in other directions, and it is more important that each spin points mostly in the direction of its neighbor. In this case, the domain wall will be very fat, as depicted on the right of Fig. 17.4.

A very simple scaling argument can give us an idea of how fat the (Bloch/Néel) wall is. Let us say that the the length of the wall is N lattice constants, so $L = Na$ is the actual length of the twist in the domain wall (See Fig. 17.4). Roughly let us imagine that the spin twists uniformly over the course of these N spins, so between each spin and its neighbor, the spin twists an angle $\delta\theta = \pi/N$. The first term $-J\mathbf{S}_i \cdot \mathbf{S_j}$ in the Hamiltonian 16.3 then can be rewritten in terms of the angle between the neighbors

$$E_{one-bond} = -J\mathbf{S_i} \cdot \mathbf{S_j} = -JS^2 \cos(\theta_i - \theta_j) = -JS^2 \left(1 - \frac{(\delta\theta)^2}{2} + \ldots\right)$$

where we have used the fact that $\delta\theta$ is small to expand the cosine. Naturally, the energy of this term is minimized if the two neighboring spins are aligned, that is $\delta\theta = 0$. However, if they are not quite aligned there is an energy penalty of

$$\delta E_{one-bond} = JS^2(\delta\theta)^2/2 = JS^2(\pi/L)^2/2$$

This is the energy per bond. So the energy of the domain wall due to this spin “stiffness” is

$$\frac{\delta E_{stiffness}}{A/a^2} = NJS^2(\pi/N)^2/2$$

Figure 17.4: Left: An infinitely sharp domain wall. This would be realized if the anisotropy energy (κ) is extremely large so the spins must point either up or down (i.e., this is a true Ising system). Right: A Bloch/Néel wall (actually this depicts a Néel wall) where the spin flips continuously from up to down over a length scale L. The anisotropy energy here is smaller so that the spin can point at intermediate angle for only a small energy penalty. By twisting slowly the domain wall will pay less spin-stiffness energy.

Here we have written the energy, per unit area A of the domain wall in units of the lattice constant a.

On the other hand, in Eq. 16.3 there is a penalty κS^2 per spin when the spins are not either precisely up or down. So we can estimate the total energy due to this term to be κS^2 per spin, or a total of

$$\frac{\delta E_{anisotropy}}{A/a^2} \approx \kappa S^2 N$$

along the length of the twist.[6] Thus the total energy of the domain wall is

$$\frac{E_{tot}}{A/a^2} = JS^2(\pi^2/2)/N + \kappa S^2 N$$

This can be trivially minimized resulting in a domain wall twist having length $L = Na$ with

$$N = C_1\sqrt{(J/\kappa)} \tag{17.1}$$

---

[6] This approximation of the energy of the κ term is annoyingly crude. To be more precise, we should instead write $\kappa S^2 \cos^2(\theta_i)$ and then sum over i. Although this makes things more complicated, it is still possible to solve the problem so long as the spin twists slowly so that we can replace the finite difference $\delta\theta$ with a derivative, and replace the sum over sites with an integral. In this case, one minimizes the function

$$E = \int dx \left[JS^2(a^2/2)(d\theta(x)/dx)^2 - \kappa S^2 \cos^2\theta(x)\right]/a$$

with a the lattice constant. Using calculus of variations the minimum of this energy is given by the solution of the differential equation

$$(Ja^2/\kappa)d^2\theta/dx^2 - \sin(2\theta) = 0$$

which has a truly remarkable solution of the form

$$\theta(x) = 2\tan^{-1}\left(\exp\left[\sqrt{2}(x/a)\sqrt{\frac{\kappa}{J}}\right]\right)$$

where we once again see the same $L \sim \sqrt{J/\kappa}$ scaling. Plugging in this solution, the total energy of the domain wall becomes $E_{tot}/(A/a^2) = 2\sqrt{2}S^2\sqrt{J\kappa}$.

and a minimum domain wall energy per unit area

$$\frac{E_{tot}^{min}}{A/a^2} = C_2 S^2 \sqrt{J\kappa}$$

where C_1 and C_2 are constants of order unity (which we will not get right here considering the crudeness our approximation, but see footnote 6). As predicted, the length increases with J/κ. In many real materials the length of a domain wall can be hundreds of lattice constants.

Since the domain wall costs an energy per unit area, it is energetically unfavorable. However, as mentioned above, this energy cost is weighed against the long-range dipolar energy which tries to introduce domain walls. The more energy the domain wall costs,the larger individual domains will be (to minimize the number of domain walls). Note that if a crystal is extremely small (or, say, one considers a single crystallite within a polycrystaline sample) it can happen that the size of the crystal is much smaller than the optimum size of the domain wall twist. In this case the spins within this crystallite always stay aligned with each other.

Finally we comment that even though the actual domain wall may be hundreds of lattice constants thick, it is easy to see that these objects still have a tendency to stick to disorder as described in section 17.1.1 above.

17.2 Hysteresis in Ferromagnets

We know from our experience with electromagnetism that, ferromagnets show a hysteresis loop with respect to the applied magnetic field, as shown in Fig. 17.5. After a large external magnetic field is applied, when the field is returned to zero there remains a residual magnetization. We can now ask why this should be true. In short, it is because there is a large activation energy for changing the magnetization.

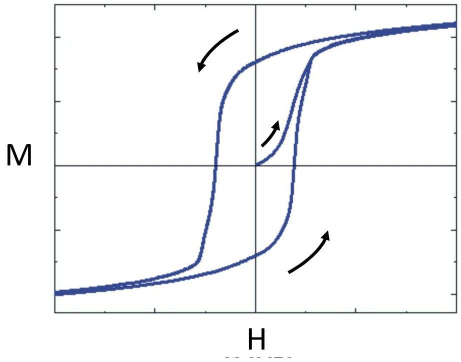


Figure 17.5: The hysteresis loop of a ferromagnet

17.2.1 Single-Domain Crystallites

For example, let us consider the case of a ferromagnet made of many small crystallites. If the crystallites are small enough then all of the moments in each crystallite point in a single direction.

(We determined in section 17.1.3 that domain walls are unfavorable in small enough crystallites). So let us imagine that all of the spins in this crystallite are locked with each other to point in the same direction. The energy of the crystallite in an external field can be written as

$$\begin{aligned} E/N &= E_0 - g\mu_B \mathbf{S}\cdot\mathbf{B} - \kappa(S^z)^2 \\ &= E_0 - g\mu_B SB\cos\theta - \kappa S^2\cos^2\theta \end{aligned} \tag{17.2}$$

where N is the number of spins in the entire crystallite, θ is the angle with respect to the $\hat{z}$ crystal axis, and we assume that the external field B is pointing along the $\hat{z}$ axis as well (although we will allow it to point either up or down). Note that we have no J term since this would just give a constant if all the spins in the crystallite are always aligned with each other.

We see that this energy is a parabola in the variable $\cos\theta$ which ranges from $+1$ to -1. The minimum of this energy is always when the spins point in the same direction as the external field (which always points in the either the $+\hat{z}$ or $-\hat{z}$ direction, corresponding to $\theta = 0$ or π). However, for small B_z the energy is not monotonic in θ. Indeed, having the spins direction is a local minimum (because the κ term favors pointing along the z-axis). This is shown schematically in Fig. 17.6. It is an easy exercise[7] to show that there will be a local minimum of the energy with the spins pointing the *opposite* direction as the applied field for $B < B_{crit}$ with

$$B_{crit} = \frac{2\kappa S}{g\mu_B}$$

So if the spins are aligned along the $-\hat{z}$ direction and a field $B < B_{crit}$ is applied in the $+\hat{z}$ direction, there is an activation barrier for the spins to flip over. Indeed, since the energy shown in Eq. 17.2 is an energy *per-spin*, the activation barrier[8] is proportional to the number of spins in the entire crystallite – i.e., it can be very large. As a result, the spins will not be able to flip until a large enough field ($B > B_{crit}$) is applied to lower the activation barrier, at which point the spins flip over. Clearly this type of activation barrier can result in hysteretic behavior as shown in Fig. 17.5.

17.2.2 Domain Pinning and Hysteresis

Domains turn out to be extremely important for determining the detailed magnetic properties of materials – and in particular for understanding hysteresis in crystals that are sufficiently large that they are not single-domain (Recall that we calculated the size L of a domain wall in Eq. 17.1. Crystals larger than this size can in principle contain a domain wall). As mentioned above, when a magnetic field is externally applied to a ferromagnet, the domain walls move to re-establish a new domain configuration (See the left two panels of Fig. 17.2) and therefore a new magnetization. However, as we discussed in section 17.1.2 above, when there is disorder in a sample, the domain walls can get pinned to the disorder: There is a low energy configuration where the domain wall intersects the disorder, and there is then an activation energy to move the domain wall. This activation energy, analogous to what we found above in section 17.2.1 results in hysteresis of the magnet.

It is frequently the case that one wants to construct a ferromagnet which retains its magnetization extremely well — i.e., where there is strong hysteresis, and even in the absence of applied

---

[7] Try showing it!

[8] In principle the spins can get over the activation barrier either by being thermally activated or by quantum tunneling. However, if the activation barrier is sufficiently large (i.e., for a large crystallite) both of these are greatly suppressed.

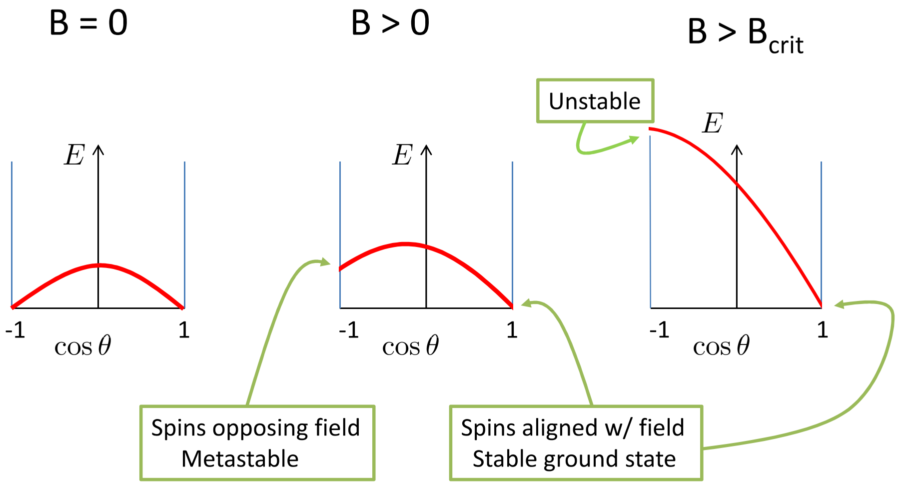


Figure 17.6: Energy of an anisotropic Ferromagnet in a magnetic field as a function of angle. Left: Due to the anisotropy, in zero field the energy is lowest if the spins point either in the $+\hat{z}$ or $-\hat{z}$ direction. When a field is applied in the $+\hat{z}$ direction the energy is lowest when the spins are aligned with the field, but there is a metastable solution with the spins pointing in the opposite direction. The spins must cross an activation barrier to flip over. Right: For large enough field, it is no longer metastable.

magnetic field there will be a large magnetization (also known as a “permanent” magnet). This is known as a “hard” magnet. It turns out that much of the trick of constructing hard-magnets is arranging to insert appropriate disorder and microstructure to strongly pin the domain walls.

17.3 Summary of Domains and Hysteresis in Ferromagnets

- Although short range interaction in ferromagnet favors all magentic moments to align, long range dipolar forces favors spins to anti-align. A compromise is reached with domains of aligned spins where different domains point in different directions. A very small crystal may be a single domain.

- The actual domain wall boundary may be a continuous rotation of the spin rather than a sudden flip over a single bond-length. The size of this spin structure depends on the ratio of the ferromagnetic energy to the anisotropy energy. (I.e., if it is very costly to have spins point in directions between up and down then the wall will be over a single bond length).

- Domain walls are lower energy if they intersect certain types of disorder in the solid. This results in the *pinning* of domain walls – they stick to the disorder.

- In a large crystal, changes in magnetization occur by changing the size of domains. In polycrystalline samples with very small crystallites, changes in magnetization occur by flipping over individual single-domain crystallites. Both of these processes can require an activation energy (domain motion requires activation energy if domain walls are pinned) and thus result in hysteretic behavior a magnetization in ferromagnets.

References

- Hook and Hall, section 8.7
- Blundell, section 6.7
- Burns, section 15.10
- Ashcroft and Mermin, end of chapter 33

Also good (but covers material in random order compared to what we want):

- Rosenberg, chapter 12
- Kittel, chapter 12

Chapter 18

Mean Field Theory

Given a Hamiltonian for a magnetic system, we are left with the theoretical task of how to predict its magnetization as a function of temperature (and possibly external magnetic field). Certainly at low temperature, the spins will be maximally ordered, and at high temperature, the spins will thermally fluctuate and will be disordered. But calculating the magnetization as a function of temperature and applied magnetic field, is typically a very hard task. Except for a few very simple exactly solvable models (like the Ising model in one dimension) we must always resort to approximations. The most important, and probably the simplest, such approximation is known as "Mean Field Theory" or "Molecular Field theory" or "Weiss Mean Field theory"[1] which we will discuss in depth in this chapter.

The general concept of mean field theory proceeds in two steps:

- First, one examines one site (or one unit cell, or small region) and treats it exactly. Any object outside the unit cell is approximated as an expectation (an average or a *mean*).
- The second step is to impose self-consistency: Every site (or unit cell, or small region) in the entire system should look the same. So the one site we treated exactly should have the same average as all of the others.

This procedure is extremely general and can be applied to problems ranging from magnetism to liquid crystal to fluid mechanics. We will demonstrate the procedure as it applies to ferromagnetism. For a homework problem we will consider how mean field theory can be applied to antiferromagnets as well (further generalizations should then be obvious).

18.1 Mean Field Equations for the Ferromagnetic Ising Model

As an example, let us consider the spin-1/2 Ising model

$$\mathcal{H} = -\frac{1}{2}\sum_{\langle i,j\rangle} J\sigma_i\sigma_j - g\mu_B B\sum_i \sigma_j$$

---

[1] The same Pierre-Ernest Weiss for whom Weiss domains are named.

where $J > 0$, and here $\sigma = \pm 1/2$ is the z-component of the spin and the magnetic field H is applied in the $\hat{z}$ direction (and as usual μ_B is the Bohr magneton). For a macroscopic system, this is a statistical mechanical system with 10^{23} degrees of freedom, where all the degrees of freedom are now coupled to each other. In other words, it looks like a hard problem!

To implement mean field theory, we focus in on one site of the problem, say, site i. The Hamiltonian for this site can be written as

$$\mathcal{H}_i = -\left(J \sum_j \sigma_j + g\mu_B B \right) \sigma_i$$

where the sum is over sites j that neighbor i. We think of the term in brackets as being caused by some effective magnetic field seen by the spin on site i, thus we define $B_{eff,i}$ such that

$$g\mu_B B_{eff,i} = J \sum_j \sigma_j + g\mu_B B$$

with again j neighboring i. Now $B_{eff,i}$ is not a constant, but is rather an operator since it contains the variables σ_j which can take several values. However, the first principle of mean field theory is that we should simply take an average of all quantities that are not site i. Thus we write the Hamiltonian of site i as

$$\mathcal{H}_i = -g\mu_B \langle B_{eff} \rangle \sigma_i$$

This is precisely the same Hamiltonian we considered when we studied paramagnetism in Eq. 15.4 above, and it is easily solvable. In short, one writes the partition function

$$Z_i = e^{-\beta g\mu_B \langle B_{eff} \rangle /2} + e^{\beta g\mu_B \langle B_{eff} \rangle /2}$$

From this we can derive the expectation of the spin on site i (compare Eq. 15.6)

$$\langle \sigma_i \rangle = \frac{1}{2} \tanh \left(\beta g\mu_B \langle B_{eff} \rangle /2 \right) \tag{18.1}$$

However, we can also write that

$$g\tilde{\mu}_B \langle B_{eff} \rangle = J \sum_j \langle \sigma_j \rangle + g\mu_B B$$

The second step of the mean field approach is to set $\langle \sigma \rangle$ to be equal on all sites of the lattice, so we obtain

$$g\mu_B \langle B_{eff} \rangle = Jz \langle \sigma \rangle + g\mu_B B \tag{18.2}$$

where z is the number of neighbors j of site i (this is known as the *coordination number* of the lattice, and this factor has replaced the sum on j). Further, again assuming that $\langle \sigma \rangle$ is the same on all lattice sites, from Eq. 18.1 and 18.2, we obtain the self-consistency equation for $\langle \sigma \rangle$ given by

$$\langle \sigma \rangle = \frac{1}{2} \tanh \left(\beta \left[Jz \langle \sigma \rangle + g\mu_B B \right] /2 \right) \tag{18.3}$$

The expected moment per site will correspondingly be given by

$$m = g\mu_B \langle \sigma \rangle \tag{18.4}$$

18.2 Solution of Self-Consistency Equation

The self-consistency equation, Eq. 18.3 is still complicated to solve. One approach is to find the solution graphically. For simplicity, let us set the external magnetic field H to zero. We then have the consistency equation

$$\langle \sigma \rangle = \frac{1}{2} \tanh\left(\frac{\beta J z}{2}\langle \sigma \rangle\right) \tag{18.5}$$

We then choose a value of the parameter $\beta Jz/2$. Let us start by choosing a value $\beta Jz/2 = 1$ that is somewhat small, i.e., a high temperature. Then in Fig. 18.1 we plot both the right hand side of Eq. 18.5 as a function of $\langle \sigma \rangle$ (in blue) and the left hand side of Eq. 18.5 (in green). Note that the left hand side is $\langle \sigma \rangle$ so the straight line is simply $y = x$. We see that there is only a single point where the two curves meet, i.e., where the left side equals the right side. This point, in this case is $\langle \sigma \rangle = 0$. From this we conclude that, for this value temperature, within mean field approximation, there is no magnetization in zero field.

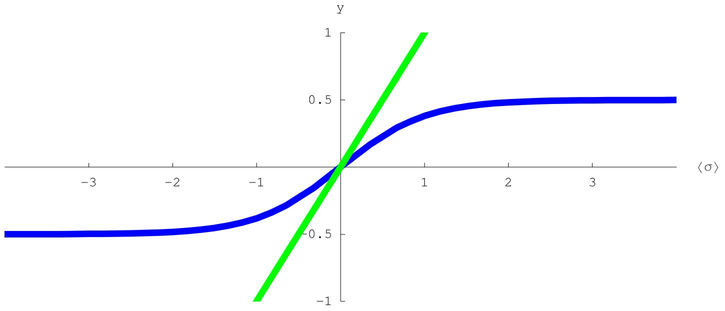


Figure 18.1: Graphical Solution of the mean field self consistency equations at relatively high temperature $\beta Jz/2 = 1$. The blue line is the tanh of Eq. 18.5. The green line is just the line $y = x$. Eq. 18.5 is satisfied only where the two curves cross – i.e., at $\langle \sigma \rangle = 0$ meaning that at this temperature, within the mean field approximation, there is no magnetization.

Let us now reduce the temperature substantially to $\beta Jz/2 = 6$. Analogously, in Fig. 18.2 we plot both the right hand side of Eq. 18.5 as a function of $\langle \sigma \rangle$ (in blue) and the left hand side of Eq. 18.5 (in green). Here, however, we see there are three possible self-consistent solutions to the equations. There is the solution at $\langle \sigma \rangle = 0$ as well as two solutions marked with arrows in the figure at $\langle \sigma \rangle \approx \pm .5$. The two nonzero solutions tell us that at low temperature this system can have nonzero magnetization even in the absence of applied field — i.e., it is ferromagnetic.

The fact that we have possible solutions with the magnetization pointing in both directions is quite natural: The Ising ferromagnet can be polarized either spin up or spin down. However, the fact that there is also a self-consistent solution with zero magnetization at the same temperature seems a bit puzzling. We will see as a homework assignment that when there are three solutions,

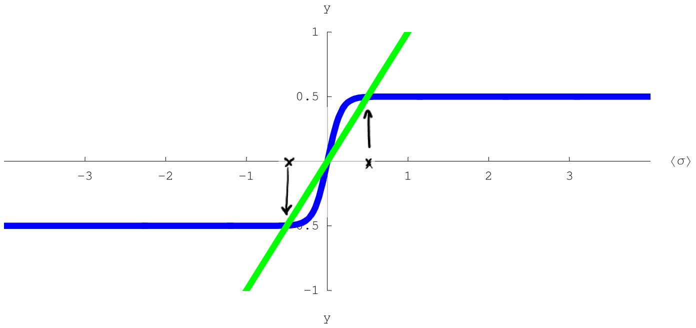


Figure 18.2: Graphical Solution of the mean field self consistency equations at relatively low temperature $\beta J z/2 = 6$. Here, the curves cross at three possible values. The fact that there is a solution of the self-consistency equations with nonzero magnetization tells us that the system is ferromagnetic (the zero magnetization solution is non-physical).

the zero magnetization solution is actually a solution of maximal free energy not minimal free energy, and therefore should be discarded[2]

Thus the picture that arises is that at high temperature the system has zero magnetization (and we will see below that it is paramagnetic) whereas at low temperature a nonzero magnetization develops and the system becomes ferromagnetic[3]. The transition between these two behaviors occurs at a temperature known as T_c, which stands for *critical temperature*[4] or *Curie temperature*[5]. It is clear from Figs. 18.1 and 18.2 that the behavior changes from one solution to three solutions precisely when the straight green line is tangent to the tanh curve, i.e., when the slope of the tanh is unity. This tangency condition thus determines the critical temperature. Expanding the tanh for small argument, we obtain the tangency condition

$$1 = \frac{1}{2}\left(\frac{\beta_c J z}{2}\right)$$

or when the temperature is

$$k_b T_c = \frac{J z}{4}$$

Using the above technique, one can solve the self-consistency equations (Eq. 18.5) at any temperature (although there is no nice analytic expression, it can be solved numerically or

---

[2] We will see (as a homework problem) that our self-consistency equations are analogous to when we find the minimum of a function by differentiation, and we may also find maxima as well.

[3] It is quite typical that at high temperatures, a ferromagnet will turn into a paramagnet, unless something else happens first – like the crystal melts.

[4] Strictly speaking it should only be called a critical temperature if the transition is second order. I.e., if the magnetization turns on continuously at this transition. For the Ising model, this is in fact true, but for some magnetic systems it is not true.

[5] Named for Pierre again.

graphically). The results are shown in Fig. 18.3. Note that at low enough temperature, all of the spins are fully aligned ($\langle\sigma\rangle = .5$ which is maximum possible for a spin-1/2). One can also, in principle, solve the self-consistency equation (Eq. 18.3) with finite magnetic field $\mathbf{B}$.

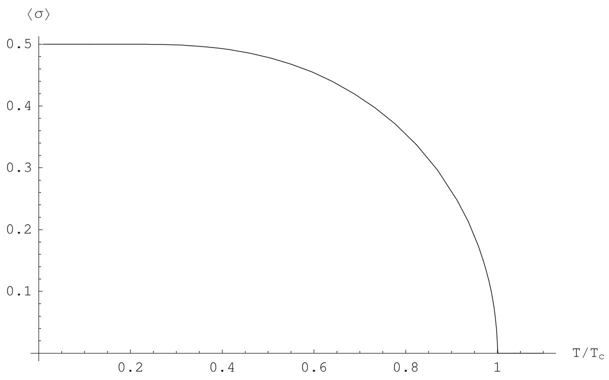


Figure 18.3: Moment per site in units of $g\mu_B$ as a function of temperature in the mean field approximation of the spin-1/2 Ising model, with zero external magnetic field applied.

18.2.1 Paramagnetic Susceptibility

At high temperature there will be zero magentization in zero externally applied field. However, at finite field, we will have a finite magnetization. Let us imagine applying a small magnetic field and solve the self-consistency equations Eq. 18.3. Since the applied field is small, we can assume that the induced $\langle\sigma\rangle$ is also small. Thus we can expand the tanh in Eq.18.3 to obtain

$$\langle\sigma\rangle = \frac{1}{2}\left(\beta\left[Jz\langle\sigma\rangle + g\mu_B B\right]/2\right)$$

Rearranging this then gives

$$\langle\sigma\rangle = \frac{\frac{1}{4}(\beta g\mu_B)B}{1 - \frac{1}{4}\beta Jz} = \frac{\frac{1}{4}(g\mu_B)B}{k_b(T - T_c)}$$

which is valid only so long as $\langle\sigma\rangle$ remains small. In terms of the moment per site is then given by (See Eq. 18.4)

$$m = g\mu_B\langle\sigma\rangle$$

which divided by the volume of a unit cell gives the magnetization M. Thus we find that the susceptibility is

$$\chi = dM/dH = \frac{\frac{1}{4}\rho(g\mu_B)^2\mu_0}{k_b(T - T_c)} = \frac{\chi_{Curie}}{1 - T_c/T} \tag{18.6}$$

where ρ is the number of spins per unit volume and χ_{Curie} is the pure Curie susceptibility of a system of (noninteracting) spin-1/2 particles (Compare Eq. 15.7). Eq. 18.6 is known as the *Curie-Weiss Law.* Thus, we see that a ferromagnet above its critical temperature is roughly a paramagnet with a slightly enhanced susceptibility. Note that the susceptibility diverges at the transition temperature when the system becomes ferromagnetic.

18.2.2 Further Thoughts

As mentioned above, the mean-field procedure is actually very general. As a homework problem we will also study the antiferromagnet. In this case, we divide the system into two sublattices — representing the two sites in a unit cell. In that example we will want to treat one spin of each sublattice exactly, but as above each spin sees only the average field from its neighbors. One can generalize even further to consider very complicated unit cells.

Aside: It is worth noting that the result of solving the Antiferromagnetic Ising model gives

$$\chi = \frac{\chi_{Curie}}{1 + T_c/T}$$

compared Eq. 18.6. It is this difference in susceptibility that pointed the way to the discovery of anitferromagnets.

The principle of mean field theory is quite general and can be applied to a vast variety of difficult problems in physics[6]. No matter what the problem, the principle remains the same — isolate some small part of the system to treat exactly, average everything outside of that small system, then demand self-consistency: that the average of the small system should look like the chosen average of the rest of the system.

While the mean field approach is merely an approximation, it is frequently a very good approximation for capturing a variety of physical phenomena. Furthermore, many of its shortcomings can be systematically improved by considering successively more corrections to the initial mean field approach[7].

18.3 Summary

- Understand the mean field theory calculation for ferromagnets. Understand how you would generalize this to any model of antiferro (homework) ferrimagnets (try it!) different spins, anisotropic models, etc etc.
- For the ferromagnet the important results of mean field theory includes:
  - a finite temperature phase transition from a low temperature ferromagnetic phase to a high temperature paramagnetic phase at a transition temperature known as the Curie temperature.
  - Above the curie temperature the paramagnetic susceptibility is $\chi = \chi_{Curie}/(1 - T_c/T)$ where χ_0 is the susceptibility of the corresponding model where the ferromagnetic coupling between sites is turned off.
  - Below T_c the magnetic moment turns on, and increases to saturation at the lowest temperature.

---

[6] In chapter 2 we already saw another example of mean field theory, when we considered the Boltzmann and Einstein models of specific heat of solids. There we considered each atom to be in a harmonic well formed by all of its neighbors. The single atom was treated exactly, whereas the neighboring atoms were treated only approximately in that their positions were essentially averaged in order to simply form the potential well – and nothing further was said of the neighbors. Another example in similar spirit was given in footnote 2 of chapter 14 where an alloy of Al and Ga with As is replaced by some averaged atom $\text{Al}_x\text{Ga}_{1-x}$ and is still considered a periodic crystal. (No unit cell is treated exactly in this case, all are replaced by the average unit cell)

[7] The motivated student might want to think about various ways one might improve mean field theory systematically. One approach is discussed in the Additional Problems.

References on Mean Field Theory

- Ibach and Luth, chapter 8 (particularly 8.6, 8.7)
- Hook and Hall, chapter 8 (particularly 8.3, 8.4)
- Kittel, Beginning of chapter 12
- Burns, section 15.5
- Ashcroft and Mermin, chapter 33

Chapter 19

Magnetism from Interactions: The Hubbard Model

So far we have only discussed ferromagnetism in the context of isolated spins on a lattice that align due to their interactions with each other. However, in fact many materials have magnetism where the magnetic moments – the aligned spins – are not pinned down, but rather can wander through the system. This phenomenon is known as *itinerant ferromagnetism*[1]. For example, it is easy to imagine a free electron gas where the number of up spins is different from the number of down spins. However, for completely free electrons it is always lower energy to have the same number of up and down spins than to have the numbers differ[2]. So how does it happen that electrons can decide, even in absence of external magnetic field, to polarize their spins? The culprit is the strong Coulomb interaction between electrons. On the other hand, we will see that Antiferromagentism can also be caused by strong interaction between electrons.

The Hubbard model[3] is an attempt to understand the magnetism that arises from interactions between electrons. It is certainly the most important model of interacting electrons in modern condensed matter physics.

The model is relatively simple to describe[4]. First we write a tight binding model for a band of electrons as we did in chapter 7 with hopping parameter t. (We can choose to do this in

---

[1] Itinerant means traveling from place to place.

[2] The total energy of having N electrons spin up in a system is proportional to $NE_F = N(N/V)^{2/d}$ where d is the dimensionality of the system (you should be able to prove this easily). We can write $E = CN^{1+a}$ with $a > 0$ and C some constant. For $N_\uparrow$ up spins and $N_\downarrow$ downspins, we have a total energy $E = CN_\uparrow^{1+a} + CN_\downarrow^{1+a} = C(N_\uparrow^{1+a} + (N - N_\uparrow)^{1+a})$ where N is the total number of electrons. Setting $dE/dN = 0$ immediately gives $N_\uparrow = N/2$ as the minimum energy configuration.

[3] John Hubbard, a British physicist, wrote down this model in 1963 and it quickly became an extremely important example in the attempt to understand interacting electrons. Despite the success of the model, Hubbard, who died relatively young in 1980, did not live to see how important his model became: In 1986, when the phenomenon of "high temperature superconductivity" was discovered by Bednorz and Müller (resulting in a Nobel prize the following year), the community quickly came to believe the an understanding of this phenomenon would only come from studying the Hubbard model. Over the next two decades the Hubbard model took on the status of being *the* most important question in condensed matter physics. Its complete solution remains elusive despite the tens of thousands of papers written on the subject. It is a shame that we do not have time to discuss superconductivity in this course.

[4] The reason most introductory books do not cover the Hubbard model is that the model is conventionally introduced using so-called "second quantized" notation — that is, using field-theoretic methods which are rather advanced. We will avoid this approach, but as a result, we cannot delve too deep into the physics of the model.

one, two, or three dimensions as we see fit[5]). We will call this Hamiltonian H_0. As we derived above (and should be easy to derive in 2 and 3d now) the full bandwidth of the band is $4dt$ in d dimensions. We can add as many electrons as we like to this band. Let us define the doping x to be the number of electrons added in the band, per site (also the $x/2$ is the fraction of k states in the band which are filled). As long as we do not fill all of the states in the band ($x < 2$), in the absence of interactions, this partially filled tight binding band is a metal. Finally we include the Hubbard interaction

$$H_{hubbard} = \sum_i U\, n_{i\uparrow}\, n_{i\downarrow} \tag{19.1}$$

where here $n_{i\uparrow}$ is the number of electrons with spin up on site i and $n_{i\downarrow}$ is the number of electrons on site i with spin down, and $U > 0$ is an energy known as the repulsive Hubbard interaction energy. This term gives an energy penalty of U whenever two electrons sit on the same site of the lattice. This short ranged interaction term is an approximate representation of the Coulomb interaction between electrons.

19.1 Ferromagnetism in the Hubbard Model

Why should this on-site interaction create magnetism? Imagine for a moment that all of the electrons in the system had the same spin state (a so-called "spin-polarized" configuration). If this were true, by the Pauli exclusion principle, no two electrons could ever sit on the same site. In this case, the expectation of the Hubbard interaction term would be zero

$$\langle \text{Polarized Spins}|H_{hubbard}|\text{Polarized Spins}\rangle = 0$$

which is the lowest possible energy that this interaction term could have. On the other hand, if we filled the band with only one spin-species, then the Fermi energy (and hence the kinetic energy of the system) would be much higher than if the electrons could be distributed between the two possible spin states. Thus, it appears that there will be some competition between the potential and kinetic energy that decides whether the spins align or not.

19.1.1 Hubbard Ferromagnetism Mean Field Theory

To try to decide quantitatively whether spins will align or not we start by writing

$$U\, n_{i\uparrow}\, n_{i\downarrow} = \frac{U}{4}(n_{i\uparrow} + n_{i\downarrow})^2 - \frac{U}{4}(n_{i\uparrow} - n_{i\downarrow})^2$$

Now we make the approximation of treating all operators $n_{i,\uparrow}$ and $n_{i,\downarrow}$ as their expectations.

$$U\, n_{i\uparrow}\, n_{i\downarrow} \approx \frac{U}{4}\langle n_{i\uparrow} + n_{i\downarrow}\rangle^2 - \frac{U}{4}\langle n_{i\uparrow} - n_{i\downarrow}\rangle^2$$

This type of approximation is a type of mean-field theory, similar to that we encountered in the previous chapter 18: We replace operators by their expectations.[6] The expectation $\langle n_{i\uparrow} + n_{i,\downarrow}\rangle$ in

---

[5] In one dimension, the Hubbard model is exactly solvable.

[6] This is a slightly different type of mean field theory from that encountered in chapter 18. Previously we considered some local degree of freedom (some local spin) which we treated exactly, and replaced all other spins by their average. Here, we are going to treat the kinetic energy term exactly, but replace the operators in the potential energy term by their averages.

the first term is just the average number of electrons on site i which is just the average number of particles per site[7] , which is equal to the doping x which we keep fixed.

Correspondingly, the second expectation, $\langle n_{i\uparrow} - n_{i\downarrow}\rangle$, is related to the magnetization of the system. In particular, since each electron carries[8] a magnetic moment of μ_B, the magnetization[9] is

$$M = (\mu_B/v)\langle n_{i\uparrow} - n_{i\downarrow}\rangle$$

with v the volume of the unit cell. We thus see that the energy of the expectation of the Hubbard interaction is given by

$$\langle H_{hubbard}\rangle \approx (V/v)(U/4)\left(x^2 - (Mv/\mu_B)^2\right) \tag{19.2}$$

where V/v is the number of sites in the system. Thus, as expected, increasing the magnetization M decreases the expectation of the interaction energy. To determine if the spins actually polarize we need to weigh this potential energy gain against the potential energy cost.

19.1.2 Stoner Criterion

Here we calculate the kinetic energy cost of polarizing the spins in our model and we balance this against the potential energy gain. We will recognize this calculation as being almost identical to the calculation we did way back in section 3.2.3 when we studied Pauli paramagnetism (but we repeat it here for clarity).

Consider a system of fermions (at zero temperature for simplicity) with the same number of spin up and spin down electrons. Let $g(E_F)$ be the total density of states at the Fermi surface per unit volume (for both spins put together). Now, let us flip over a small number of spins so that the spin up and spin down fermi surfaces have slightly different energies[10].

$$\begin{aligned} E_{F,\uparrow} &= E_F + \delta\epsilon/2 \\ E_{F,\downarrow} &= E_F - \delta\epsilon/2 \end{aligned}$$

The difference in the number density of up and down electrons is then

$$\rho_\uparrow - \rho_\downarrow = \int_0^{E_F+\delta\epsilon/2} dE\, \frac{g(E)}{2} - \int_0^{E_F-\delta\epsilon/2} dE\, \frac{g(E)}{2}$$

where we have used the fact that the density of states per unit volume for either the spin-up or spin-down species is $\frac{g(E)}{2}$.

The difference in the number of up and down electrons is related to the magnetization of the system by[8]

$$M = \mu_B(\rho_\uparrow - \rho_\downarrow)$$

Similarly, we can write the total kinetic energy per unit volume as

$$K = \int_0^{E_F+\delta\epsilon/2} dE\, E\, \frac{g(E)}{2} + \int_0^{E_F-\delta\epsilon/2} dE\, E\, \frac{g(E)}{2}$$

---

[7] This assumes that the system remains homogeneous – that is, that all sites have the same average number of electrons.

[8] Assuming an electron g-factor of $g = 2$ and an electron spin of 1/2.

[9] Recall magnetization is moment per unit volume.

[10] If we were being very careful we would adjust E_F to keep the overall electron density $\rho_\uparrow + \rho_\downarrow$ fixed as we change $\delta\epsilon$. For small $\delta\epsilon$ E_F remains unchanged as we change $\delta\epsilon$ but this is not true for larger $\delta\epsilon$.

Although we could carry forward at this point and try to perform the integrals generally for arbitrary $\delta\epsilon$ (indeed we will have a homework problem on this) it is enough for our present discussion to consider the simpler case of very small $\delta\epsilon$. In this case, we have

$$\rho_\uparrow - \rho_\downarrow = \delta\epsilon \frac{g(E)}{2}$$

so

$$M = \mu_b \delta\epsilon \frac{g(E)}{2}$$

The kinetic energy per unit volume, is a bit more tricky. We write

$$\begin{aligned} K &= 2\int_0^{E_F} dE\ E\ \frac{g(E)}{2} + \int_{E_F}^{E_F+\delta\epsilon/2} dE\ E\ \frac{g(E)}{2} - \int_{E_F-\delta\epsilon/2}^{E_F} dE\ E\ \frac{g(E)}{2} \\ &\approx K_{M=0} + \frac{g(E_F)}{2}\left[\left(\frac{(E_F+\delta\epsilon/2)^2}{2} - \frac{E_F^2}{2}\right) - \left(\frac{E_F^2}{2} - \frac{(E_F-\delta\epsilon/2)^2}{2}\right)\right] \\ &= K_{M=0} + \frac{g(E_F)}{2}(\delta\epsilon/2)^2 \\ &= K_{M=0} + \frac{g(E_F)}{2}\left(\frac{M}{\mu_B g(E_F)}\right)^2 \end{aligned} \tag{19.3}$$

Where $K_{M=0}$ is the kinetic energy per unit volume for a system with no net magentization (even numbers of spin-up and spin-down electrons).

We can now add this result to Eq. 19.2 to give the total energy of the system per unit volume

$$E_{tot} = E_{M=0} + \left(\frac{M}{\mu_B}\right)^2 \left[\frac{1}{2g(E_F)} - vU\right]$$

with v the volume of the unit cell. We thus see that for

$$U > 2g(E_F)/v$$

the energy of the system is lowered by increasing the magnetization from zero. This condition for itinerant ferromagnetism is known as the *Stoner criterion*[11].

Aside: We did a lot of work to arrive at Eq. 19.3. In fact, we could have almost have written it down with no work at all based on the calculation of the Pauli susceptibility we did back in section 3.2.3. Recall first that when an external magnetic field is applied in the up direction to a system, there is an energy induced from the coupling of the spins to the field which is given by $\mu_B(\rho_\uparrow - \rho_\downarrow)B = -MB$ (with positive M being defined in the same direction as positive B so that having the two aligned is low energy). Also recall in section 3.2.3 that we derived the (Pauli) susceptibility of an electron system is

$$\chi_{Pauli} = \mu_0 \mu_B^2 g(E_F)$$

which means that when a magnetic field B is applied, a magnetization $\chi_{Pauli}B/\mu_0$ is induced. Thus we can immediately conclude that the energy of such a system in an external field *must* be of the form

$$E(M) = \frac{M^2\mu_0}{2\chi_{Pauli}} - MB$$

---

[11] Edmund Stoner was a British physicist who, among other things, figured out the Pauli Exclusion principle in 1924 a year before Pauli. However, Stoner's work focused on the spectra, and behavior, of atoms, and he was not bold enough to declare the exclusion was a fundamental property of electrons. Stoner was diagnosed with diabetes in 1919 at 20 years of age and grew progressively weaker for the next eight years. In 1927 Insulin treatment became available, saving his life. He died in 1969.

To see that this is correct, we minimize the energy with respect to M at a given B and we discover that this properly gives us $M = \chi_{Pauli} B/\mu_0$. Thus, at zero applied B, the energy should be

$$E(M) = \frac{M^2 \mu_0}{2\chi_{Pauli}} = \frac{M^2}{2\mu_B^2 g(E_F)}$$

exactly as we found in Eq. 19.3!

19.2 Mott Antiferromagnetism in the Hubbard Model

In fact, the Hubbard model is far more complex than the above mean field calculation would lead one to believe. Let us now consider the case where the doping is such that there is exactly one electron per site of the lattice. For noninteracting electrons, this would be a half-filled band, and hence a conductor. However, if we turn on the Hubbard interaction with a large U the system becomes an insulator. To see this, imagine one electron sitting on every site. In order for an electron to move, it must hop to a neighboring site which is already occupied. This process therefore costs energy U, and if U is large enough, it cannot happen. This is precisely the physics of the Mott insulator which we discussed above in section 12.3.

With one immobile electron on each site we can now ask which way the spins align. For a square or cubic lattice, there are two obvious options: either the spins want to be aligned with their neighbors or they want to be anti-aligned with their neighbors (ferromagnetism or antiferromagnetism). It turns out that antiferromagnetism is favored! To see this, consider the antiferromagnetic state $|GS_0\rangle$ shown on the left of Fig. 19.1. In the absence of hopping this state is an eigenstate with zero energy (as is any other state where there is precisely one electron on each site). We then consider adding the hopping perturbatively. Because the hopping hamiltonian allows an electron to hop from site to site (with hopping amplitude $-t$), the electron can make a "virtual" hop to a neighboring site, as shown in the right of Fig. 19.1. The state on the right $|E\rangle$ is of higher energy, (in the absence of hopping it has energy U because of the double occupancy). Using second order perturbation theory we obtain

$$\begin{aligned} E(|GS_0\rangle + \text{hopping}) &= E(|GS_0\rangle) + \sum_E \frac{|\langle E|H_{hop}|GS_0\rangle|^2}{E_{GS_0} - E_E} \\ &= E(|GS_0\rangle) - \frac{NZ|t|^2}{U} \end{aligned}$$

In the first line the sum is over all $|E\rangle$ states that can be reached in a single hop from the state $|GS_0\rangle$. In the second line, we have counted the number of such terms to be NZ where Z is the coordination number (number of nearest neighbors) and N is the number of sites. Further we have inserted $-t$ for the amplitude of hopping from one site to the next. Note that if the spins were all aligned, no virtual intermediate state $|E\rangle$ could exist since it would violate the Pauli exclusion principle (hopping of electrons conserves spin state, so spins cannot flip over during a hop, so there is strictly no double occupancy). Thus we conclude that the antiferromagnetic state has its energy lowered compared to the ferromagnetic state in the limit of large U in a Mott insulating phase.

Admittedly the above argument appears a bit handwaving (It is correct though!). To make the argument more precise, one should be much more careful about how one represents states with multiple electrons. This typically requires field theoretic techniques. A very simple example of how this is done (without more advanced techniques) is presented in the appendix to this chapter.

Nonetheless, the general physics of why the antiferromagnetic Mott insulator state should be lower energy than its ferromagnetic counterpart can be understood qualitatively without resorting

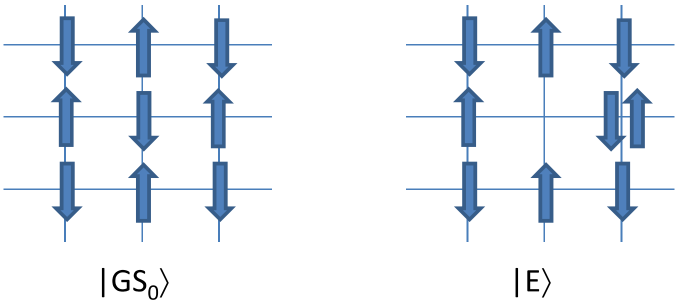


Figure 19.1: Left: The proposed antiferromagnetic ground state in the limit that t is very small. Right: A higher energy state in the limit of small t which can occur by an electron from one site hopping onto a neighboring site. The energy penalty for double occupancy is U.

to the more precise arguments. On each site one can think of an electron as being confined by the interaction with its neighbors to that site. In the ferromagnetic case, the electron cannot make any excursions to neighboring sites because of the Pauli exclusion principle (these states are occupied). However, in the antiferromagnetic case, the electron can make excursions, and even though when the electron wanders onto neighboring sites, the energy is higher, there is nonetheless some amplitude for this to happen.[12] Allowing the electron wavefunction to spread out always lowers its energy[13].

Indeed, in general a Mott insulator (on a square or cubic lattice) is typically an antiferromagnet (unless some other interesting physics overwhelms this tendency). It is generally believed that there is a substantial range of t, U and doping x where the ground state is antiferromagnetic. Indeed, many real materials are thought to be examples of antiferromagnetic Mott insulators. Interestingly, it turns out that in the limit of very very strong on-site interaction $U \to \infty$ adding even a single additional hole to the half-filled Mott insulator will turn the Mott antiferromagnet into a ferromagnet! This rather surprising result due to Nagaoka and Thouless[14] (one of the few key results about the Hubbard model which is known as a rigorous theorem) shows the general complexity of this model.

19.3 Summary of Hubbard Model

- Hubbard model includes tight-binding hopping t and on-site "Hubbard" interaction U

---

[12] Similar to when a particle is in a potential well $V(x)$, there is some amplitude to find the electron at a position such that $V(x)$ is very large.

[13] By increasing Δx we can decrease Δp and thus lower the kinetic energy of the particle, as per the Heisenberg uncertainty principle.

[14] David Thouless, born in Scotland, is one of the most prominent names in modern condensed matter physics. He has not yet won a Nobel prize, but he is frequently mentioned as a high contender. Yosuki Nagaoka is a prominent Japanese theorist.

- For partially filled band, the repulsive interaction (if strong enough) makes the system an (itinerant) ferromagnet: aligned spins can have lower energy because they do not double occupy sites, and therefore are lower energy with respect to U although it costs higher kinetic energy to align all the spins.

- For a half-filled band, the repulsive interaction makes the Mott insulator antiferromagnetic: virtual hopping lowers the energy of anti-aligned neighboring spins.

References on Hubbard Model

Unfortunately there are essentially no references that I know of that are readable without background in field theory and second quantization.

19.4 Appendix: The Hubbard model for the Hydrogen Molecule

Since my above perturbative calculation showing antiferromagnetism is very hand-waving, I thought it useful to do a real (but very simple) calculation showing how, in principle, these calculations are done more properly. This appendix is certainly nonexaminable, but if you are confused about the above discussion of antiferromagnetism in the Hubbard model, this appendix might be enlightening to read.

The calculation given here will address the Hubbard model for the Hydrogen molecule. Here we consider two nuclei A and B near each other, with a total of two electrons – and we consider only the lowest spatial orbital (the s-orbital) for each atom[15]. There are then four possible states which an electron can be in:

$$A\uparrow \qquad A\downarrow \qquad B\uparrow \qquad B\downarrow$$

To indicate that we have put electron 1 in, say the $A\uparrow$ state, we write the wavefunction

$$|A\uparrow\rangle \quad\longleftrightarrow\quad \varphi_{A\uparrow}(1)$$

(Here φ is the wavefunction (1) is shorthand for the position $\mathbf{r}_1$ as well as the spin S_1 coordinate).

For a two electron state, we are only allowed to write wavefunctions that are overall antisymmetric. So given two single electron orbitals α and β (α and β take values in the four possible orbitals $A\uparrow, A\downarrow, B\uparrow, B\downarrow$) we write so-called Slater determinants to describe the antisymmetric two particle wavefunctions

$$|\alpha;\beta\rangle = \frac{1}{\sqrt{2}}\det\begin{vmatrix} \alpha(1) & \beta(1) \\ \alpha(2) & \beta(2) \end{vmatrix} = (\alpha(1)\beta(2) - \beta(1)\alpha(2))/\sqrt{2} = -|\beta;\alpha\rangle$$

Note that this slater determinant can be generalized to write a fully antisymmetric wavefunction for any number of electrons. Note also that if the two orbitals are the same, then the wavefunction vanishes (as it must by Pauli exclusion).

---

[15] This technique can in principle be used for any number of electrons in any number of orbitals although exact solution becomes difficult as the Schroedinger matrix becomes very high dimensional and hard to diagonalize exactly — although approximations remain possible.

For our proposed model of the Hydrogen molecule, we thus have six possible states for the two electrons

$$
\begin{aligned}
|A\uparrow; A\downarrow\rangle &= -|A\downarrow; A\uparrow\rangle \\
|A\uparrow; B\uparrow\rangle &= -|B\uparrow; A\uparrow\rangle \\
|A\uparrow; B\downarrow\rangle &= -|B\downarrow; A\uparrow\rangle \\
|A\downarrow B\uparrow\rangle &= -|B\uparrow; A\downarrow\rangle \\
|A\downarrow; B\downarrow\rangle &= -|B\downarrow; A\downarrow\rangle \\
|B\uparrow; B\downarrow\rangle &= -|B\downarrow; B\uparrow\rangle
\end{aligned}
$$

The Hubbard interaction energy (Eq. 19.1) is diagonal in this basis – it simply gives an energy penalty U when there are two electrons on the same site. We thus have

$$\langle A\uparrow; A\downarrow|H_{hubbard}|A\uparrow; A\downarrow\rangle = \langle B\uparrow; B\downarrow|H_{hubbard}|B\uparrow; B\downarrow\rangle = U$$

and all other matrix elements are zero.

To evaluate the hopping term we refer back to where we introduced tight binding in section 4.3.2 and chapter 7. Analogous to there, we see that the hopping term with amplitude $-t$ turns an $A\uparrow$ orbital into a $B\uparrow$ orbital or vice versa, and similarly turns a $A\downarrow$ into a $B\downarrow$ and vice versa (The hopping does not change the spin). Thus, for example, we have

$$\langle A\downarrow; B\uparrow|H_{hop}|A\downarrow; A\uparrow\rangle = -t$$

where here the hopping term turned the B into an A. Note that this implies similarly that

$$\langle A\downarrow; B\uparrow|H_{hop}|A\uparrow; A\downarrow\rangle = t$$

since $|A\downarrow; A\uparrow\rangle\rangle = -|A\uparrow; A\downarrow\rangle$.

Since there are six possible basis states, our most general Hamiltonian can be expressed as a six by six matrix. We thus write our Schroedinger equation as

$$
\begin{pmatrix}
U & 0 & -t & t & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
-t & 0 & 0 & 0 & 0 & -t \\
t & 0 & 0 & 0 & 0 & t \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -t & t & 0 & U
\end{pmatrix}
\begin{pmatrix}
\psi_{A\uparrow A\downarrow} \\ \psi_{A\uparrow B\uparrow} \\ \psi_{A\uparrow B\downarrow} \\ \psi_{A\downarrow B\uparrow} \\ \psi_{A\downarrow B\downarrow} \\ \psi_{B\uparrow B\downarrow}
\end{pmatrix}
= E
\begin{pmatrix}
\psi_{A\uparrow A\downarrow} \\ \psi_{A\uparrow B\uparrow} \\ \psi_{A\uparrow B\downarrow} \\ \psi_{A\downarrow B\uparrow} \\ \psi_{A\downarrow B\downarrow} \\ \psi_{B\uparrow B\downarrow}
\end{pmatrix}
$$

where here we mean that the full wavefunction is the sum

$$
\begin{aligned}
|\Psi\rangle &= \psi_{A\uparrow A\downarrow}|A\uparrow; A\downarrow\rangle + \psi_{A\uparrow B\uparrow}|A\uparrow; B\uparrow\rangle + \psi_{A\uparrow B\downarrow}|A\uparrow; B\downarrow\rangle \\
&+ \psi_{A\downarrow B\uparrow}|A\downarrow; B\uparrow\rangle + \psi_{A\downarrow B\downarrow}|A\downarrow; B\downarrow\rangle + \psi_{B\uparrow B\downarrow}|B\uparrow; B\downarrow\rangle
\end{aligned}
$$

We note immediately that the Hamiltonian is block diagonal. We have eigenstates

$$|A\uparrow; B\uparrow\rangle \qquad |A\downarrow; B\downarrow\rangle$$

both with energy $E = 0$ (hopping is not allowed and there is no double occupancy so no Hubbard interaction either). The remaining four by four Shroedinger equation is then

$$
\begin{pmatrix}
U & t & -t & 0 \\
t & 0 & 0 & t \\
-t & 0 & 0 & -t \\
0 & t & -t & U
\end{pmatrix}
\begin{pmatrix}
\psi_{A\uparrow A\downarrow} \\ \psi_{A\uparrow B\downarrow} \\ \psi_{A\downarrow B\uparrow} \\ \psi_{B\uparrow B\downarrow}
\end{pmatrix}
= E
\begin{pmatrix}
\psi_{A\uparrow A\downarrow} \\ \psi_{A\uparrow B\downarrow} \\ \psi_{A\downarrow B\uparrow} \\ \psi_{B\uparrow ; B\downarrow}
\end{pmatrix}
$$

We find one more eigenvector $\propto (0,1,1,0)$ with energy $E=0$ corresponding to the state[16]

$$\frac{1}{\sqrt{2}}\left(|A\uparrow;B\downarrow\rangle + |A\downarrow;B\uparrow\rangle\right)$$

The remaining eigenstates are more complicated, and have energies U and $\frac{1}{2}\left(U \pm \sqrt{U^2+16t^2}\right)$. The ground state, always has energy

$$E_{ground} = \frac{1}{2}\left(U - \sqrt{U^2+16t^2}\right)$$

In the limit of t/U becoming zero, the ground state wavefunction becomes very close to

$$\frac{1}{\sqrt{2}}\left(|A\uparrow;B\downarrow\rangle - |A\downarrow;B\uparrow\rangle\right) + \mathcal{O}(t/U) \tag{19.4}$$

with amplitudes of order t/U for the two electrons to be on the same site. In this limit the energy goes to

$$E_{ground} = -4t^2/U$$

which is almost in agreement with our above perturbative calculation — the prefactor differs from that mentioned in the above calculation by a factor of 2. The reason for this discrepancy is that the ground state is not just ↑ on one site and ↓ on the other, but rather a superposition between the two. This superposition can be thought of as a (covalent) chemical bond (containing two electrons) between the two atoms.

In the opposite limit, $U/t \to 0$ the ground state wavefunction for a single electron is the symmetric superposition $(|A\rangle + |B\rangle)/\sqrt{2}$ (see section 4.3.2) assuming $t>0$. This is the so-called "bonding" orbital. So the ground state for two electrons is just the filling of this bonding orbital with both spins — resulting in

$$\begin{aligned}\frac{|A\uparrow\rangle + |B\uparrow\rangle}{\sqrt{2}} \otimes \frac{|A\downarrow\rangle + |B\downarrow\rangle}{\sqrt{2}} &= \frac{1}{2}\left(|A\uparrow;A\downarrow\rangle + |A\uparrow;B\downarrow\rangle + |B\uparrow;A\downarrow\rangle + |B\uparrow;B\downarrow\rangle\right) \\ &= \frac{1}{2}\left(|A\uparrow;A\downarrow\rangle + |A\uparrow;B\downarrow\rangle - |A\downarrow;B\uparrow\rangle + |B\uparrow;B\downarrow\rangle\right)\end{aligned}$$

Note that eliminating the double occupancy states (simply crossing them out)[17] yields precisely the same result as Eq. 19.4. Thus, as the interaction is turned on it simply suppresses the double occupancy in this case.

---

[16] The three states with $E=0$ are in fact the $S_z=-1,0,1$ states of $S=1$. Since the Hamiltonian is rotationally invariant, these all have the same energy.

[17] Eliminating doubly occupied orbitals by hand is known as *Gutzwiller projection* (after Martin Gutzwiller) and is an extremely powerful approximation tool for strongly interacting systems.

Chapter 20

Magnetic Devices

This is the chapter on magnetic devices. It is *NONEXAMINABLE* It is also *NONFINISHED*. I hope to finish this soon!

Indices

These notes have two indices[1].

In the index of people, Nobel laureates are marked with *. There are over 50 of them. To be fair, a few of these Nobel laureates (such as Fredrick Sanger) are mentioned offhandedly in these notes, but for random cultural reasons having little to do with the content of this course. Nonetheless, the total count of Nobel laureates is easily over 50. (A few stray celeb pop stars got into the index as well).

A few people whose names are mentioned did not end up in the index because the use of their name is so common that it is not worth indexing them as people as well. A few examples are Coulomb's law; Fourier transform; Boltzmann's constant; Taylor expansion; Hamiltonian; Jacobian, and so forth. But then again, I did index Schroedinger Equation and Fermi Statistics under Schroedinger and Fermi respectively. So I'm not completely consistent. So sue me.

The index of topics was much more difficult to put together. It was hard to figure out what the most useful division of topics is to put in the index. I tried to do this so that the Index would be maximally useful - but I'm not sure how good a job I did. Most book indices are not very useful, and now I know why — it is hard to predict why a reader is going to want to look something up.

---

[1] Making it a tensor. har har.

Index of People

Index of Topics