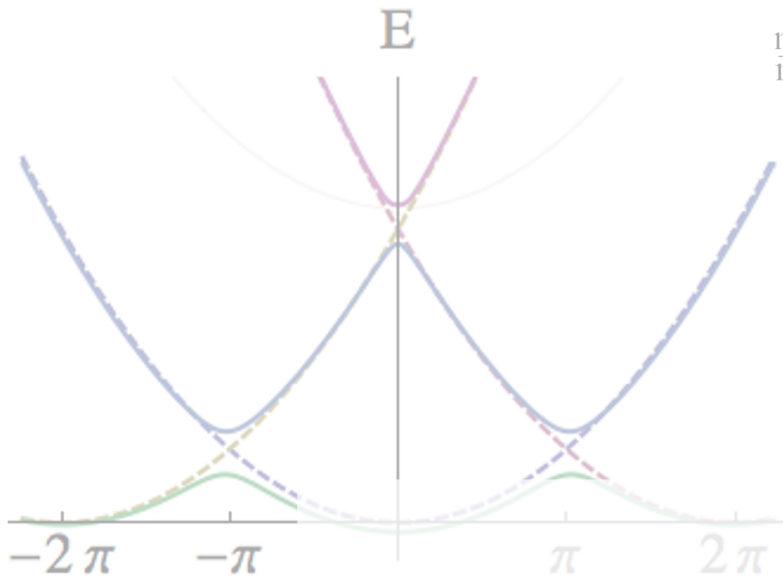
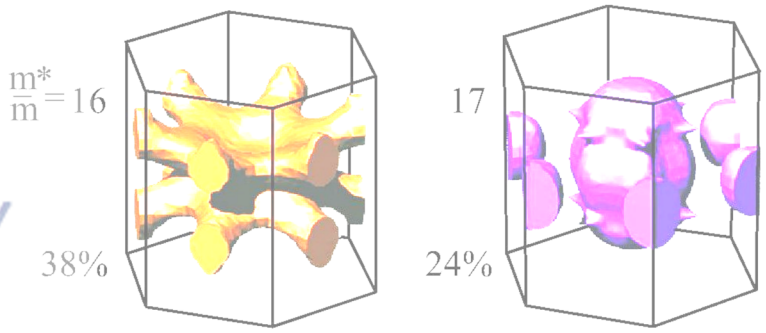
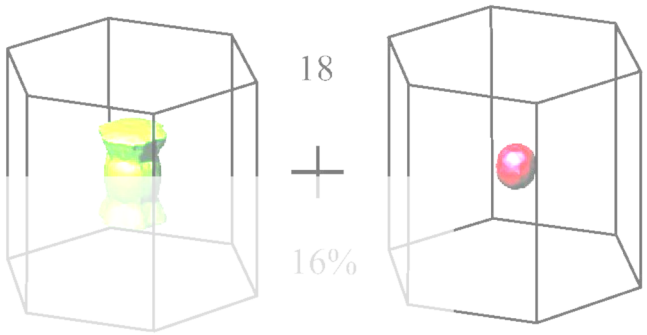
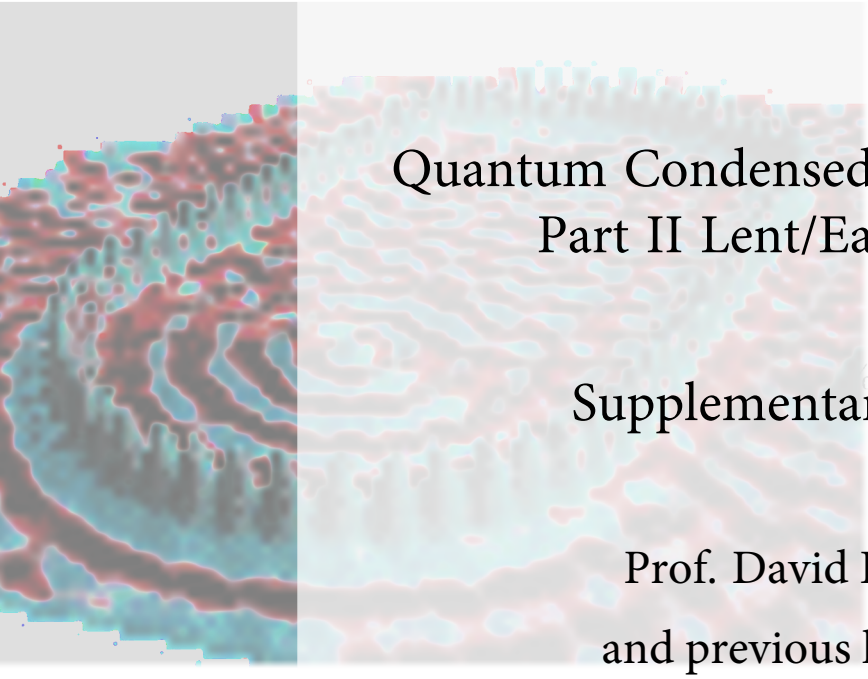








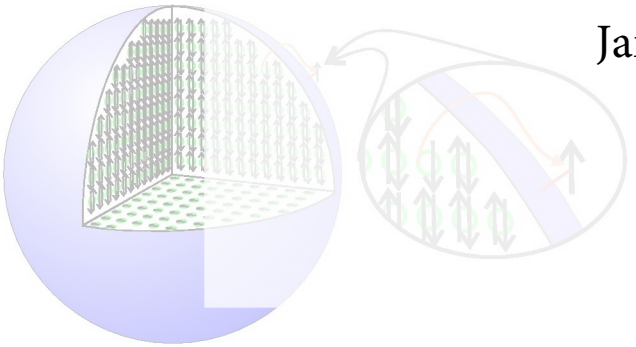
Quantum Condensed Matter Physics
Part II Lent/Easter 2021

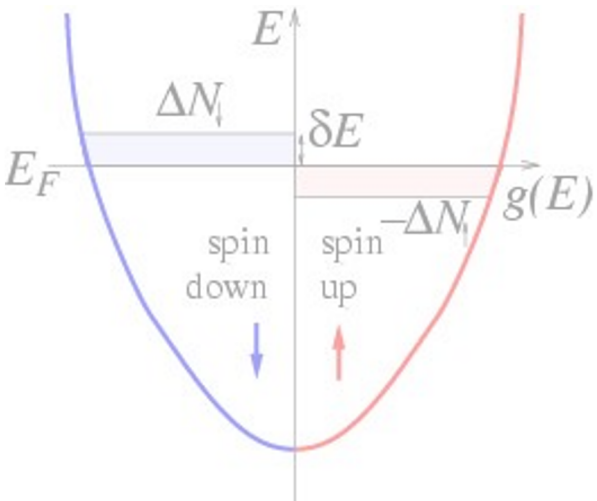
Supplementary notes

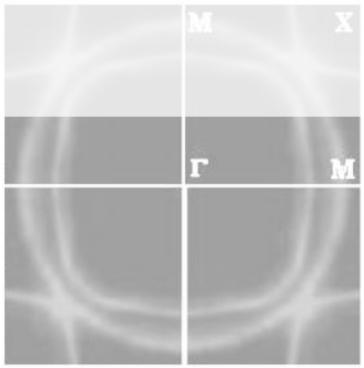
Prof. David Ritchie

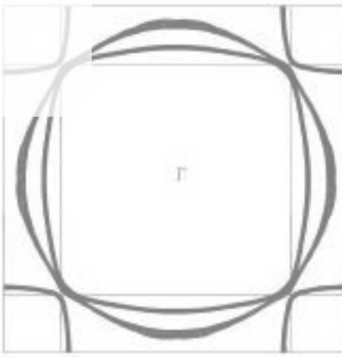
and previous lecturers

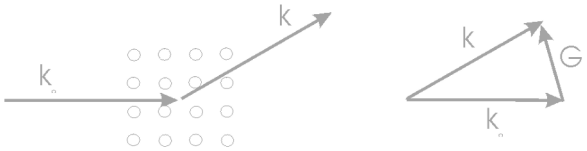
January 2021











Contents

Preface

Books

There are many good books on solid state and condensed matter physics, but the subject is rich and diverse enough that each of these contains both much more and much less than the topics covered in this course. The two classic textbooks are Kittel, and Ashcroft and Mermin. They are both at the correct level for the course, and have the virtue of clear exposition, many examples, and lots of experimental data. Slightly more concise, though in places a little more formal is Ziman. Grosso and Parravicini has a somewhat wider coverage of material, but much of it goes well beyond the level of detail required for this course. Marder is at about the right level (though again with more detail than we shall need), and has a nice blend of quantum properties with statistical and classical properties. A well illustrated modern treatment of most topics in this course is also given by Ibach and Lüth. OUP have recently issued a series of short texts in condensed matter physics. They are more detailed than needed for this course, but are quite accessible and excellent for reference. **The most relevant for this course is Singleton.**

- C. Kittel, *Introduction to Solid State Physics, 7th edition*, Wiley, NY, 1996.
- N. W. Ashcroft and N.D.Mermin, *Solid State Physics*, Holt-Saunders International Editions, 1976.
- J. M. Ziman, *Principles of the Theory of Solids*, CUP, Cambridge, 1972.
- H. Ibach and H. Lüth, *Solid State Physics*, Springer 1995.
- J. Singleton, *Band Theory and the Electronic Properties of Solids*, OUP 2001.
- M. P. Marder, *Condensed Matter Physics*, Wiley, NY, 2000. Covers both quantum matter and mechanical properties.
- G. Grosso and G. P. Parravicini, *Solid State Physics*, AP, NY, 2000. A wide coverage of material, very bandstructure oriented, very detailed.
- A very good book, though with a focus on statistical and "soft" condensed matter that makes it less relevant for this course, is
  P. M. Chaikin and T. Lubensky, *Principles of Condensed Matter Physics*, CUP, Cambridge, 1995.

These notes

Treat these notes with caution. If you are looking for text book quality, then you should look at text books.

Polished and optimised treatments of most of the topics covered here have been published in a number of excellent books, listed above. For much of its duration the course follows the book by Singleton, and where it does not, the books by Ashcroft&Mermin and Kittel give excellent support. Reading up in text books is not only useful revision of the lecture material, it also

provides important background information and context which, however, is outside the scope of the lectures. Now that you have made it into the second half of the third year, you should really give it a try!

What is the purpose of these notes, then? (i) They may help you to fill in some gaps or make corrections to your personal lecture notes, when the lecture moved too quickly to keep accurate notes. (ii) They may contain some information which was only superficially touched on during the lectures but not explicitly written down. (iii) They may contain some alternative approaches, which were not given in the lectures but which may be interesting to know, as understanding comes from combining and reconciling many approaches to the same topic. The notes have a *complementary* function. Do not attempt to learn condensed matter physics from these notes alone. The lectures will be presented using more qualitative and physical descriptions. In some places, treatments given in the lectures are simpler and more direct those you will find in the notes. The lecture overheads together with your personal notes from the lectures form the backbone of this course – and where they do not suffice, text books and these notes may help.

Notation

Do not be confused if wavevector or frequency dependencies in these notes are sometimes expressed in terms of sub-scripts and sometimes as function arguments. For example $\epsilon_\omega = \epsilon(\omega)$, and $V_{\mathbf{q}} = V(\mathbf{q})$. The two notations are used interchangeably in this handout.

Outline

"Any new discovery contains the germ of a new industry." *(J. J. Thomson)*

In this course, we will take a first step towards working with electrons in solids. This requires skills you have acquired in other courses, in particular in electromagnetism, quantum mechanics and statistical physics.

The course is structured as follows: roughly the first half of Lent term consists of a progression of more and more sophisticated models, each fixing the deficiencies of its predecessor. We will start by considering electrons in insulators (Lorentz oscillator model) and metals (Drude model) by very intuitive classical approaches that get quite a number of things right. We will then introduce quantum statistics (Sommerfeld model) to correct those things which go most spectacularly wrong in the classical approach. Next, we will turn to the crystalline lattice, which the Sommerfeld model does not take into account, and ponder its symmetry properties and its vibrations. Arguably the toughest part of the course is Bloch's theorem and the calculation of electronic energy levels (or band structure) in the presence of a periodic potential arising from the lattice of atoms. This is done in two ways, (i) using the nearly free electron gas approach, and (ii) using linear combinations of atomic orbitals (or tight binding). Having reached this point, it is downhill again, considering band structures of real materials and how band structures can be determined experimentally. Lent term will conclude with an introduction to how all of this can be applied to build semiconductor-based devices.

The models discussed up to this point in the course neglect correlations between the electrons: having computed the single-electron energy eigenstates of a solid, we then fill up these states with the available electrons, ignoring the mutual interaction between the electrons. We will apply these models in order to understand semiconductor-based electronic devices such as *p-n* junction based diodes, field effect transistors and solar cells.

It is surprising but true that a reasonable description of many phenomena involving electrons in materials can be obtained while neglecting the (strong repulsive) interactions between electrons. Metals can be regarded as *dense electron liquids*, and we find that in many metals, the strong electron-electron interaction lead to new *collective phenomena*. The most prominent examples are magnetism, of which there are many varieties, and superconductivity, but there are many other possible forms of electronic quantum self-organisation. We will take a first step towards understanding the driving force behind such collective states by considering charge and spin instabilities in metals, and we will discuss a very general conceptual framework for working with strongly interacting fermionic systems, in many ways the standard model of condensed matter physics, namely Landau's Fermi liquid theory.

Chapter 1

Classical models for electrons in solids

We start by considering intuitive, classical models for condensed matter, which were first put forward in the late 19th century. These models are the Lorentz, or dipole oscillator model, in which the oscillations of charges around their average positions is incorporated, and the Drude model, in which electrons are treated similarly to an ideal gas.

As the shortcomings of these models become apparent, we will progress to more sophisticated models. In particular, we find that electrons form a degenerate quantum gas, in which states are occupied according to the Pauli exclusion principle. This leads to the Sommerfeld model of the free electron gas. Next, we consider the effect of the periodic potential produced by the lattice atoms on the electronic states. This leads to the full electronic band structure and the possibility of producing energy gaps. Finally, in the Easter term, we begin to include the effects of interactions between the electrons.

For now, however, we concentrate on how the electromagnetic response of an insulator is modelled using classical physics ...

1.1 Lorentz oscillator model

We consider the effect of incoming electromagnetic waves on the charges present in an insulator. For frequencies up to far above the optical range, the wavelengths are much larger than the distances between atoms, so we are effectively in the long wavelength limit and can assume that the electric field across a single atom is uniform.

We model the atoms as consisting of a positively charged nucleus and a negatively charged electron cloud. An applied electric field causes displacement of the electron cloud a by distance u. For small displacements, we can linearise the restoring force and assume that the restoring force is proportional to the displacement. This leads to a model of the electron cloud as a damped harmonic oscillator,

$$m\ddot{u} + m\gamma\dot{u} + m\omega_T^2 u = qE \quad , \tag{1.1}$$

where q is the charge on the electron ($= -e$), ω_T is the natural frequency, given by the force constant and mass, and γ is a damping rate. Within our classical model, it is difficult to attribute γ quantitatively to an actual physical damping mechanism – we might think of radiative

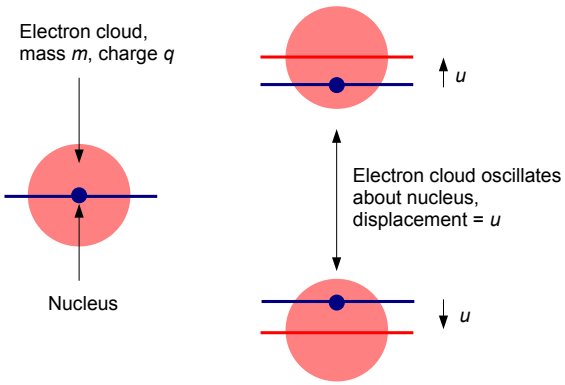


Figure 1.1: Classical picture of the electron cloud oscillating around a stationary ion

.

losses or interference with the electron clouds on neighbouring atoms – but even if we do not have a theory to calculate γ, we can include it in our model as a phenomenological parameter.

In the presence of an oscillating electric field $E(t) = E_\omega e^{-i\omega t}$, the electron cloud will oscillate with a displacement $u(t) = u_\omega e^{-i\omega t}$. The resulting dipole moment per atom at angular frequency ω is $p_\omega = qu_\omega$, which gives rise to a polarisation (=dipole moment density) $P_\omega = \epsilon_0 \chi_\omega E_\omega$, where the polarisability χ_ω is obtained from the equation of motion (1.1) as

$$\chi_\omega = \frac{N}{V} \frac{q^2}{m\epsilon_0(\omega_T^2 - \omega^2 - i\omega\gamma)} \quad , \tag{1.2}$$

where N/V is the number density of dipoles.

The relative permittivity is $\epsilon_\omega = 1 + \chi_\omega$. The typical frequency dependence of the permittivity is illustrated in Fig. 1.2.

The analogy with a damped harmonic oscillator tells us that the power absorbed by the electron cloud is determined by the imaginary part of ϵ: it is $\frac{1}{2}\omega\epsilon_0|E_\omega|^2\mathrm{Im}(\epsilon_\omega)$. This is a simple way to think about absorption lines in optical spectra and the origin of colours.

Also, note that the presence of a resonance at higher frequency makes itself felt even well below the resonance frequency. In the low frequency limit it causes an enhancement of the permittivity:

$$\epsilon(\omega \to 0) = 1 + \frac{N}{V} \frac{q^2}{m\epsilon_0\omega_T^2} \tag{1.3}$$

.

This helps us understand why different materials can have very different static (low frequency) permittivities. Moreover, a mismatch in dielectric permittivity between two media gives rise to reflection. We write the reflectivity at the interface between two media, if the permeability μ is the same in both media, as

$$r = \frac{\sqrt{\epsilon_1} - \sqrt{\epsilon_2}}{\sqrt{\epsilon_1} + \sqrt{\epsilon_2}} \quad , \tag{1.4}$$

and the power reflection coefficient as $R = |r|^2$.

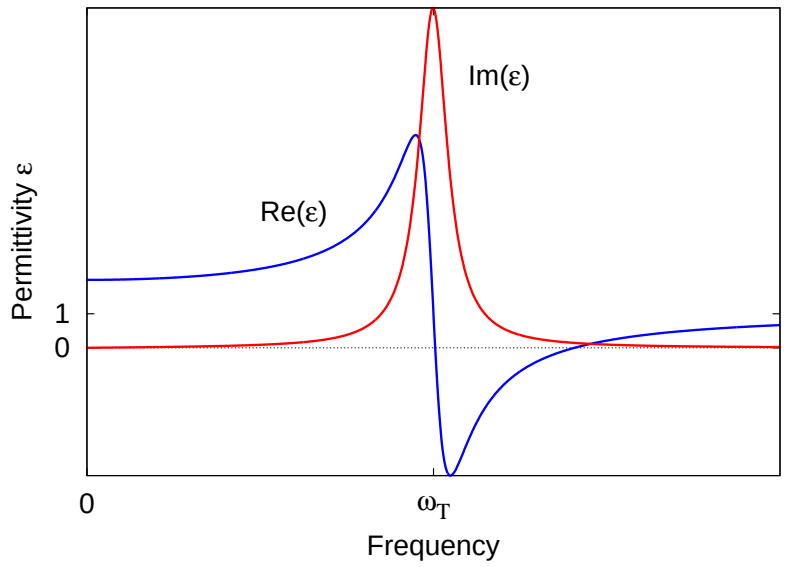


Figure 1.2: Real and imaginary part of the relative permittivity within the Lorentz (or dipole) oscillator model. Note that $\epsilon = \chi + 1$ and the polarisability χ_ω has the frequency dependence typical for a damped harmonic oscillator.

Note: we need to be careful in taking the Lorentz oscillator model at face value. Our calculation is not yet fully self-consistent, because the electric field experienced by the electron cloud on one atom is not only the applied field; it is also modified by the polarisation and the associated electric field due to other atoms in the vicinity. It can be shown, but it is outside the scope of this course, that the resulting corrections do not change the functional form of ϵ_ω but they shift the apparent resonance frequencies ω_T from those expected purely from atomic parameters such as the effective spring constant and electronic mass. In short, while the above treatment is correct for dilute gases, it needs to be modified for solids, but these modifications do not change the overall form of the results.

Of course, this classical model cannot be the whole story. For example, the equipartition theorem would tell us that each dipole oscillator should contribute $2\times\frac{1}{2}k_B$ to the heat capacity of the solid, when measurements show that the contribution due to the electrons in an insulator is vanishingly small. Also, the model gives us no handle on calculating the resonance frequencies. It works, however, as a phenomenological description of the optical response functions. Fig. 1.3 explains why: For two sharply defined energy levels E_a and E_b, time-dependent perturbation theory gives $\chi_\omega \propto (E_b - E_a - \hbar\omega)^{-1} + 2\pi i\delta(E_b - E_a - \hbar\omega)$. The imaginary part in this expression corresponds to the transition rate. This expression can also be written as $\chi_\omega \propto \frac{1}{E_b - E_a - \hbar\omega - i\hbar\gamma/2}$ with infinitesimal γ. As the energy levels broaden into bands, causing γ to become finite, this expression becomes similar to the Lorentz model close to the resonance frequency $\omega_T = (E_b - E_a)/\hbar$, if we multiply top and bottom by $\omega_T + \omega$: $\chi_\omega \propto \frac{\omega_T + \omega}{\omega_T^2 - \omega^2 - i(\omega_T + \omega)\gamma/2} \simeq \frac{2\omega_T}{\omega_T^2 - \omega^2 - i\omega\gamma}$, if we approximate ω by ω_T wherever the sum of the two frequencies occurs.

In general, atomic spectra can give rise to multiple allowed transitions at energies $\hbar\omega_{T1}$, $\hbar\omega_{T2}$, ... , $\hbar\omega_{Ti}$, etc. Usually, these occur at high frequency, at least in the optical range. The resulting frequency-dependent permittivity can be obtained by adding the responses associated with each transition: $\epsilon(\omega) = 1 + \sum \chi_i(\omega)$ (Fig. 1.4).

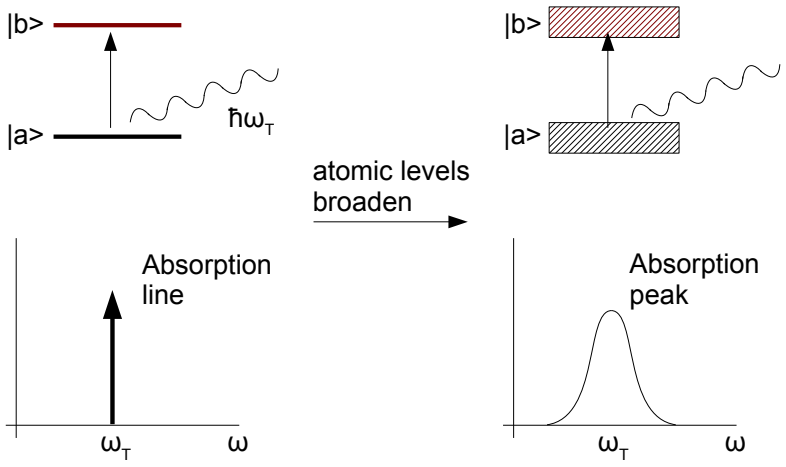


Figure 1.3: Absorption spectra produced by electronic transitions in atoms (discrete levels) and solids (levels are broadened). In solids, the sharp absorption peaks found in atomic spectra broaden out into resonances with a finite width. The resulting absorption spectra have a similar, Lorentzian form to that expected from the dipole oscillator model

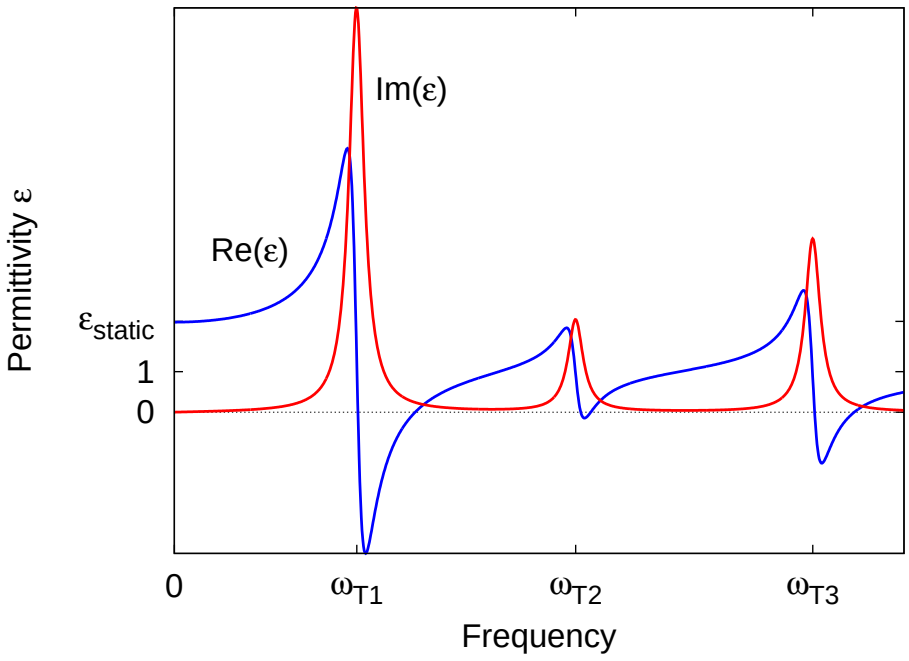


Figure 1.4: The electromagnetic response of an insulator can very generally be modelled by superimposing dipole oscillator responses with different natural frequencies, each scaled by a suitable oscillator strength. The low frequency, static permittivity then includes contributions from the low frequency tails of all the individual oscillator responses.

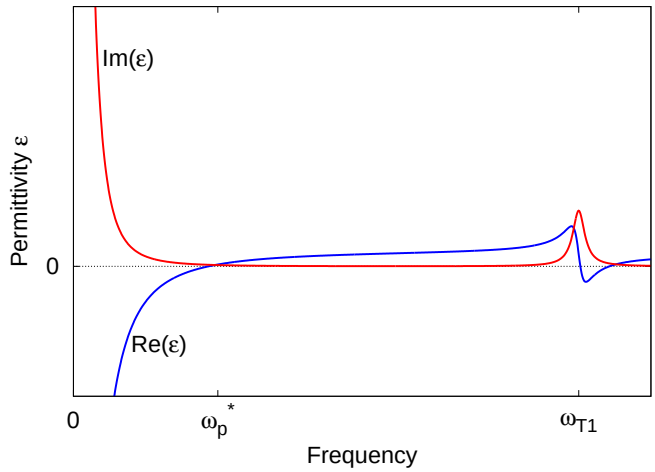


Figure 1.5: Real and imaginary part of the dielectric permittivity in the Drude model. The peak at ω_{T_1} illustrates the possibility of additional resonances due to the bound, core electrons.

1.2 Drude model

The Lorentz, or dipole oscillator model extends naturally to metals, if we imagine that some of the electrons are no longer bound to the ions. The remaining inner, or core electrons are still closely bound and will continue to contribute to the permittivity according to the Lorentz oscillator model. The outer, or conduction electrons, however, have been cut loose from the ions and are now free to roam around the entire piece of metal. This corresponds to the spring constant in their linearised force law going to zero, and hence the natural frequency ω_T vanishes. To model their contribution to ϵ_ω we can then simply use our earlier expressions for the frequency dependent permittivity, but drop ω_T: the resonance peak now occurs at zero frequency. This picture, in which some of the electrons are cut loose from the ionic cores, is called the Drude model.

1.2.1 Optical properties of metals in the Drude model

Setting ω_T in the dipole oscillator response (1.2) to zero and inserting a background permittivity ϵ_∞ to take account of the polarisability of the bound core electrons, leads to the Drude response

$$\epsilon_\omega = \epsilon_\infty - \frac{N}{V}\frac{q^2}{m\epsilon_0(\omega^2 + i\omega\gamma)} = \epsilon_\infty - \frac{\omega_p^2}{\omega^2 + i\omega\gamma} \quad , \tag{1.5}$$

where $\omega_p^2 = \frac{ne^2}{m\epsilon_0}$ (defining $n = N/V$, the number density of mobile electrons). ω_p is called the *plasma frequency*.

We can draw three immediate conclusions from the above form of the permittivitiy (Fig. 1.2):

1. $|\epsilon_\omega|$ diverges for $\omega \to 0 \implies$ metals are highly reflecting at low frequency.
2. The imaginary part of ϵ_ω peaks at $\omega = 0$, giving rise to enhanced absorption at low frequency, the 'Drude peak'.
3. ϵ_ω crosses zero and approaches unity at a high frequency $\omega_P^* \implies$ metals become transparent in the ultraviolet.

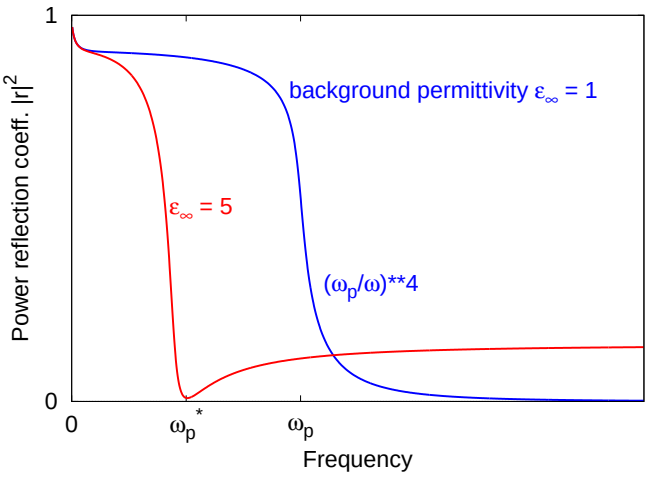


Figure 1.6: Frequency dependence of the power reflection coefficient $R = |r|^2$ in the Drude model.

Moreover, the reflection coefficient at an air/metal interface has an interesting frequency dependence. Substituting the Drude form for ϵ_ω into Eq. 1.4 allows us to analyse this in some detail (see question on problem sheet 1). We find that the power reflection coefficient R reaches 1 in the limit $\omega \to 0$, then assumes a weakly frequency dependent value less than 1 over a wide frequency range, and drops off as $R \propto \omega^{-4}$ at high frequency. A finite background polarisability (caused by the core electrons), which gives rise to $\epsilon_\infty > 1$, causes R to dip to zero at finite frequency (Fig. 1.6).

1.2.2 Plasma oscillations

These results are directly analogous to some of those seen in the Part 1B Electromagnetism course on the topic of plasmas. According to the Drude model, electrons in metals behave like a plasma, i.e. a classical charged gas moving in an oppositely charged environment. The electrons act in many ways like an ideal gas, but whereas the molecules of an ideal gas are meant to scatter off each other, we take the electrons as completely non-interacting. Our electrons do scatter off defects in the solid, however, which includes thermally excited lattice vibrations (phonons), and this gives them a mean free path ℓ and a scattering rate, which is the inverse of the relaxation time τ. We assume that a scattering event completely randomises the momenta of the electrons. As will be explained in more detail below, we can identify the damping rate γ in Eq. 1.5 with the scattering rate τ^{-1}.

One of the findings of the Part 1B Electromagnetism course was the occurrence of **plasma oscillations.** The occurrence of free oscillations in the plasma is surprising at first, because we have reduced the restoring force due to the ionic cores to zero ($\omega_T \to 0$) to obtain the Drude model. Where do these oscillations come from?

Consider probing a slab of material (the sample) by applying an oscillating field (see Fig. 1.7). The free charges brought into the vicinity of our sample to probe its properties produce a displacement, or $\mathbf{D}-$ field. Because $D_\perp$ is continuous across the interface, this can be translated to the electric field, or $\mathbf{E}-$field inside the sample, and the polarisation $\mathbf{P}$ can be determined:

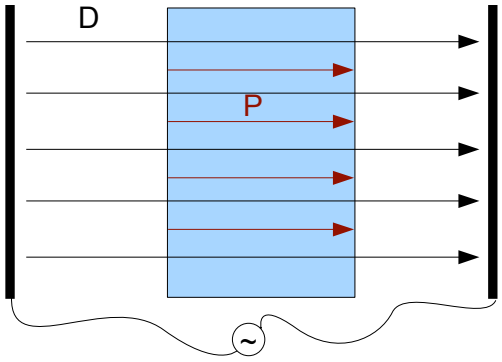


Figure 1.7: Exciting plasma oscillations by applying an oscillating electric field

$$\begin{aligned} \epsilon_0 E &= \epsilon^{-1} D = D - P \\ \implies P &= D(1 - \epsilon^{-1}) \, . \end{aligned}$$

Hence, the response to an oscillating applied D-field is given by ϵ^{-1}, which we can compute from (1.5). In metals, $\omega_p >> \gamma = \tau^{-1}$ and usually $\epsilon_\infty \simeq 1$. In this case, we can write $1/\epsilon$ as

$$\epsilon(\omega)^{-1} = \frac{\omega^2 + i\omega\gamma}{\omega^2 - \omega_p^2 + i\omega\gamma} \quad . \tag{1.6}$$

(Remember that $\omega_p^2 = \frac{ne^2}{m\epsilon_0}$, where we define $n = N/V$).

We see that for $\epsilon_\infty = 1$, the inverse permittivity $1/\epsilon_\omega$ peaks at the plasma frequency: $\epsilon(\omega_p) \to 0$. More generally, $\epsilon \to 0$ at $\omega_p^* = \omega_p/\sqrt{\epsilon_\infty}$. This implies (because $D = \epsilon_0 \epsilon_\omega E$) a finite amplitude of oscillation for E and P despite the zero forcing field D, so that at $\omega = \omega_p^*$ ('Plasma frequency'), the polarisation can oscillate without even having to apply a driving field. These are modes of *free oscillation*: solutions of the form $u_0 e^{-i\omega_p t}$ with high resonance frequency corresponding to energies in the eV range. This free resonance of the conduction electrons on top of the positively charged ionic charge background is called a *plasma oscillation*. Any polarisation in the metal causes surface charges to build up, which generate a restoring force that can drive oscillations. The entire electron gas in the metal oscillates back and forth in synchrony. The peak width of this resonance is given by τ^{-1}.

Plasma oscillations can be detected by measuring the optical absorption (which give us $\mathrm{Im}(\epsilon_\omega)$), or they can be probed by inelastic scattering of charged particles such as electrons. When high energy electrons pass through the metal, they can excite plasma oscillations and thereby lose some of their energy. By comparing the incident and final energies, we can deduce the energy of the plasma oscillations. This is called *Electron Energy Loss Spectroscopy* (EELS). As in any driven oscillator, energy is dissipated at or near the resonant frequency (with the frequency width depending on the damping of the oscillator). An EELS spectrum will therefore have a peak near the plasma frequency (Fig. 1.8).

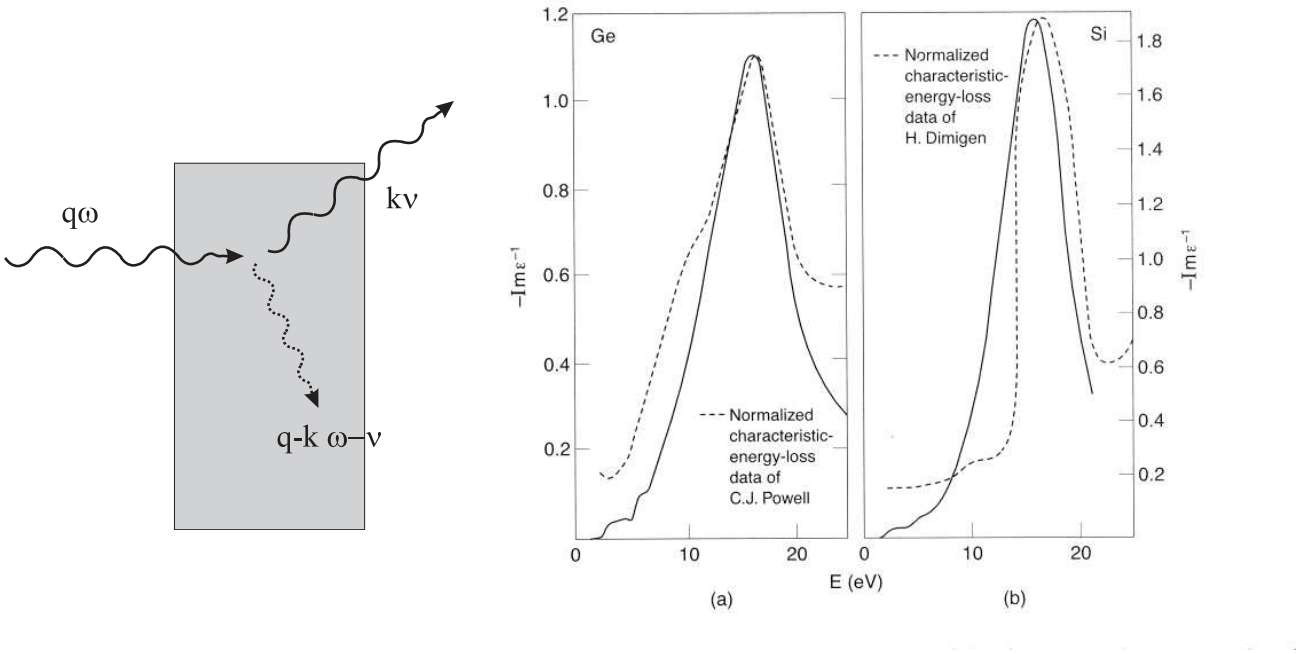


Figure 1.8: Detection of plasma oscillations by electron energy loss spectroscopy. Left panel: Generic diagram of an inelastic scattering experiment. The incident particle – in this case an electron – is scattered to a final state of different energy and momentum. Electrons enter the sample at wavevector $\mathbf{q}$ and energy $\hbar\omega$. The outgoing electrons have a different wavevector $\mathbf{k}$ and energy $\hbar\nu$. By comparing the incident and scattered energies, one deduces the *energy loss spectrum* of the internal collective excitations in the medium. For high energy electrons - typically used in an EELS experiment - the momentum loss $(\mathbf{q} - \mathbf{k})$ is small. Centre and right panels: Electron energy loss spectrum for Ge and Si (dashed lines) compared to values of $Im(1/\epsilon)$ extracted directly from measurements of the optical absorption. The energy difference between incoming and outgoing electrons is attributed to the excitation of plasma oscillations (or 'plasmons') with energy $\hbar(\omega - \nu)$. [From H.R. Philipp and H. Ehrenreich, Physical Review **129**, 1550 (1963)

1.2.3 Frequency dependent conductivity in the Drude model

So far we have worked out the frequency dependent permittivity in metals using the Lorentz oscillator model and simply setting the dipole resonance frequency ω_T to zero. This approach is conceptually somewhat unsatisfactory, primarily for two reasons:

1. If the electrons are cut loose from the ionic cores, their average position, around which they are meant to oscillate, is no longer defined. The displacement $\mathbf{u}$ for an individual electron acquires an arbitrary offset.

2. Our picture of scattering processes (electrons occasionally hit an obstacle and thereby randomise their momentum) prevents us from ascribing the same velocity to all the electrons – we need a statistical description.

These issues can be addressed by two modifications to our approach, which do not change any of our results so far, but help us interpret these results and make further progress:

1. Rather than starting with an equation of motion for the electronic displacement, we consider the rate of change of the velocity $\mathbf{v} = \dot{\mathbf{u}}$.

2. Instead of considering the velocity of an individual electron, we average over a large number electrons, and look for a differential equation for the *average* velocity of the electrons. Even if the individual electrons have wildly different velocities, the average, or *drift* velocity will be seen to follow a simple equation of motion, and it is the drift velocity that determines the optical and transport properties of the metal.

Writing $\rho = qN/V$ for the charge density, we can express the current density $\mathbf{j}$ in terms of the average velocity $\mathbf{v}$ of the mobile, or conduction electrons, and we can link $\mathbf{j}$ to $\mathbf{E}$ via the conductivity σ and Ohm's law:

$$\mathbf{j} = \rho\mathbf{v} = \sigma\mathbf{E} \tag{1.7}$$

Because the polarisation $P = (N/V)uq$, we can link the current density to the time-derivative of the polarisation due to the conduction electrons, $\mathbf{P}_c$:

$$\mathbf{j} = \dot{\mathbf{P}}_c \tag{1.8}$$

. Note that while there can be some ambiguity about P due to the ambiguity about the displacement u for travelling electrons, this drops out of the time derivative, so the expression for j is correct. [1]

Adding in the polarisation of the core electrons, which is given by the background polarisability χ_∞, we obtain

$$\dot{\mathbf{P}} = \mathbf{j} + \epsilon_0\chi_\infty\dot{\mathbf{E}} \tag{1.9}$$

. Considering, as before, the oscillatory response $\mathbf{v}_\omega e^{-i\omega t}$, $\mathbf{j}_\omega e^{-i\omega t}$, $\mathbf{P}_\omega e^{-i\omega t}$ to an oscillating electric field $\mathbf{E}_\omega e^{-i\omega t}$, we find for the Fourier components

$$\mathbf{j}_\omega = \sigma_\omega\mathbf{E}_\omega = -i\omega\epsilon_0(\chi_\omega - \chi_\infty)\mathbf{E}_\omega \quad , \tag{1.10}$$

from which we obtain the key expressions

$$\begin{aligned} \sigma_\omega &= -i\omega\epsilon_0(\epsilon_\omega - \epsilon_\infty) \\ \epsilon_\omega &= i\frac{\sigma_\omega}{\epsilon_0\omega} + \epsilon_\infty. \end{aligned} \tag{1.11}$$

These expressions relate the imaginary part of the permittivity (which determines optical absorption) to the real part of the frequency dependent conductivity. This means that we can indirectly determine the conductivity of a metal at high frequencies by optical measurements.

If the atomic (or background) polarisability is zero, then $\epsilon_\infty = 1$ and $\sigma_\omega = -i\omega\epsilon_0(\epsilon_\omega - 1)$.

Differential equation for the drift velocity

Rather than simply inserting our earlier expression (1.5) into (1.11), which would indeed give us a correct expression for the frequency dependent conductivity σ_ω within the Drude model, we

[1] An alternative approach, which avoids introducing the field $\mathbf{u}(\mathbf{r}, t)$ altogether, could consider $\nabla\mathbf{P}_c = -\rho_c$ and $\dot{\rho}_c = -\nabla\mathbf{j}$ to find $\mathbf{j} = \dot{\mathbf{P}}_c + \mathbf{j}_0$, where $\nabla\mathbf{j}_0 = 0$ so that the boundary conditions on the sample fix $\mathbf{j}_0 = 0$.

can also derive an expression for σ_ω by considering an equation of motion for the drift velocity of the electrons, or for the resulting current density. This answers the conceptual issues raised earlier and gives a more precise meaning to the relaxation time τ. The drift velocity relates to the total momentum of the electron system via $\mathbf{v} = \frac{\mathbf{p}}{Nm}$, where N is the number of electrons in the system and m is the electronic mass. The momentum, in turn, changes in the presence of an applied electric or magnetic field. If there were no collisions, which can remove momentum from the electron system, we would have $\dot{\mathbf{p}} = N\mathbf{f}(t) = N\frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B})$.

Collisions, or scattering introduce a further term that represents the decay of the electron momentum in the absence of an external force. Note that electron-electron collisions do not give rise to a decay of momentum in any obvious way: they would appear to conserve the momentum of the electron system. It turns out that at a more advanced level of analysis, they do contribute to the relaxation of momentum, but let us for the moment neglect this contribution. The electron momentum decays due to collisions of the electrons with lattice imperfections such as impurities, dislocations etc., and – in the wider sense – lattice distortions caused by lattice vibrations.

We could model the influence of electron scattering events by making two simplifying assumptions:

- Electron collisions randomise the electron momenta, so that – on average – the contribution of an electron to the total momentum is zero after a collision.
- The probability for a collision to occur, P, is characterised by a single relaxation time τ: $P(\text{collision in } [t, t+dt]) = dt/\tau$.

From these assumptions, we find that the probability that a particular electron has not scattered in the time interval $[t, t+dt]$ is $1 - dt/\tau$. As only the electrons which have not scattered contribute to the total momentum (the momentum of the others randomises to zero on average), and these electrons continue to be accelerated by the applied force, we obtain a total momentum after time $t + dt$ of:

$$\mathbf{p}(t+dt) = (1 - dt/\tau)(\mathbf{p}(t) + N\mathbf{f}(t)dt) + ... \tag{1.12}$$

.

This gives rise to a differential equation for the momentum:

$$\left(\frac{d}{dt} + \frac{1}{\tau}\right)\mathbf{p} = N\mathbf{f}(t) \tag{1.13}$$

and, substituting $\mathbf{f} = q\mathbf{E}$, we obtain for the drift velocity and for the current density:

$$\begin{aligned} \left(\frac{d}{dt} + \frac{1}{\tau}\right)\mathbf{v} &= \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \\ \left(\frac{d}{dt} + \frac{1}{\tau}\right)\mathbf{j} &= \frac{N}{V}\frac{q^2}{m}\mathbf{E}(t) + \frac{q}{m}\mathbf{j} \times \mathbf{B}. \end{aligned} \tag{1.14}$$

We neglect $\mathbf{B}$ for the time being, as the effects of the magnetic field accompanying an electromagnetic wave on the electron system are much weaker than those of the electric field.

However, **B** will be included in transport phenomena at low frequencies, below (Hall effect). Then, we obtain the simplified equations

$$\begin{aligned} \left(\frac{d}{dt}+\frac{1}{\tau}\right)\mathbf{v} &= q\mathbf{E}(t)/m \\ \left(\frac{d}{dt}+\frac{1}{\tau}\right)\mathbf{j} &= \frac{N}{V}\frac{q^2}{m}\mathbf{E}(t) \end{aligned} \tag{1.15}$$

To extract the frequency dependent conductivity from (1.15), we again insert a trial solution of the form $\mathbf{j} = \mathbf{j}_\omega e^{-i\omega t}$ and a time-varying field of the form $\mathbf{E} = \mathbf{E}_\omega e^{-i\omega t}$. This transforms the above differential equation into an algebraic equation:

$$\mathbf{j}_\omega = \frac{ne^2}{m}\frac{1}{1/\tau - i\omega}\mathbf{E}_\omega \tag{1.16}$$

(Remember that $q = -e$ is the electronic charge, and we write the number density N/V as n). Using the definition of the conductivity $\mathbf{j} = \sigma\mathbf{E}$, we obtain:

$$\sigma_\omega = \frac{ne^2}{m}\frac{1}{1/\tau - i\omega} \quad , \tag{1.17}$$

In the low frequency limit, this expression for the conductivity tends to the DC conductivity

$$\sigma_0 = \frac{ne^2\tau}{m} \tag{1.18}$$

At frequencies larger than $1/\tau$ the conductivity falls off rapidly:

$$\mathrm{Re}(\sigma_\omega) = \frac{\sigma(0)}{1+\omega^2\tau^2} \quad . \tag{1.19}$$

The DC ($\omega = 0$) conductivity can also be written in terms of the *mobility* $\mu = e\tau/m$

$$\sigma = ne\mu = \frac{ne^2\tau}{m} \tag{1.20}$$

Inserting (1.17) into Eqn. 1.11 gives

$$\epsilon(\omega) = \epsilon_\infty - \frac{\omega_p^2}{\omega^2 + i\omega/\tau} \quad , \tag{1.21}$$

which is exactly the expression we obtained right at the start (Eq. 1.5) by setting $\omega_T \to 0$, if we identify $1/\tau$ with γ. Here, as before, ω_p is the *Plasma frequency*:

$$\omega_p^2 = \frac{ne^2}{\epsilon_0 m} \tag{1.22}$$

.

1.2.4 Low frequency (DC) transport properties in the Drude model

We now use the differential equation we obtained earlier for the current density in applied electric and magnetic fields, (1.14), in order to study the electrical transport in a transverse magnetic field. We consider an arrangement, in which a static magnetic field $\mathbf{B}$ is applied along the $\hat{z}$ direction, and static currents and electrical fields are constrained to the $x-y$ plane.

The equations of motion for charge carriers with charge q are now [2].

$$\begin{aligned} \left(\partial_t + \tau^{-1}\right)\mathbf{j}_x &= \frac{q^2 n}{m}\left(E_x + v_y B\right) \\ \left(\partial_t + \tau^{-1}\right)\mathbf{j}_y &= \frac{q^2 n}{m}\left(E_y - B v_x\right) \\ \left(\partial_t + \tau^{-1}\right)\mathbf{j}_z &= \frac{q^2 n}{m} E_z \end{aligned} \tag{1.23}$$

In steady state, we set the time derivatives $\partial_t = d/dt = 0$, and get the three components of the current density

$$\begin{aligned} j_x &= qn\left(\frac{q\tau}{m}E_x + \beta v_y\right) \\ j_y &= qn\left(\frac{q\tau}{m}E_y - \beta v_x\right) \\ j_z &= qn\frac{q\tau}{m}E_z \end{aligned} \tag{1.24}$$

where the dimensionless parameter $\beta = \frac{qB}{m}\tau = \omega_c\tau = \mu B$ is the product of the cyclotron frequency ($\omega_c = qB/m$) and the relaxation time, or of the mobility and the applied field.

Hall effect

Consider now the rod-shaped geometry of Fig. 1.9. The current is forced by geometry to flow only in the x-direction, so that $\mathbf{j}_y = 0$, $\mathbf{v}_y = 0$. Since there is no flow in the normal direction, there must be an electric field $\mathbf{E} = -\mathbf{v} \times \mathbf{B}$, which exactly counterbalances the Lorentz force on the carriers. This is the *Hall effect*. We find

$$\mathbf{v}_x = \frac{q\tau}{m}\mathbf{E}_x \ , \tag{1.25}$$

and

$$\mathbf{E}_y = \beta \mathbf{E}_x \tag{1.26}$$

(Remember $\beta = \frac{qB}{m}\tau$). It turns out that for high mobility materials and large magnetic fields, it is not hard to reach large values of $|\beta| \gg 1$, so that the electric fields are largely normal to the electrical currents.

The *Hall coefficient* is defined by

$$R_H = \frac{\mathbf{E}_y}{\mathbf{j}_x \mathbf{B}} = \frac{1}{nq} \tag{1.27}$$

[2] Again, we define e to be the *magnitude* of the charge of an electron. Note, also, that the particle mass m, may in general differ from the mass of an electron in vacuum, m_e. We will see later that in solids the effective charge carrier mass depends on details of the electronic structure.

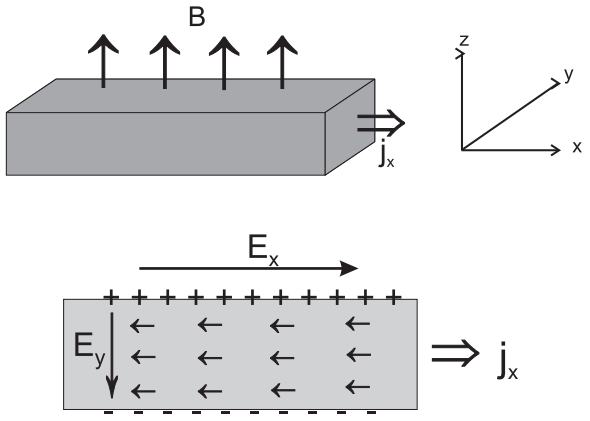


Figure 1.9: The upper figure shows the geometry of a *Hall bar*, with the current flowing uniformly in the x-direction, and the magnetic field in z. The lower figure shows the *steady state* electron flow (arrows) in a section normal to $\hat{z}$. When a voltage E_x is first applied, and E_y is not yet established, the electrons will deflect and move in the (downward) y-direction. The y-surfaces of the crystal then become charged, producing the field E_y which exactly cancels the Lorentz force $-ev_xB$.

Notice that it is *negative* for electrons, but importantly it is independent of the effective mass, and increases with decreasing carrier density. For holes of charge $+e$ the sign is positive.

The Hall effect is an important diagnostic for the density and type of carriers transporting the electrical current in a semiconductor. The simple picture presented here works quite well for alkali metals, where the predicted Hall coefficient is within a few percent of the expected value for a parabolic free electron band. But Be, Al, and In all have *positive* Hall coefficients - accounted for by a band-structure with hole pockets that dominates the Hall effect. In still more complicated cases, contributions from both positive and negative carriers, attributed to different electronic bands, combine in a non-trivial way to determine the Hall effect.

Thermal conductivity of metals

Particles with velocity v, mean free path ℓ and specific heat C are expected to yield a thermal conductivity $K = Cv\ell/3$. For a free Fermi gas, we get the correct answer from this formula by using the electronic specific heat, the characteristic carrier velocity v_F, and the mean free path for carriers on the Fermi surface $\ell = v_F\tau$. Hence, using the relationship between the Fermi velocity and the Fermi energy $E_F = mv_F^2/2$, we obtain

$$K_{el} = \frac{\pi^2}{2}\frac{nk_B^2T}{E_F} \cdot v_F \cdot v_F\tau = \frac{\pi^2 nk_B^2T\tau}{3m} \tag{1.28}$$

Almost invariably, the electronic thermal conductivity is larger than that due to the lattice. K and σ are of course closely related, being both proportional to the scattering time and the density, as is natural. The ratio

$$\frac{K}{\sigma T} = \frac{\pi^2}{3}\left(\frac{k_B}{e}\right)^2 \tag{1.29}$$

is expected to be constant, independent of material parameters. This proportionality is the *Wiedemann-Franz* law, which works strikingly well for simple metals.

1.2.5 Summary of key results in Drude theory

Plasma frequency

$$\omega_p^2 = \frac{ne^2}{\epsilon_0 m}$$

Frequency dependence of the relative permittivity and of the electrical conductivity (taking into account the polarisability of the atomic cores, $\chi_\infty = \epsilon_\infty - 1$ causes slight modifications):

$$\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma} \xrightarrow{\epsilon_\infty \neq 1} \epsilon_\infty - \frac{\omega_p^2}{\omega^2 + i\omega\gamma}$$

$$\sigma(\omega) = -i\omega\epsilon_0(\epsilon(\omega) - 1) \xrightarrow{\epsilon_\infty \neq 1} -i\omega\epsilon_0(\epsilon(\omega) - \epsilon_\infty)$$

Useful expressions for electrical conductivity and Hall coefficient:

$$\sigma(\omega) = \frac{ne^2\tau}{m(1 - i\omega\tau)} \xrightarrow{\omega \to 0} \frac{ne^2\tau}{m} = ne\mu$$

$$R_H = \frac{1}{nq}$$

(q is carrier charge, n is carrier density, τ is relaxation time $= 1/\gamma$)

Chapter 2

Sommerfeld theory – electrons as a degenerate quantum gas

2.1 The problems with Drude theory

Just as the Lorentz oscillator model is successful at describing the optical response of insulators, Drude theory works surprisingly well in modelling the optical and transport properties of metals. Both theories fail dramatically, however, when the thermodynamic properties are considered. Applying the equipartition theorem to the dipole oscillator model, we would expect a contribution of k_B to the heat capacity of each oscillator. Similarly, within the Drude model, which treats the conduction electrons like a classical ideal gas, we would expect a contribution to the heat capacity of $\frac{3}{2}k_B$ per conduction electron. In reality, the measured heat capacities for both insulators and metals are far beyond those values (Fig. 2.1).

The reason for this is the same as the reason why the heat capacity due to lattice vibrations falls below the Dulong-Petit limit at low temperature: electronic motion is largely frozen out, because in a quantum mechanical model for the electrons, the energy required to excite them exceeds the thermal energy available.

More specifically, in insulators, the atomic energy levels are separated by large energy gaps of the order of electron Volts ($\simeq k_B \times 11,000$ K). Therefore, the specific heat contribution due to electronic excitations in insulators will only become noticeable at temperatures of thousands of Kelvin.

In metals, on the other hand, low energy excitations are always possible, but only for a small fraction of the electrons. As we will see (and has been shown in the *Thermal and Statistical Physics* course), the conduction electrons form a degenerate Fermi gas, in which only the fraction $\sim k_B T/E_F$ (where the Fermi energy $E_F \sim$ eV) are close enough to the chemical potential so that they can contribute to the heat capacity, which is therefore proportional to temperature. Within the Fermi gas picture, most of the electrons travel at very high speeds, dictated by the wavevector of the quantum state they occupy. These speeds can reach 10^6 m/s, in contrast to the far lower velocities that appear in the Drude model. This contrast emphasises, again, the need to interpret the velocity $v = \dot{u}$ in the Drude model as a *drift velocity*, averaged over many particles.

By using quantum statistics, as introduced in the *Thermal and Statistical Physics* course, we

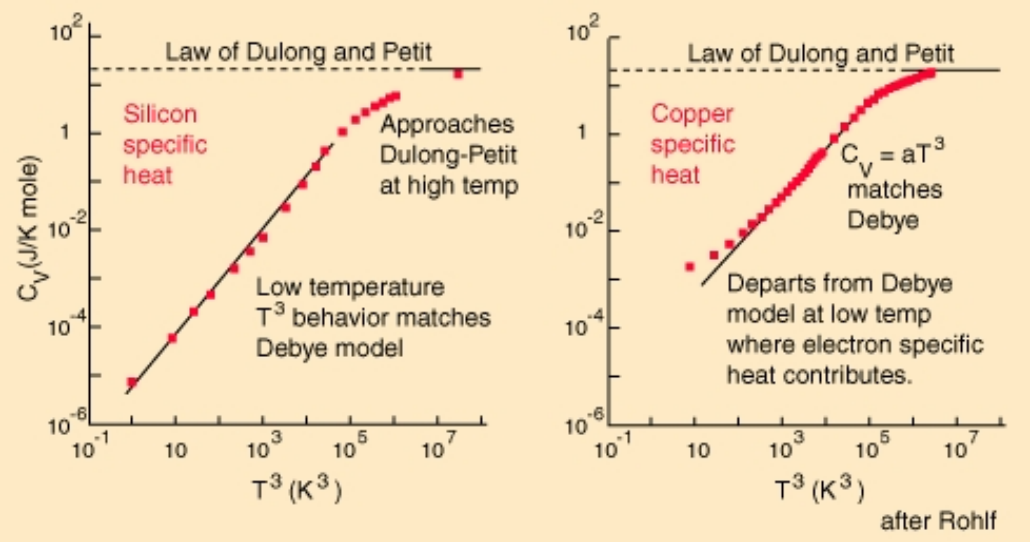


Figure 2.1: Temperature dependence of the molar heat capacity of silicon (left) and copper (right). At high temperatures, the heat capacity tends towards the Dulong-Petit limit $C = 3R$, which attributes the heat capacity entirely to atomic vibrations, using the equipartition theorem (3 degrees of freedom, potential and kinetic energy $\rightarrow 6 \times \frac{1}{2}k_B$. At low temperatures, the heat capacity due to lattice vibrations is frozen out, as predicted by the Debye theory of solids ($C \propto T^3$). The electronic contribution is clearly far smaller than would be expected, if the electrons behaved like a classical ideal gas ($C_e = \frac{3}{2}k_B$), and is only visible at very low temperatures.

can resolve the difficulties of the Drude model and achieve an understanding of thermodynamic properties of solids.

2.2 Free electron gas in three-dimensions

Consider a free electron gas, confined to a three-dimensional box of side L. The free particle Schrödinger equation is

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial y^2}\right)\psi(\mathbf{r}) = \epsilon\psi(\mathbf{r}) \tag{2.1}$$

which has the following eigenstates:

$$\psi_{\mathbf{k}}(\mathbf{r}) = \mathcal{N}\sin(k_x x)\sin(k_y y)\sin(k_z z) \tag{2.2}$$

with energy

$$\epsilon_{\mathbf{k}} = \frac{\hbar^2|k|^2}{2m} \tag{2.3}$$

Owing to the restriction to the box $(0 < x < L, 0 < y < L, 0 < z < L))$ the allowed values of k are discrete.

$$\mathbf{k} = \frac{\pi}{L}(n_x, n_y, n_z) \tag{2.4}$$

where n_x, n_y, n_z are **positive** integers.

It is more convenient to introduce wavefunctions that satisfy periodic boundary conditions, namely

$$\psi(x+L,y,z)=\psi(x,y,z) \tag{2.5}$$

and similarly for the y and z directions. These are of the form of a plane wave

$$\psi_k(\mathbf{r})=\exp(i\mathbf{k}\cdot\mathbf{r}) \tag{2.6}$$

where the eigen-energies are identical to (2.4) but the restriction on momentum being

$$\mathbf{k}=\frac{2\pi}{L}(n_x,n_y,n_z) \tag{2.7}$$

with n_x, n_y, n_z **positive or negative** integers.

Together with the spin quantum number m, the components of $\mathbf{k}$ are the good quantum numbers of the problem.

2.3 Fermi surface and density of states

In the ground state at zero temperature, the Fermi gas can be represented by filling up all the low energy states up to a maximum energy ϵ_F (the Fermi energy) corresponding to a sphere of radius the Fermi momentum k_F in k-space.

Each triplet of quantum numbers k_x, k_y, k_z accounts for two states (spin degeneracy) and occupies a volume $(2\pi/L)^3$.

The total number of occupied states inside the Fermi sphere is

$$N=2\cdot\frac{4/3\pi k_F^3}{(2\pi/L)^3} \tag{2.8}$$

so that the Fermi wave-vector is written in term of the electron density $n=N/V$ as

$$k_F=(3\pi^2 n)^{1/3}\ . \tag{2.9}$$

We are often interested in the *density of states*, $g(E)$, which is the number of states per unit energy range. This is calculated by determining how many states are enclosed by a thin shell of energy width dE, viz.

$$g(E)dE=2\cdot\frac{\text{Volume of shell in k}-\text{space}}{\text{Volume of k}-\text{space per state}}=2\cdot\frac{4\pi k^2dk}{(2\pi)^3/V}\quad , \tag{2.10}$$

hence

$$g(E)=2\frac{V}{(2\pi)^3}4\pi k^2\frac{dk}{dE}=\frac{V}{\pi^2}\frac{m}{\hbar^2}\left(\frac{2mE}{\hbar^2}\right)^{\frac{1}{2}}\quad . \tag{2.11}$$

The factor of 2 is for spin degeneracy. Often, the density of states is given per unit volume, so the factor of V disappears.

2.4 Thermal properties of the electron gas

The occupancy of states in thermal equilibrium in a Fermi system is governed by the Fermi distribution

$$f(E) = \frac{1}{e^{(E-\mu)/k_B T} + 1} \tag{2.12}$$

where the chemical potential μ can be identified (at zero temperature) with the Fermi energy E_F of the previous section.

The number density of particles is

$$\begin{aligned} n = N/V &= \frac{1}{V}\sum_i f(E_i) = \frac{2}{V}\sum_{\mathbf{k}} f(\epsilon_{\mathbf{k}}) \\ &= \frac{1}{4\pi^3}\int d\mathbf{k} f(\epsilon_{\mathbf{k}}) \\ &= \int dE\ g(E) f(E) \end{aligned} \tag{2.13}$$

The internal energy density $u = U/V$ can then be written in the same fashion:

$$u = \int dE\ E g(E) f(E) \tag{2.14}$$

Eq. (2.14) will be used to derive the electronic specific heat $c_v = \partial u/\partial T|_v$ at constant volume. The estimation is made much simpler by realising that in almost all cases of interest, the energy scale set by temperature $k_B T$ (≈ 0.025 eV at room temperature) is much less than the Fermi energy E_F (a few eV in most metals).

From (2.14)

$$c_v = \int dE\ E g(E) \frac{\partial f(E)}{\partial T} \tag{2.15}$$

Notice that the Fermi function is very nearly a step-function, so that the temperature-derivative is a function that is sharply-peaked for energies near the chemical potential. The contribution to the specific heat then comes only from states within $k_B T$ of the chemical potential and is much less than the $3/2 k_B$ per particle from classical distinguishable particles. From such an argument, one guesses that the specific heat per unit volume is of order

$$c_v \approx \frac{N}{V}\frac{k_B T}{E_F} k_B \tag{2.16}$$

Doing the algebra is a little tricky, because it is important to keep the number density fixed ((2.13)) — which requires the chemical potential to shift (a little) with temperature since the density of states is not constant. A careful calculation is given by Ashcroft and Mermin.

But to the extent that we can take the density of states to be a constant, we can remove the factors $g(E)$ from inside the integrals. Notice that with the change of variable $x = (E-\mu)/k_B T$,

$$\frac{df}{dT} = \frac{e^x}{(e^x+1)^2} \times \left[\frac{x}{T} + \frac{1}{k_B T}\frac{d\mu}{dT}\right] \tag{2.17}$$

The number of particles is conserved, so we can write

$$\frac{dn}{dT} = 0 = g(E_F) \int dE \; \frac{\partial f(E)}{\partial T} \tag{2.18}$$

which on using (2.17) becomes

$$0 = g(E_F) k_B T \int_{-\infty}^{\infty} dx \frac{e^x}{(e^x+1)^2} \times \left[\frac{x}{T} + \frac{1}{k_B T} \frac{d\mu}{dT} \right] \quad . \tag{2.19}$$

The limits can be safely extended to infinity: the factor $\frac{e^x}{(e^x+1)^2}$ is even, and hence at this level of approximation $d\mu/dT = 0$.

To the same level of accuracy, we have

$$\begin{aligned} c_v &= g(E_F) \int dE \; E \frac{\partial f(E)}{\partial T} \\ &= g(E_F) k_B T \int_{-\infty}^{\infty} dx \; (\mu + k_B T \, x) \frac{e^x}{(e^x+1)^2} \frac{x}{T} \\ &= g(E_F) k_B^2 T \int_{-\infty}^{\infty} dx \; \frac{x^2 e^x}{(e^x+1)^2} \\ &= \frac{\pi^2}{3} k_B^2 T g(E_F) \end{aligned} \tag{2.20}$$

The last result is best understood when rewritten as

$$c_v = \frac{\pi^2}{2} \frac{k_B T}{E_F} n k_B \tag{2.21}$$

confirming the simple argument given earlier and providing a numerical prefactor.

The calculation given here is just the leading order term in an expansion in powers of $(k_B T/E_F)^2$. To next order, one finds that the chemical potential is indeed temperature-dependent:

$$\mu = E_F \left[1 - \frac{1}{3} (\frac{\pi k_B T}{2E_F})^2 + O(k_B T/E_F)^4 \right] \tag{2.22}$$

but this shift is small in metals at room temperature, and may usually be neglected.

2.5 Screening and Thomas-Fermi theory

One of the most important characteristics of the metallic state is the phenomenon of screening. If we insert a positive test charge into a metal, it attracts a cloud of electrons around it, so that at large distances away from the test charge the potential is perfectly screened - there is *zero* electric field inside the metal. Notice that this is quite different from a dielectric, in which the form of the electrostatic potential is unchanged but the magnitude is reduced by the dielectric constant ϵ.

Screening involves a length-scale: a perturbing potential is not screened perfectly at very short distances. Why not? In a classical picture, one might imagine that the conduction electrons simply redistribute in such a way as to cancel any perturbing potential perfectly. This would require precise localisation of the electrons, however, which in quantum mechanics would incur too high a penalty in kinetic energy. Just as in the hydrogen atom, the electron cannot sit right on top of the proton, a balance is reached in metals between minimising potential and kinetic energy. This leads to charge density building up in the vicinity of a perturbing potential, which will screen the potential over a short but finite distance.

Response to an external potential

The aim of this calculation is to estimate the response of a free electron gas to a perturbing potential. The perturbing potential could be caused by charges outside the metal, but it could also be due to extra charges placed inside the metal.

We begin with the free electron gas in a metal, without an externally applied perturbing potential. The electrostatic potential in the metal, $V_0(\mathbf{r})$ is connected to the charge distribution $\rho_0(\mathbf{r})$ via

$$\nabla^2 V_0(\mathbf{r}) = -\frac{\rho_0(\mathbf{r})}{\epsilon_0} \tag{2.23}$$

In the simplest model of a metal, we consider the positive background charge to be smeared out homogeneously throughout the metal. The electron gas moves on top of this positive background. This is the plasma or 'Jellium' model for a metal. $\rho_0 = 0$ everywhere in this case.[1]

In the presence of a perturbing potential $V_{ext}(\mathbf{r})$, the electron charge density $\rho(\mathbf{r})$ will redistribute, $\rho(\mathbf{r}) = \rho_0(\mathbf{r}) + \delta\rho(\mathbf{r})$, causing a correction to the potential $V(\mathbf{r}) = V_0(\mathbf{r}) + \delta V(\mathbf{r})$:

$$\nabla^2 \delta V(\mathbf{r}) = -\frac{\delta\rho(\mathbf{r})}{\epsilon_0} \tag{2.24}$$

In order to make progress, we need to link the charge density redistribution $\delta\rho$ to the applied potential V_{ext}. For long-wavelength perturbations, it is plausible that in a region surrounding the position $\mathbf{r}$ the perturbing potential effectively just shifts the free electron energy levels, which is equivalent to assuming a spatially varying Fermi energy. This is the essence of the *Thomas Fermi approximation.*

Thomas-Fermi approximation

We shall treat the case of "jellium", where the ionic potential is spread out uniformly to neutralise the electron liquid. Note: the average charge density is therefore always zero! The metal is neutral. An external potential will, however, cause a redistribution of charge, leading to local accumulation of positive or negative charge, which will tend to screen the external potential. The net effect will be that the total potential seen by an individual electron in the Schrödinger equation is less than the external potential. We wish to calculate the charge density induced by such an external potential $\rho_{ind}([V_{ext}(\mathbf{r})])$.

Jellium. The potential in the problem is the *total* potential (external plus induced, $V_{tot} = V_{ext} + \delta V$) produced by the added charge and by the non-uniform screening cloud (see Fig. 2.2)

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + (-e)(\delta V(\mathbf{r}) + V_{ext}(\mathbf{r}))\psi(\mathbf{r}) = E\psi(\mathbf{r}) \quad . \tag{2.25}$$

Slowly varying potential. Assume that the induced potential is slowly varying enough that the energy eigenvalues of (2.25) are still indexed by momentum, but just shifted by the potential as a function of position:

$$E(\mathbf{k}, \mathbf{r}) = E_0(\mathbf{k}) - eV_{tot}(\mathbf{r}) \quad , \tag{2.26}$$

[1] The correction to the charge density, $\delta\rho$, does not include those charges (ρ_{ext}) which may have been placed inside the metal to set up the perturbing potential. They obey $\nabla^2 V_{ext} = -\rho_{ext}/\epsilon_0$.

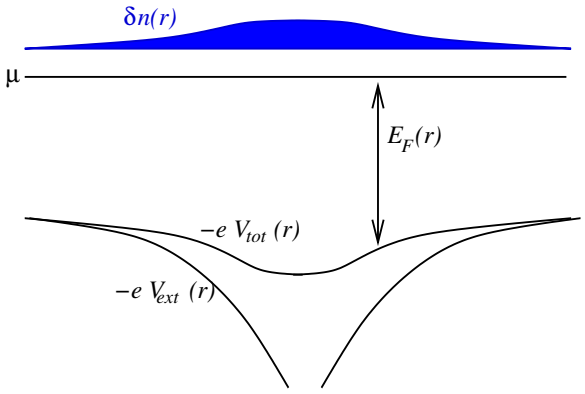


Figure 2.2: Thomas-Fermi approximation

where $E_0(\mathbf{k})$ follows the free electron, parabolic dispersion $\frac{\hbar^2 k^2}{2m}$. This only makes sense in terms of wavepackets, but provided the potential varies slowly enough on the scale of the Fermi wavelength $2\pi/k_F$, this approximation is reasonable.

Constant chemical potential. Keeping the electron states filled up to a constant energy μ requires that we adjust the local Fermi energy $E_F(\mathbf{r})$ (as measured from the bottom of the band) such that[2]

$$\mu = E_F(\mathbf{r}) - eV_{tot}(\mathbf{r}) \quad , \tag{2.27}$$

Local density approximation. We assume that E_F depends only on the local electron number density n via the density of states per unit volume $g_V(E)$:

$$\int^{E_F} g_V(E)dE = n \quad . \tag{2.28}$$

This means that a small shift in the Fermi energy, δE_F gives rise to a change in the number density $\delta n = g_V(E_F)\delta E_F$. The Fermi energy shift, in turn, is linked (via Eqn. 2.27) to V_{tot} as $\delta E_F = e(\delta V + V_{ext})$, from which we obtain

$$\delta n = e g_V(E_F)(\delta V + V_{ext}) \quad . \tag{2.29}$$

Linearised Thomas-Fermi. When the added potential V_{ext} is small, the induced number density δn is small, and therefore the number density n cannot differ very much from the density n_0 of the system without the potential ($n = n_0 + \delta n$). We may then express Eqn.2.24 in a linearised form with respect to the perturbing potential:

$$\nabla^2 \delta V(\mathbf{r}) = \frac{e^2 g_V(E_F)}{\epsilon_0}(\delta V(\mathbf{r}) + V_{ext}(\mathbf{r})) \tag{2.30}$$

Density response. This is solved by Fourier transformation, for instance by assuming an oscillatory perturbing potential $V_{ext} = V_{ext}(\mathbf{q})e^{i\mathbf{q}.\mathbf{r}}$ and a resulting oscillatory induced potential

[2]One is often sloppy about using E_F and μ interchangeably; here is a place to take care.

$\delta V = \delta V(\mathbf{q})e^{i\mathbf{q}.\mathbf{r}}$:

$$\delta V(\mathbf{q}) = -\frac{e^2 g_V(E_F)/\epsilon_0}{q^2 + e^2 g_V(E_F)/\epsilon_0} V_{ext}(\mathbf{q}) = -\frac{q_{TF}^2}{q^2 + q_{TF}^2} V_{ext}(\mathbf{q}) \quad , \tag{2.31}$$

where we have collected $e^2 g_V(E_F)/\epsilon_0$ into the *Thomas Fermi wave vector* $q_{TF} = (e^2 g_V(E_F)/\epsilon_0)^{\frac{1}{2}}$, which for the free electron gas is

$$q_{TF}^2 = \frac{1}{\pi^2}\frac{me^2}{\epsilon_0\hbar^2}k_F = \frac{4}{\pi}\frac{k_F}{a_B} = (\frac{2.95}{\sqrt{r_s}}\text{Å}^{-1})^2 \ . \tag{2.32}$$

Here, $a_B = \frac{4\pi\hbar^2\epsilon_0}{me^2} \simeq 0.53$ Å is the Bohr radius and r_s is the Wigner-Seitz radius, defined by $(4\pi/3)r_S^3 = n^{-1}$.

For the induced number density we obtain:

$$n_{ind}(\mathbf{q}) = \frac{\epsilon_0 q^2}{e}\frac{V_{ext}(\mathbf{q})}{[1 + q^2/q_{TF}^2]} \ , \tag{2.33}$$

Dielectric permittivity. In general, this phenomenon is incorporated into electromagnetic theory through the generalised wavevector dependent dielectric function $\epsilon(\mathbf{q})$. The dielectric function relates the electric displacement D to the electric field $\mathbf{E}$, in the form $\epsilon_0\epsilon(\mathbf{q})\mathbf{E}(\mathbf{q}) = \mathbf{D}(\mathbf{q})$. While the gradient of the total potential $V_{tot} = V_0 + \delta V + V_{ext} = \delta V + V_{ext}$ ($V_0 = 0$ for Jellium) gives the $\mathbf{E}-$ field, the gradient of the externally applied potential V_{ext} gives the displacement field $\mathbf{D}$. As $\mathbf{E}$ and $\mathbf{D}$ are related via the relative permittivity, ϵ, the potentials from which they derive are also connected by ϵ:

$$V_{ext}(\mathbf{q}) = \epsilon(\mathbf{q})\left(\delta V(\mathbf{q}) + V_{ext}(\mathbf{q})\right) \tag{2.34}$$

Using Eq. 2.31 we find

$$V_{tot}(\mathbf{q}) = V_{ext}(\mathbf{q})\frac{q^2}{q^2 + q_{TF}^2} \ , \tag{2.35}$$

and for $\epsilon(q)$:

$$\epsilon^{TF}(q) = 1 + \frac{q_{TF}^2}{q^2} \quad . \tag{2.36}$$

Screening. $\epsilon_{TF} \propto q^{-2}$ at small q (long distances), so the long range part of the Coulomb potential (also $\propto 1/q^2$) is *exactly* cancelled. In real space, if $v_{ext} = Q/r$ is Coulombic (long range), $V(r) = (Q/r)e^{-q_{TF}r}$ is a short-range Yukawa, or screened potential[3]. In a typical metal, r_s is in the range $2-6$, and so potentials are screened over a distance comparable to the interparticle spacing; the electron gas is highly effective in shielding external charges.

[3]This form is originally due to P. Debye and E. Hückel, *Zeitschrift für Physik* **24**, 185, (1923) and was derived for the theory of electrolytes; it appears also in particle theory under the name of the Yukawa potential; the physics in these cases is identical

Chapter 3

From atoms to solids

What holds a solid together? Cohesion is ultimately produced by the interactions between the nuclei and the electrons, which give rise to an effective interaction potential between atoms. We distinguish between a number of distinct mechanisms (or types of bonds), however, which can contribute to this.

3.1 The binding of crystals

Inert gases

The inert gases have filled electron shells and large ionisation energies. Consequently, the electronic configuration in the solid is close to that of separated atoms. Since the atoms are neutral, the interaction between them is weak, and the leading attractive force at large distances arises from the van der Waals interaction, which gives an attractive potential proportional to $1/R^6$.

This form can be loosely derived by thinking of an atom as an oscillator, with the electron cloud fluctuating around the nucleus as if on a spring. The centre of the motion lies on top of the atom, but if the cloud is displaced, there will be a small dipole induced, say p_1. Such displacements occur as a result of zero-point motion of the electron cloud in the potential of the nucleus. A distance R away from the atom there is now an induced electric field $\propto p_1/R^3$. A second atom placed at this point will then have a dipole induced by the electric field of the first: $p_2 \propto \alpha p_1/R^3$, where α is the atomic polarizability. The second dipole induces an electric field at the first, which is now

$$E_1 \propto p_2/R^3 \propto \alpha p_1/R^6. \tag{3.1}$$

The energy of the system is then changed by an amount

$$\Delta U = \langle -p_1 \cdot E_1 \rangle \propto -\alpha \left\langle p_1^2 \right\rangle / R^6. \tag{3.2}$$

Notice that ΔU depends on the expectation value of the square of the dipole moment $< p_1^2 >$, which is non-zero, and not the square of the expectation value $< p_1 >^2$, which would be zero.

If the atoms move together so that the electron charge distributions begin to overlap, repulsive forces come into play. While there is of course a contribution from the direct electrostatic

Figure 3.1: Two dipoles represent model atoms that are arranged along a line, with the positive charges (+e) fixed at the positions $0, R$, and the negative charges (-e) at the points $x_1, R + x_2$.

repulsion of the electrons, more important is the Pauli exclusion principle that prevents two electrons having equal quantum numbers. The effect of Pauli exclusion can be seen by an extreme example, of overlapping two Hydrogen atoms entirely, with the electrons for simplicity assumed to be in the same spin state. In this case, while two separated atoms may be both in the $1S$ ground state, the combined molecule must have a configuration $1s2s$, and thus is higher by the promotion energy.

Calculations of the repulsive interaction are complex but the answer is clearly short-ranged. They are often modelled empirically by an exponential form e^{-R/R_o}, or a power law with a large power. A commonly used empirical form to fit experimental data for inert gases is the Lennard-Jones potential

$$U(R) = -\frac{A}{R^6} + \frac{B}{R^{12}} \tag{3.3}$$

with $theA$ and B atomic constants obtained from gas-phase data.

With the exception of He, the rare gases from close-packed (face-centered cubic) solids with a small cohesive energy, and low melting temperatures. Helium is special because zero-point motion of these light atoms is substantial enough that they do not solidify at zero pressure down to the absolute zero of temperature. The quantum fluids ^3He and ^4He have a number of extraordinary properties, including superfluidity.

Ionic Crystals

Given the stability of the electronic ground state configurations of a rare gas, atoms that are close to having a filled shell will have a tendency to lose or gain electrons to fill the shell.

- The energy for the reaction $M \to M^+ + e^-$ in the gas phase is the ionization energy I.
- The energy for the reaction $X + e^- \to X^-$ in the gas phase is the electron affinity A.
- Although it costs an energy $I + A$ to form an ionic molecules, a greater energy reduction can be obtained from the electrostatic attraction between the charges, e^2/R.
- In a solid, the electrostatic interaction energy for a diatomic crystal[1] is

$$U_{electrostatic} = \frac{1}{2}\sum_i \sum_j U_{ij} \tag{3.4}$$

[1]Note the factor of 1/2, which avoids double counting the interaction energy. The energy of a single ion i due to interaction with all the other ions is $U_i = \sum_{j \neq i} U_{ij}$; the total energy is $\frac{1}{2}\sum_i U_i$.

where $U_{ij} = \pm q^2/R_{ij}$ is the sum of all Coulomb forces between ions. If the system is on a regular lattice of lattice constant R, then we write the sum

$$U_{electrostatic} = -\frac{1}{2}\frac{\alpha_M q^2}{R} \tag{3.5}$$

where α_M is a dimensionless constant that depends only on the crystal structure.

- The evaluation of α_M is tricky, because the sum converges slowly. Three common crystal structures are $NaCl$ ($\alpha_M = 1.7476$), $CsCl$ (1.7627), and cubic ZnS or Zincblende (1.6381).
- To the attractive Madelung term must be added the repulsive short range force, and we now have the added caveat that ions have different sizes, explaining why $NaCl$ has the rocksalt structure, despite the better electrostatic energy of the $CsCl$ structure.

Covalent crystals

The covalent bond is the electron pair or single bond of chemistry.

Model Hydrogen. Two overlapping atomic orbitals on identical neighbouring atoms will hybridise. Because the Hamiltonian must be symmetric about a point centered between the ions, the eigenstates must have either even or odd parity about this center. If we have a simple system of two one electron atoms - model hydrogen - which can be approximated by a basis of atomic states $\phi(r - R)$ (assumed real) centered on the nucleus R, then two states of even and odd parity are

$$\psi_\pm(r) = \phi(r - R_a) \pm \phi(r - R_b) \tag{3.6}$$

ψ_+ has a substantial probability density between the atoms, where ψ_- has a node. Consequently, for an attractive potential $E_+ < E_-$, and the lower (*bonding*) state will be filled by two electrons of opposite spin. The *antibonding* state ψ_- is separated by an energy gap $E_g = E_- - E_+$ and will be unfilled. The cohesive energy is then approximately equal to the gap E_g.[2].

Covalent semiconductors. If we have only s-electrons, we clearly make molecules first, and then a weakly bound molecular solid, as in H_2. Using p, d, orbitals, we may however make *directed* bonds, with the classic case being the sp^3 hybrid orbitals of C, Si, and Ge. These are constructed by hybrid orbitals $s + p_x + p_y + p_z$ + 3 other equivalent combinations, to make new orbitals that point in the four tetrahedral directions: (111), $(\bar{1}\bar{1}1)$, $(\bar{1}1\bar{1})$, $(1\bar{1}\bar{1})$. These directed orbitals make bonds with neighbours in these tetrahedral directions, with each atom donating one electron. The open tetrahedral network is the familiar diamond structure of C, Si and Ge.

Ionic semiconductors. In $GaAs$ and cubic ZnS the total electron number from the pair of atoms satisfies the "octet" rule, and they have an identical tetrahedral arrangement found in diamond, but with the atomic types alternating. This is called the *zincblende* structure. The cohesion in these crystals is now partly ionic and partly covalent. There is another locally tetrahedral arrangement called *wurtzite* which has a hexagonal lattice and is favoured in more ionic systems. With increasing ionic components to the bonding, the structures change to

[2]Actually twice (two electrons) half the gap, if we assume that $E_\pm = E_{atom} \pm \frac{1}{2}E_g$

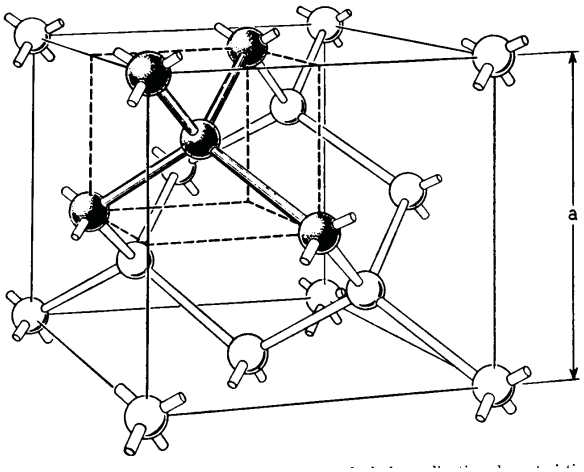


Figure 3.2: Tetrahedral bonding in the diamond structure. The zincblende structure is the same but with two different atoms per unit cell

reflect the ionicity: group IV Ge (diamond), III-V $GaAs$ (Zincblende), II-VI ZnS (zincblende or wurtzite), II-VI $CdSe$ (wurtzite), and I-VII $NaCl$ (rocksalt).

Metals

Metals are generally characterised by a high electrical conductivity, arising because the electrons are relatively free to propagate through the solid.

Close packing. Simple metals (e.g., alkalis such as Na, and $s-p$ bonded metals such as Mg and Al) are usually highly coordinated (i.e., fcc or hcp with 12 nearest neighbours, or sometimes bcc with 8 nearest neighbours), since the proximity of many neighbouring atoms facilitates hopping between neighbours. Remember that the Fermi energy of a free electron gas (i.e., the average kinetic energy per particle) is proportional to $k_F^2 \propto a^{-2} \propto n^{2/3}$ (here a is the lattice constant and n the density; the average Coulomb interaction of an electron in a solid with all the other electrons and the other ions is proportional to $a^{-1} \propto n^{1/3}$. Thus the higher the density, the larger the kinetic energy relative to the potential energy, and the more *itinerant* the electrons.[3] By having a high coordination number, one can have relatively

[3]Note the contrast to classical matter, where solids are stabilised at higher density, and gases/liquids at lower density.

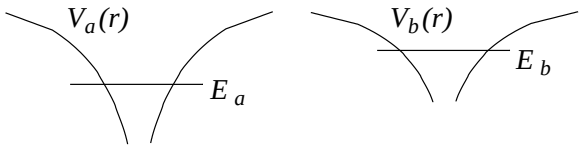


Figure 3.3: A simple model of a diatomic molecule. The atomic Hamiltonian is $H_i = T + V_i(r)$, with T the kinetic energy $-\hbar^2\nabla^2/2m$ and V_i the potential. We keep just one energy level on each atom.

large distances between neighbours - minimising the kinetic energy cost - in comparison to a loose-packed structure of the same density.

Screening. Early schooling teaches us that a metal is an equipotential (i.e., no electric fields). We shall see later that this physics in fact extends down to scales of the screening length $\lambda \approx 0.1nm$, i.e., about the atomic spacing (although it depends on density) - so that the effective interaction energy between two atoms in a metal is not Z^2/R (Z the charge, R the separation), but $Z^2 e^{-R/\lambda}/R$ and the cohesion is weak.

Trends across the periodic table. As an $s-p$ shell is filled (e.g., Na, Mg, Al, Si) the ion core potential seen by the electrons grows. This makes the density of the metal tend to increase with atomic number. Eventually, the preference on the right-hand side of the periodic table is for covalent semiconducting (Si, S) or insulating molecular (P, Cl) structures, because the energy is lowered by making tightly bound directed bonds.

Transition metals. Transition metals and their compounds involve both the outer $s-p$ electrons as well as inner d-electrons in the binding. The d-electrons are more localised and often are spin-polarised in the $3d$ shell when they have a strong atomic character (magnetism will be discussed later in the course). For $4d$ and $5d$ transition metals, the d-orbitals are more strongly overlapping from atom to atom and this produces the high binding energy of metals like W (melting point 3700 K) in comparison to alkali metals such as Cs (melting point 300 K).

3.2 Complex matter

Simple metals, semiconductors, and insulators formed of the elements or binary compounds like $GaAs$ are only the beginning of the study of materials. Periodic solids include limitless possibilities of chemical arrangements of atoms in compounds. Materials *per se*, are not perhaps so interesting to the physicist, but the remarkable feature of condensed matter is the wealth of physical properties that can be explored through novel arrangements of atoms.

Many new materials, often with special physical properties, are discovered each year. Even for the element carbon, surely a familiar one, the fullerenes (e.g., C_{60}) and nanotubes (rolled up graphitic sheets) are recent discoveries. Transition metal oxides have been another rich source of discoveries (e.g., high temperature superconductors based on La_2CuO_4, and ferromagnetic metals based on $LaMnO_3$). f-shell electron metals sometimes produce remarkable electronic properties, with the electrons within them behaving as if their mass is 1000 times larger than the free electron mass. Such quantum fluid ground states (metals, exotic superconductors, and superfluids) are now a rich source of research activity. The study of artificial meta-materials begins in one sense with doped semiconductors (and especially layered heterostructures grown by *molecular beam epitaxy* or MBE), but this subject is expanding rapidly due to an influx of new tools in nano-manipulation and biological materials.

Many materials are of course not crystalline and therefore not periodic. The physical description of *complex* and *soft* matter requires a separate course.

Glasses

If one takes a high temperature liquid (e.g., of a metal) and quenches it rapidly, one obtains a *frozen* structure that typically retains the structure of the high-temperature liquid. Melt-quenched alloys of ferromagnets are often prepared this way because it produces isotropic magnetic properties. For most materials the amorphous phase is considerably higher in energy than the crystalline one, so the system has to be frozen rapidly, far from its equilibrium configuration. A few materials make glassy states readily, and the most common example is vitreous silica (SiO_2). Many crystalline forms of silica exist consisting of network structures in which each Si atom bonds to four oxygen neighbours (approximately tetrahedrally) and each O atom is bonded to two Si atoms. Since the O^{2-} ion is nearly isotropic, the orientation of one tetrahedral group with respect to a neighbouring group about the connecting $Si-O-Si$ bond is not fixed, and this allows for many possible crystalline structures, especially for the entropic stabilisation of the glass phase. Whatever the arrangement of the atoms, all the electrons are used up in the bonding, so glass is indeed a good insulator. The characteristic feature of a strong or network glass is that on cooling the material becomes increasingly viscous, often following the *Vogel-Fulcher* law,

$$\eta \propto e^{\frac{C}{T-T_0}} \tag{3.7}$$

implying a divergence in the viscosity η at a temperature T_0. Once η reaches about 10^{12} Pa s, it is no longer possible to follow the equilibrium behaviour. Consequently, debates still rage about whether or not the glass transition is a "true" phase transition, or indeed whether or no the temperature T_0 has physical meaning.

Polymers

The classic polymers are based on carbon, relying on its remarkable ability to adopt a variety of local chemical configurations. Polyethylene is built from repeating units of CH_2, and more complex polymers are constructed out of more complex subunits. Because the chains are long, and easily deformed or entangled, most polymers are glassy in character, and therefore their physical properties are largely dominated by entropic considerations. The elasticity of rubber is produced by the decrease in entropy upon stretching, not by the energetic cost of stretching the atomic bonds. Many simple polymers are naturally insulating (e.g., the alkanes) or semiconducting, but it is sometimes possible to "dope" these systems so that there are unfilled electronic states. They have become interesting in technology and fundamental science. Because a simple polymer chain can often be modelled as a one-dimensional wire, they provide a laboratory for the often unusual properties of one-dimensional electronic systems. Because the tools of organic chemistry allow one to modify the physical properties of polymers in a wide range of ways (for example, by adding different side chains to the backbones). One can attempt to tune the electronic and optical properties of heterogeneous polymer structures to make complex devices (solar cells, light-emitting diodes, transistors) using a very different medium from inorganic semiconductors.

Liquid crystals

Polymers are isotropic, because they are very long and they curl up. Shorter rod-shaped molecules, however, have an obvious orientational axis, and when combined together to make

a liquid crystal one can construct matter whose properties are intermediate between a liquid and solid.

Nematics. An array of rods whose centres are arranged randomly has no long-range positional order (just like a liquid), but if the rods are oriented parallel to each other has *long-range orientational order*, as in a molecular crystal. This is a nematic liquid crystal. The direction in space of the orientational order is a vector $\hat{n}$ called the *director*. The refractive index of the material will now be different for light polarized parallel and perpendicular to the director.

Cholesterics. It turns out that if the molecule is *chiral* the director need not always point in the same direction, and in a cholesteric liquid crystal the direction of $\hat{n}$ twists slowly in a helix along an axis that is perpendicular to it. Usually the pitch of the twist is much longer than size of the rod, is a strong function of temperature, and frequently close to the wavelength of visible light.

Smectics. Smectics additionally have long-range positional order along one direction, usually to be thought of as having layers of molecules. So called *smectic A* has the director parallel to the planes, whereas in *smectic C* the director is no longer perpendicular (and may indeed rotate as a function of position). In *smectic B* the molecules in the plane have a crystalline arrangement, but different layers fall out of registry. This is a kind of quasi-2D solid.

Quasicrystals

As a last piece of exotica, the classic group theory of crystal structures proves the impossibility of building a Bravais lattice with five-fold symmetry. Nature is unaware of this, and a series of metallic alloys have been found that indeed have crystals with axes of three, five, and ten-fold symmetry. These materials are in fact physical representations of a mathematical problem, introduced by Penrose, of tiling of a plane with, e.g., two rhombus shaped tiles that have corner angles of $2\pi/10$ and $2\pi/5$. A complete tiling of the plane is possible, though the structure is not a periodic lattice (it never repeats).

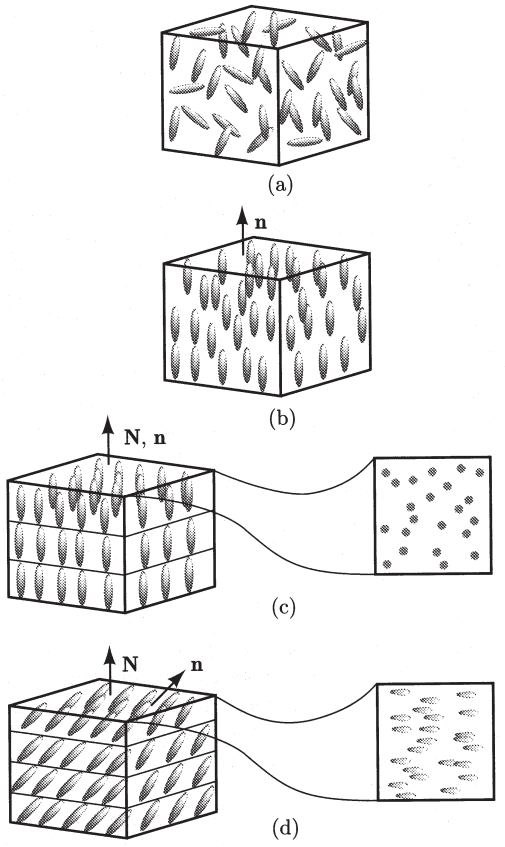


Figure 3.4: Liquid crystal structures
Schematic representation of the position and orientation of anisotropic molecules in: (a) the isotropic phase; (b) the nematic phase; (c) the smectic-A phase; and (d) the smectic-C phase. [From Chaikin and Lubensky]

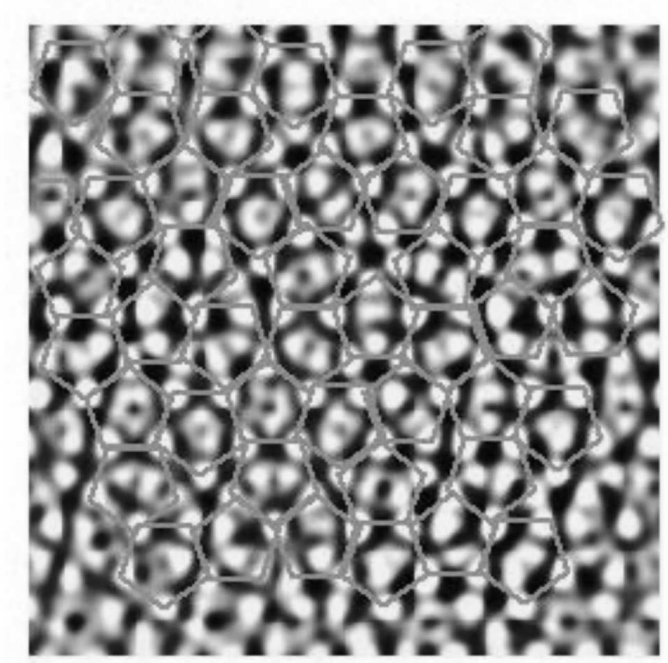


Figure 3.5: Scanning tunnelling microscope image of a 10 nm^2 quasicrystal of $AlPdMn$ with a Penrose tiling overlaid. [Ledieu et al. Phys.Rev.B **66**, 184207 (2002)]

3.3 The description of periodic solids

An ideal crystal is constructed from an infinite repetition of identical structural units in space. The repeating structure is called the *lattice*, and the group of atoms that is repeated is called the *basis*. The basis may be as simple as a single atom, or as complicated as a polymer or protein molecule. This section discusses briefly some important definitions and concepts. For a more complete description with examples, see any of the textbooks recommended in the introduction.

Lattice. The lattice is defined by three fundamental (called *primitive*) translation vectors $\mathbf{a}_i$, $i = 1, 2, 3$, which define primitive lattice translation operations. An arbitrary lattice translation operation can be written as

$$\mathbf{T} = \sum_i n_i \mathbf{a}_i \tag{3.8}$$

The atomic arrangement looks the same from equivalent points in the unit cell:

$$\mathbf{r}' = \mathbf{r} + \sum_i n_i \mathbf{a}_i \quad \forall \text{ integer } n_i \, . \tag{3.9}$$

The lattice so formed is called a *Bravais* lattice.

Primitive unit cell. A **unit cell** is a part of the crystal, which – when translated repeatedly by the primitive translation vectors – can fill the entire volume of the crystal. However, there may be overlapping regions. For example, a face centred cubic lattice (fcc) can have a simple cubic unit cell. Because this is the simplest one to draw and to work with, it is called the **conventional unit cell**. It contains more than one lattice site (four, in fact). When this cell is translated by one of the primitive translation vectors (e.g., $a(1/2, 1/2, 0)$), there will be overlaps. A **primitive unit cell**, however, would exactly fill the volume of the crystal, when repeatedly translated. It contains exactly one lattice point. One way of constructing a primitive unit cell, is by forming a parallelepiped from a set of primitive translation vectors $\mathbf{a}_i$.

Wigner-Seitz cell. A convenient alternative primitive unit cell to use is the *Wigner-Seitz* cell. This is the region in space around a lattice point, which is closer to this lattice point than to any other lattice point. It can be constructed as follows: Draw lines to connect a given lattice point to all of its near neighbours. Then draw planes normal to each of these lines from the midpoints of the lines. The smallest volume enclosed in this way is the Wigner-Seitz primitive unit cell. You can show that the Wigner-Seitz cell has the same symmetry properties as the lattice itself, which is not true of all the choices of primitive unit cell. The first Brillouin zone is the Wigner-Seitz cell in reciprocal space.

Point group. The symmetry operations on a lattice consist of translations, rotations and reflections. The set of symmetry operations which, when applied about a lattice point, map the lattice onto itself is the *point group* of the lattice. This includes reflections and rotations; for example a 2D square lattice is invariant under reflections about the x and y axes, as well as through axes at an angle of $\pi/4$ to the x and y axes, and rotations through any multiple of $\pi/2$. Remember that adding a basis to a primitive lattice may destroy some of the point group symmetry operations.

Space group. The translational symmetry and the point group symmetries are subgroups of the full symmetry of the lattice which is the *space* group. Every operation in the space group

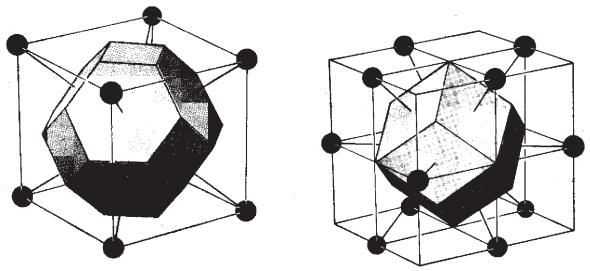


Figure 3.6: . The Wigner-Seitz cell for the BCC and FCC lattices

consists of a rotation, reflection, or inversion followed by a translation. However, the space group is not necessarily just the sum of the translational symmetries and the point symmetries, because there can be space group symmetries that are the sum of a proper rotation and a translation, neither of which are independently symmetries of the lattice.

Lattice types. For *Bravais lattices*, there are five distinct lattice types in two dimensions, and 14 in three dimensions. These are made up of seven crystal classes, each of which can have face centred, body centred or side centred variants: (i) cubic (simple, face centred, body centred), (ii) tetragonal (simple, body centred), (iii) orthorhombic (i.e., brick-shaped: simple, face centred, body centred), (iv) hexagonal, (v) monoclinic (i.e., rectangular footprint, but oblique out of the plane, $\alpha = \gamma = 90°, \beta \neq 120°$: simple, side-centred), (vi) triclinic (all angles $\neq 90°$, all lengths different), and (vii) trigonal (all angles $= 90°$, all lengths the same).

The number of possible **lattices with bases** is large but finite. In three dimensions, for lattices with bases, there are 32 distinct point groups, and 230 possible space groups. Two of the important lattices that we shall meet frequently are the body-centred and face-centred cubic lattices, shown in Fig. 3.6.

Index system for crystal planes

Know the coordinates of three lattice points (not collinear) is enough to define a crystal plane. Suppose you chose each point to lie along a different crystal axis, the plane is then specified by giving the coordinates of the three intersection points as

$$\{\mathbf{r}_i\} = \{x\mathbf{a}_1, y\mathbf{a}_2, z\mathbf{a}_3\} \tag{3.10}$$

The triad (xyz) need not be integers. However, one can always generate a set of parallel planes by translating the original plane by a lattice vector. One of these parallel planes will intersect the axes at integer multiples of the lattice vectors.

The set of three integers (hkl) where $xh = yk = zl =$ integer is called the *Miller index* of the plane. For instance, if $x = 1/2$, $y = 1/2$, $z = 1$, then $h = 2$, $k = 2$, $l = 1$ (as in Fig. 3.7) and $hx = ky = lz = 1$. Often, the Miller index is simply obtained from the inverses of the axis intercepts. For $x = 2/3$, $y = 4/3$, $z = 1$, however, the situation is more complicated, and we find $h = 6$, $k = 3$, $l = 4$, and the product $hx = ky = lz = 4$. The definition of the Miller index is more transparent when considering reciprocal lattice vectors (below and question on problem sheet), where it emerges that the reciprocal space vector $h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3$ (where the

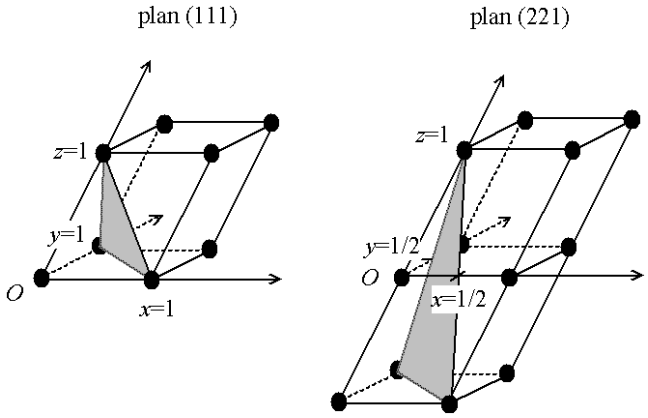


Figure 3.7: Definition of lattice planes and Miller indices. If a plane intersects the axes at $x\mathbf{a}_1$, $y\mathbf{a}_2$ and $z\mathbf{a}_3$, then the Miller index (hkl) is given by the smallest set of values h, k, l for which $hx = ky = lz =$ integer.

$\mathbf{b}_i$ are primitive lattice vectors in reciprocal space) is normal to the lattice plane with index (hkl) defined as above.

When we refer to a set of planes that are equivalent by symmetry, we use a curly bracket notation: $\{100\}$ for a cubic crystal denotes the six equivalent symmetry planes (100), (010), (001), $(\bar{1}00)$, $(0\bar{1}0)$, $(00\bar{1})$, with the overbar used to denote negation.

3.4 The reciprocal lattice and diffraction

The reciprocal lattice as a concept arises from the theory of the scattering of waves by crystals. You should be familiar with the diffraction of light by a 2-dimensional periodic object - a diffraction grating. Here an incident plane wave is diffracted into a set of different directions in a Fraunhofer pattern. An infinite periodic structure produces outgoing waves at particular angles, which are determined by the periodicity of the grating. What we discuss now is the generalisation to scattering by a three-dimensional periodic lattice.

First calculate the scattering of a single atom (or more generally the basis that forms the unit cell) by an incoming plane wave, which should be familiar from elementary quantum mechanics. An incoming plane wave of wavevector $\mathbf{k}_o$ is incident on a potential centred at the point $\mathbf{R}$. At large distances the scattered wave take the form of a circular wave. (See figure Fig. 3.8).

The total field (here taken as a scalar) is then

$$\psi \propto e^{i\mathbf{k}_o\cdot(\mathbf{r}-\mathbf{R})} + cf(\hat{r})\frac{e^{ik_o|\mathbf{r}-\mathbf{R}|}}{|\mathbf{r}-\mathbf{R}|}. \tag{3.11}$$

All of the details of the scattering are buried in the *form factor* $f(\hat{r})$ which is a function of the scattering angle, the arrangement and type of atom, etc. The total scattered intensity is just set by the coefficient c and we will assume it is small (for this reason we do not consider multiple scattering by the crystal).

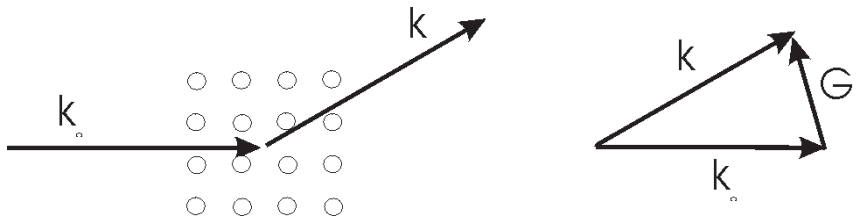


Figure 3.8: Illustration of Bragg scattering from a crystal

For sufficiently large distance from the scatterer, we can write

$$k_o|\mathbf{r}-\mathbf{R}| \approx k_o r - k_o \frac{\mathbf{r}\cdot\mathbf{R}}{r} \tag{3.12}$$

Define the scattered wavevector

$$\mathbf{k} = k_o \frac{\mathbf{r}}{r} \tag{3.13}$$

and the momentum transfer

$$\mathbf{q} = \mathbf{k}_o - \mathbf{k} \tag{3.14}$$

we then have for the waveform

$$\psi \propto e^{i\mathbf{k}_o\cdot\mathbf{r}}\left[1 + cf(\hat{r})\frac{e^{i\mathbf{q}\cdot\mathbf{R}}}{r}\right] \quad . \tag{3.15}$$

Now sum over all the identical sites in the lattice, and the final formula is

$$\psi \propto e^{i\mathbf{k}_o\cdot\mathbf{r}}\left[1 + c\sum_i f_i(\hat{r})\frac{e^{i\mathbf{q}\cdot\mathbf{R}_i}}{r}\right] \quad . \tag{3.16}$$

Away from the forward scattering direction, the incoming beam does not contribute, and we need only look at the summation term. We are adding together terms with different phases $\mathbf{q}\cdot\mathbf{R}_i$, and these will lead to a cancellation unless the Bragg condition

$$\mathbf{q}\cdot\mathbf{R} = 2\pi m \tag{3.17}$$

for all $\mathbf{R}$ in the lattice is satisfied, and with m an integer (that depends on $\mathbf{R}$). The special values of $\mathbf{q} \equiv \mathbf{G}$ that satisfy this requirement lie on a lattice, which is called the *reciprocal lattice.* [4]

One can check that the following prescription for the reciprocal lattice will satisfy the Bragg condition. The primitive vectors $\mathbf{b}_i$ of the reciprocal lattice are given by

$$\mathbf{b}_1 = 2\pi\frac{\mathbf{a_2}\wedge\mathbf{a_3}}{\mathbf{a}_1\cdot\mathbf{a}_2\wedge\mathbf{a}_3} \quad \text{and cyclic permutations} \quad . \tag{3.18}$$

[4]We can be sure that they are on a lattice, because if we have found any two vectors that satisfy (3.17), then their sum also satisfies the Bragg condition.

3.5 Diffraction conditions and Brillouin zones

For elastic scattering, there are two conditions relating incident and outgoing momenta. Conservation of energy requires that the magnitudes of k_o and k are equal, and the Bragg condition requires their difference to be a reciprocal lattice vector $\mathbf{k} - \mathbf{k}_o = \mathbf{G}$. The combination of the two can be rewritten as

$$\mathbf{k} \cdot \frac{\mathbf{G}}{2} = \left(\frac{G}{2}\right)^2 \quad . \tag{3.19}$$

(3.19) defines a plane constructed perpendicular to the vector $\mathbf{G}$ and intersecting this vector at its midpoint. The set of all such planes defines those incident wavevectors that satisfy the conditions for diffraction (see Fig. 3.9).

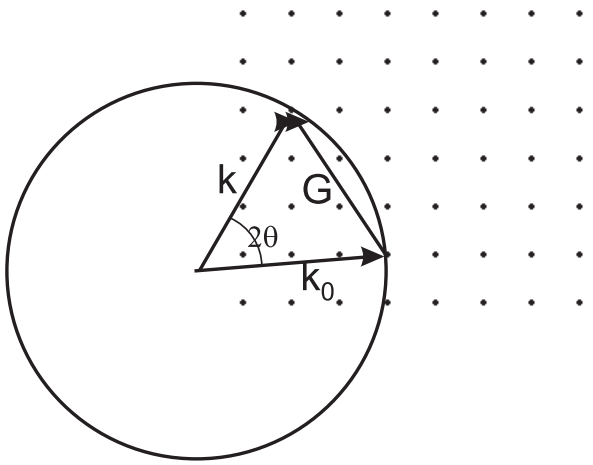


Figure 3.9: Ewald construction. The points are the reciprocal lattice of the crystal. k_0 is the incident wavevector, with the origin chosen so that it terminates on a reciprocal lattice point. A sphere of radius $|k_0|$ is drawn about the origin, and a diffracted beam will be formed if this sphere intersects any other point in the reciprocal lattice. The angle θ is the Bragg angle of (3.21)

This condition is familiar as *Bragg's Law*. The condition (3.19) may also be written as

$$\frac{2\pi}{\lambda} \sin\theta = \frac{\pi}{d} \tag{3.20}$$

where $\lambda = 2\pi/k$, θ is the angle between the incident beam and the crystal planes perpendicular to $\mathbf{G}$, and d is the separation between the plane and the origin.

Since the indices that define an actual crystal plane may contain a common factor n, whereas the definition used earlier for a *set* of planes removed it, we should generalise (3.20) to define d to be the spacing between adjacent parallel planes with indices h/n, k/n, l/n. Then we have

$$2d \sin\theta = n\lambda \tag{3.21}$$

which is the conventional statement of Bragg's Law.

To recap:

- The set of crystal planes that satisfy the Bragg condition can be constructed by finding those planes in reciprocal space which are perpendicular bisectors of every reciprocal lattice vector $\mathbf{G}$. A wave whose wavevector drawn from the origin terminates in any of these planes satisfies the condition for elastic diffraction.
- The planes constructed in this way divide reciprocal space up into cells. The one closest to the origin is called the first Brillouin zone. The n^{th} Brillouin zone consists of all the fragments exterior to the $(n-1)^{th}$ plane (measured from the origin) but interior to the n^{th} plane.
- The first Brillouin zone is the Wigner-Seitz cell of the reciprocal lattice. This will play an important role in the discussion of electronic states in a periodic potential.
- The volume of each Brillouin zone (adding up the fragments) is equal to the volume of the primitive unit cell of the reciprocal lattice, which is $(2\pi)^3/\Omega_{cell}$ where Ω_{cell} is the volume of the primitive unit cell of the crystal.

3.6 Lattice dynamics and phonons

One-dimensional monatomic chain

Our model consists of identical atoms connected by springs, shown in Fig. 3.10

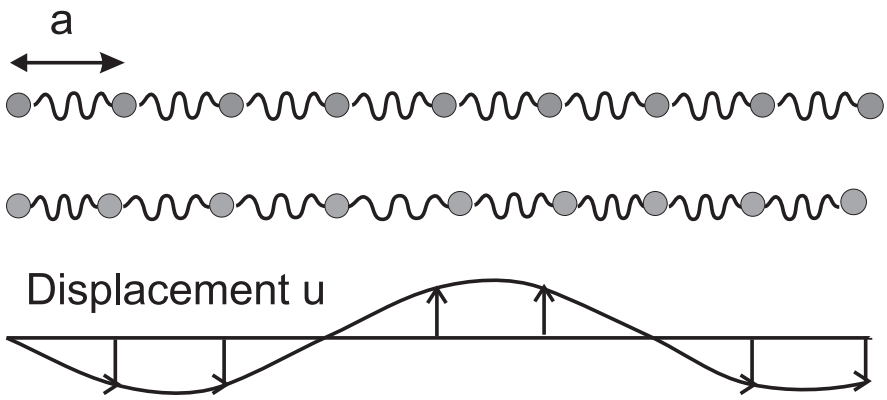


Figure 3.10: A one-dimensional linear chain. The atoms are shown in their equally spaced equilibrium conditions in the top row, and with a periodic distortion below. In the bottom figure the displacements u_n are plotted as arrows, and the curve shows sinusoidal variations with a period $6a$, in this case.

In equilibrium, the atoms are uniformly spaced at a distance a, and we now look for oscillations about the equilibrium position. We assume the crystal is harmonic, so that the spring restoring varies linearly with the extension. If we take the displacement of the n^{th} atom (which is at the point $r_n = na$) to be u_n, its equation of motion is

$$m\frac{\partial^2 u_n}{\partial t^2} = K(u_{n+1} - u_n) + K(u_{n-1} - u_n) \tag{3.22}$$

We guess that the solution is a wave, of the form

$$u_n(t) = u_o \cos(qr_n - \omega(q)t) \tag{3.23}$$

Here the wavelength is $\lambda = 2\pi/q$, and the period is $T = 2\pi/\omega(q)$; to check that this is a solution, and to determine the frequency, we substitute into the equation of motion. This is left as an exercise, and a few lines of algebra will show that the solution (3.23) exists provided that

$$m\omega^2(q) = 2K(1 - \cos(qa)) = 4K\sin^2(\frac{qa}{2}) \tag{3.24}$$

so that

$$\omega(q) = 2(K/m)^{1/2}\sin\left(\frac{qa}{2}\right) \tag{3.25}$$

(3.24) is called a dispersion relation — the relation between the frequency of the mode and its wavevector, or equivalently the relationship between the wavelength and the period.

The wavevector q is inversely related to the wavelength; note that for long wavelength modes (i.e., $q \to 0$), the relationship is linear, viz

$$\omega(q) = (K/m)^{1/2}(qa) \tag{3.26}$$

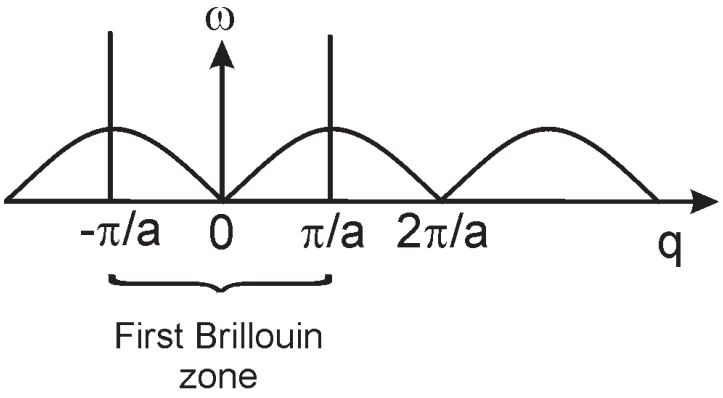


Figure 3.11: Dispersion relation between frequency and wavevector for a one-dimensional monatomic chain

, which is the same as for a wire with tension Ka and density m/a. In the long wavelength limit, we have compressive sound waves that travel with a velocity $v = a(K/m)^{1/2}$. Because this kind of wave behaves like a sound wave, it is called an acoustic mode.

The dispersion is not linear for larger values of q, and is in fact periodic (Fig. 3.11). The periodicity can easily be understood by reference to (3.23). Suppose we choose $q = 2\pi/a$. Note then that

$$qr_n = \frac{2\pi}{a} \times na = 2\pi n \tag{3.27}$$

so that all the atoms displace together, just as if $q = 0$. In general it is straightforward to show that if one replaces q by $q + \text{integer} \times 2\pi/a$, then the displacements are unchanged – so we may simplify our discussion by using only q vectors in the range

$$-\frac{\pi}{a} \leq q \leq \frac{\pi}{a} \quad . \tag{3.28}$$

This is called the first Brillouin zone.

One-dimensional diatomic chain

The monatomic chain contains only acoustic modes, but the phonon spectrum becomes more complex if there are more atoms per unit cell. As an illustration, we look at the diatomic chain.

For simplicity, we use again a phenomenological model of balls and springs, but now with two different atoms in the unit cell, two different masses and two different spring constants (see Fig. 3.12). We can now write down two equations of motion, one for each type of atom:

Figure 3.12: Diatomic chain

$$\begin{aligned} m_A \frac{\partial^2 u_{nA}}{\partial t^2} &= K(u_{nB} - u_{nA}) + K'(u_{n-1,B} - u_{nA}) \\ m_B \frac{\partial^2 u_{nB}}{\partial t^2} &= K'(u_{n+1A} - u_{nB}) + K(u_{n,A} - u_{nB}) \end{aligned} \tag{3.29}$$

The solution of this is a little more complicated than before, but we can now intuitively see that there ought to be a new type of phonon mode by considering a particular limit of the parameters. Suppose the two atoms are quite strongly bound together in pairs, as sketched in the figure above: then we might expect that $K \gg K'$, and to a first approximation the pairs can be treated as independent molecules. (We will also simplify the analysis by taking $m_A = m_B = m$.) Then every molecule will have a vibrational mode where the two atoms oscillate out of phase with each other with a frequency

$$\omega_o^2 = 2K/m \quad . \tag{3.30}$$

The corresponding coordinate which undergoes this oscillation is

$$u_{opt}(q = 0) = u_A - u_B \tag{3.31}$$

where I have explicitly remarked that this is at $q = 0$ if each molecule undergoes the oscillation in phase with the next.

We can of course make a wavelike solution by choosing the correct phase relationship from one unit cell to the next — as sketched in Fig. 3.13, but if $K' \ll K$ this will hardly change the restoring force at all, and so the frequency of this so-called optical phonon mode will be almost independent of q.

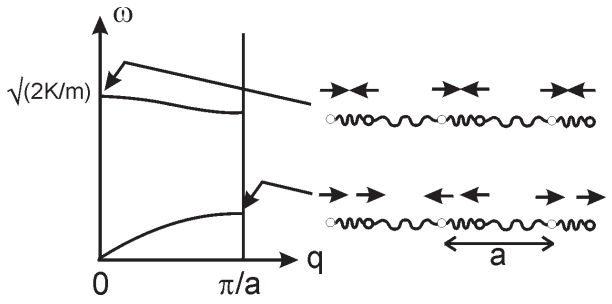


Figure 3.13: Dispersion of the optical and acoustic phonon branches in a diatomic chain, and a schematic picture of the atomic displacements in the optical mode at $q = 0$.

There are now two branches of the dispersion curve, along one of which the frequency vanishes linearly with wavevector, and where the other mode has a finite frequency as $q \to 0$(see Fig. 3.14). The name "optical" arises because at these long wavelengths the optical phonons can interact (either by absorption, or scattering) with light, and are therefore prominent features in the absorption and Raman spectra of solids in the infrared spectrum.

Phonons in three-dimensional solids

The descriptions above are not too hard to generalise to three- dimensional solids, although the algebra gets overloaded with suffices.

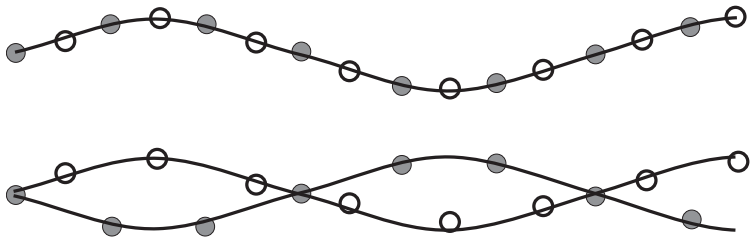


Figure 3.14: Pattern of atomic displacements for an acoustic and an optical phonon of the same wavevector.

Rather than a one-dimensional wavevector k corresponding to the direction of the 1D chain, there is now a three-dimensional dispersion relation $\omega(\mathbf{k})$, describing waves propagating in different directions.

Also, there are not just compressional waves, but also transverse, or shear waves, that have a different dispersion from the longitudinal (compressional) waves. (These exist in a crystal in any dimension, including our 1D chain, where they can be visualised with displacements perpendicular to the chain direction.) Quite generally, for each atom in the unit cell, one expects to find three branches of phonons (two transverse, and one longitudinal); always there are three acoustic branches, so a solid that has m atoms in its unit cell will have $3(m-1)$ optical modes. And again, each optical modes will be separated into two transverse branches and one longitudinal branch.[5]

Density of states

Just as for the electron gas problem we need to write down the density of states for phonons. First, we need to count how many modes we have and understand their distribution in momentum space.

In the 1D monatomic chain containing N atoms (assume N very large), there are just N degrees of freedom (for the longitudinal vibration) and therefore N modes. This tells us (and we can see explicitly by looking at boundary conditions for an N-particle chain) that the allowed k-points are discrete, viz

$$k_n = \frac{2\pi}{L} n \;\; ; \;\; n = (-\frac{N}{2}, -\frac{N}{2}+1, ..., \frac{N}{2}] \;\; , \tag{3.32}$$

so that k runs from $-\pi/a$ to π/a, with $a = N/L$, the lattice constant. Notice this is the same spacing of k-states for the electron problem, and the only difference is that because the atoms are discrete, there is a maximum momentum (on the Brillouin zone boundary) allowed by counting degrees of freedom.

By extension, in three dimensions, each branch of the phonon spectrum still contains N states in total, but now $N = L^3/\Omega_{cell}$ with Ω_{cell} the volume of the unit cell, and $L^3 = V$ the volume of the crystal. The volume associated with each allowed k-point is then

$$\Delta k = \frac{(2\pi)^3}{L^3} \tag{3.33}$$

[5]The separation between longitudinal and transverse is only rigorously true along lines of symmetry in **k**-space.

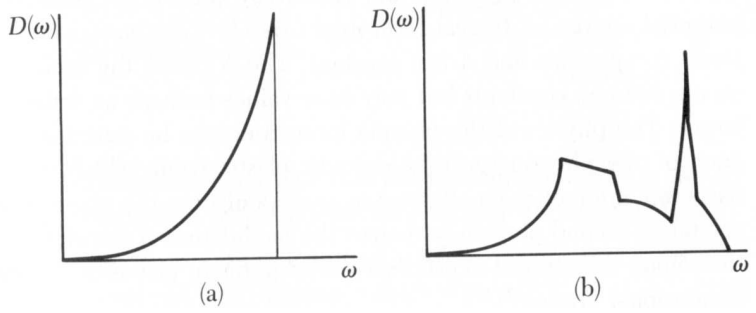


Figure 3.15: Comparison of Debye density of states (a) with that of a real material (b).

There are 3 acoustic branches, and $3(m-1)$ optical branches.

It is convenient to start with a simple description of the optical branch(es), the *Einstein* model, which approximates the branch as having a completely flat dispersion $\omega(\mathbf{k}) = \omega_0$. In that case, the density of states in frequency is simply

$$D_E(\omega) = N\delta(\omega - \omega_0) \quad . \tag{3.34}$$

We have a different result for the acoustic modes, which disperse linearly with momentum as $\omega \to 0$. Using a dispersion $\omega = vk$, and following the earlier argument used for electrons, we obtain the *Debye* model

$$D_D(\omega) = \frac{4\pi k^2}{(2\pi/L)^3}\frac{dk}{d\omega} = \frac{V\omega^2}{2\pi^2 v^3} \quad . \tag{3.35}$$

Of course this result cannot apply once the dispersion curves towards the zone boundary, and there must be an upper limit to the spectrum. In the Debye model, we cut off the spectrum at a frequency ω_D, that is determined such that the total number of states (N) is correctly counted, i.e., by choosing

$$\int_0^{\omega_D} d\omega D_D(\omega) = N \tag{3.36}$$

which yields

$$\omega_D^3 = \frac{6\pi^2 v^3 N}{V} \quad . \tag{3.37}$$

Note that this corresponds to replacing the correct cutoff in momentum space (determined by intersecting Brillouin zone planes) by a sphere of radius

$$k_D = \omega_D / v \quad . \tag{3.38}$$

3.7 Lattice specific heat

Phonons obey Bose-Einstein statistics, but their number is not conserved and so the chemical potential is zero, leading to the Planck distribution

$$n(\omega) = \frac{1}{\exp(\hbar\omega/k_B T) - 1} \quad . \tag{3.39}$$

The internal energy is

$$U = \int d\omega D(\omega) n(\omega) \hbar\omega \tag{3.40}$$

For the Einstein model

$$U_E = \frac{N\hbar\omega_o}{e^{\hbar\omega_o/k_BT} - 1} \tag{3.41}$$

and the heat capacity is

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V = Nk_B \left(\frac{\hbar\omega_o}{k_BT}\right)^2 \frac{e^{\hbar\omega_o/k_BT}}{(e^{\hbar\omega_o/k_BT} - 1)^2} \quad . \tag{3.42}$$

At low temperatures, this grows as $\exp -\hbar\omega_o/k_BT$ and is very small, but it saturates at a value of Nk_B (the Dulong and Petit law) above the characteristic temperature $\theta_E = \hbar\omega_o/k_B$.[6]

At low temperature, the contribution of optical modes is small, and the Debye spectrum is appropriate. This gives

$$U_D = \int_0^{\omega_D} d\omega \frac{V\omega^2}{2\pi^2 v^3} \frac{\hbar\omega}{e^{\hbar\omega/k_BT} - 1} \quad . \tag{3.43}$$

Power counting shows that the internal energy then scales with temperature as T^4 and the specific heat as T^3 at low temperatures. The explicit formula can be obtained as

$$C_V = 9Nk_B \left(\frac{T}{\theta_D}\right)^3 \int_0^{\theta_D/T} dx \frac{x^4 e^x}{(e^x - 1)^2} \quad , \tag{3.44}$$

where the *Debye temperature* is $\theta_D = \hbar\omega/k_B$. We have multiplied by 3 to account for the three acoustic branches.

[6]This is per branch of the spectrum, so it is multiplied by 3 in three dimensions.

Chapter 4

Electrons in a periodic potential

Our modelling of electrons in solids – both in terms of the classical Drude model and the quantum mechanical Sommerfeld model – has so far ignored the presence of the electrostatic potential caused by the positively charged ions. In crystalline lattices, the spatial dependence of this potential has the same symmetry as the lattice, and this greatly simplifies the problem. In particular, the potential is subject to discrete translational symmetry

4.1 Schrödinger equation in a periodic potential

We consider first a formal treatment in terms of a complete set of basis functions, namely the set of all plane wave states which satisfy the periodic boundary conditions. The results from this treatment can be used to obtain Bloch's theorem, which is one of the cornerstones of electronic structure in solids. Next, we will approach Bloch's theorem from a more abstract but also more elegant direction, which uses the translational symmetry of the lattice directly.

We are looking for solutions to $\hat{\mathrm{H}}\,|\psi\rangle = (\frac{\hat{p}^2}{2m} + V)\,|\psi\rangle = E\,|\psi\rangle$, where $V(\mathbf{r})$ is periodic. Because $V(\mathbf{r})$ has the same periodicity as the lattice, it can be Fourier-expanded. We define its Fourier components at reciprocal lattice vectors $\mathbf{G}$ as

$$V_{\mathbf{G}} = \frac{1}{\text{Vol.}} \int d^3\mathbf{r}\, e^{-i\mathbf{G}\cdot\mathbf{r}} V(\mathbf{r}) = \frac{1}{\text{Vol. per cell}} \int_{\text{unit cell}} d^3\mathbf{r}\, e^{-i\mathbf{G}\cdot\mathbf{r}}\, V(\mathbf{r})\ , \tag{4.1}$$

and conversely expand the spatial dependence of the potential as

$$V(\mathbf{r}) = \sum_{\mathbf{G}} V_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}} \quad . \tag{4.2}$$

Since the potential is real, $V^*_{\mathbf{G}} = V_{-\mathbf{G}}$. The Fourier component for $\mathbf{G} = 0$, $V_{\mathbf{G}} = V_0$ is the average of the potential, which we set to zero. If it were not zero, this would simply add a constant to all the energy values obtained in the following.

We build the eigenstate $|\psi\rangle$ from the plane wave states $|\mathbf{k}\rangle$, defined such that $\langle\mathbf{r}|\mathbf{k}\rangle = e^{i\mathbf{k}.\mathbf{r}}$. These form a complete set of basis vectors for 'well-behaved' functions, as is shown in functional analysis (for example, the completeness of this set of functions is the reason why Fourier transforms work).

$$|\psi\rangle = \sum_{\mathbf{k}} c_{\mathbf{k}} |\mathbf{k}\rangle$$

When we apply the Hamiltonian $\hat{\mathrm{H}}$ to this, we find

$$\sum_{\mathbf{k}} E^0_{\mathbf{k}} c_{\mathbf{k}} e^{i\mathbf{kr}} + \left[\sum_{\mathbf{G}} V_{\mathbf{G}} e^{i\mathbf{Gr}}\right]\left[\sum_{\mathbf{k}} c_{\mathbf{k}} e^{i\mathbf{kr}}\right] = E \sum_{\mathbf{k}} c_{\mathbf{k}} e^{i\mathbf{kr}}$$

(where $E^0_{\mathbf{k}} = \frac{\hbar^2 k^2}{2m}$),
which can be rewritten as

$$\sum_{\mathbf{k}} E^0_{\mathbf{k}} c_{\mathbf{k}} |\mathbf{k}\rangle + \sum_{\mathbf{G},\mathbf{k}} V_{\mathbf{G}} c_{\mathbf{k}} |\mathbf{G}+\mathbf{k}\rangle = E \sum_{\mathbf{k}} c_{\mathbf{k}} |\mathbf{k}\rangle$$

We now relabel the $\mathbf{k}$'s in the middle sum, $\mathbf{G}+\mathbf{k} \rightarrow \mathbf{k}$, to obtain:

$$\sum_{\mathbf{k}} E^0_{\mathbf{k}} c_{\mathbf{k}} |\mathbf{k}\rangle + \sum_{\mathbf{G},\mathbf{k}} V_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} |\mathbf{k}\rangle = E \sum_{\mathbf{k}} c_{\mathbf{k}} |\mathbf{k}\rangle$$

From this, we can extract an equation for the coefficients $c_{\mathbf{k}}$ by left multiplying with a single plane wave state. This gives the eigenvalue equation

$$\left(E^0_{\mathbf{k}} - E\right) c_{\mathbf{k}} + \sum_{\mathbf{G}} V_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} = 0 \tag{4.3}$$

This is a key result. Here, $\mathbf{k}$ can be anywhere in reciprocal space. We can go one step further and relate the general wavevector $\mathbf{k}$ to a vector $\mathbf{q}$ which lies in the first Brillouin zone, by shifting through a reciprocal lattice vector $\mathbf{G}'$: $\mathbf{q} = \mathbf{k} + \mathbf{G}'$, where $\mathbf{q}$ lies in the first Brillouin zone. If we now replace the sum over all $\mathbf{G}$ by one over all $\mathbf{G}'' = \mathbf{G} + \mathbf{G}'$, then we find

$$\left(\frac{\hbar^2}{2m}(\mathbf{q}-\mathbf{G}')^2 - E\right) c_{\mathbf{q}-\mathbf{G}'} + \sum_{\mathbf{G}''} V_{\mathbf{G}''-\mathbf{G}'} c_{\mathbf{q}-\mathbf{G}''} = 0 \tag{4.4}$$

Bloch's theorem from considering a plane wave basis

Although Eqn. (4.4) appears to be single equation, it is really an infinite set of simultaneous equations. For a given wavevector in the first Brillouin zone, $\mathbf{q}$, we need to consider all the coefficients $c_{\mathbf{q}-\mathbf{G}'}$ that are associated with plane wave states that can be connected with $|\mathbf{q}\rangle$ via a reciprocal lattice vector, because they contribute to the sum in the second term of (4.4). It is an eigenvector/eigenvalue problem.

We can in principle solve (4.4) to find the set of coefficients $c_{\mathbf{q}-\mathbf{G}}$. This set of coefficients is a distinct sub-set of all $c_{\mathbf{k}}$. It allows us to find a particular eigenfunction of $\hat{\mathrm{H}}$:

$$\psi_{\mathbf{q}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{q}-\mathbf{G}} e^{i(\mathbf{q}-\mathbf{G}).\mathbf{r}}.$$

By taking out a factor $e^{i\mathbf{q}.\mathbf{r}}$, this can be rewritten as

$$\psi_{\mathbf{q}}(\mathbf{r}) = e^{i\mathbf{q}.\mathbf{r}} \sum_{\mathbf{G}} c_{\mathbf{q}-\mathbf{G}} e^{-i\mathbf{G}.\mathbf{r}} = e^{i\mathbf{q}.\mathbf{r}} u_{j,\mathbf{q}}(\mathbf{r})$$

where the function $u_{j,\mathbf{q}}(\mathbf{r})$ is built from periodic function $e^{-i\mathbf{G}.\mathbf{r}}$, and must therefore have the same periodicity as the lattice.

This is Bloch's theorem:

Eigenstates of the one-electron Hamiltonian can be chosen to be a plane wave multiplied by a function with the periodicity of the Bravais lattice.

4.2 Bloch's theorem from discrete translational symmetry

Another way of thinking about Bloch's theorem is to consider what happens to an eigenstate of the Hamiltonian (kinetic energy plus periodic lattice potential), if it is translated in space. An arbitrary translation operation will not necessarily produce an eigenstate, because the new state, generated by this translation, may not match the lattice potential correctly. If, however, the translation operation is matched to the lattice, the resulting state is also an eigenstate of the Hamiltonian. The reason for this lies in the connection between symmetry and quantum mechanics, which is discussed in quantum mechanics courses: if an operator (such as the Hamiltonian) is unchanged under a change of coordinate system (i.e., a symmetry operation such as translation, rotation, etc.), then applying a symmetry operation on the eigenstate of such an operator produces another eigenstate of the operator, *with the same eigenvalue* as the original one.

Either the two eigenstates produced by the symmetry operation are actually the same and differ only by a complex prefactor, or they are different, in which case we are dealing with a set of degenerate eigenstates. In the first case, it is clear that the original eigenstate is also an eigenstate of the symmetry operation. In the second case, it can be shown that we can always *choose* from the subspace of degenerate eigenstates a set of eigenstates that are also eigenstates of the symmetry operation.

Lattices are symmetric under discrete translation of the coordinate system by lattice vectors. Accordingly, the eigenstates of the Hamiltonian can be chosen to be eigenstates of the discrete lattice translation operation. This is the underlying origin of Bloch's theorem, which we will now explore in more detail.

4.2.1 Symmetry in quantum mechanics – applied to the lattice

Consider a symmetry operator $\hat{\mathrm{T}}$, e.g., the translation $\left\langle \mathbf{r}|\hat{\mathrm{T}}\psi \right\rangle = \psi(\mathbf{r}+\mathbf{a})$. If the Hamiltonian $\hat{\mathrm{H}}$ commutes with the symmetry operator $\hat{\mathrm{T}}$, this implies that $\hat{\mathrm{T}}$ maps one eigenstate of the Hamiltonian $\hat{\mathrm{H}}$ onto another eigenstate of $\hat{\mathrm{H}}$ with the same energy:

$$\hat{\mathrm{H}}\left|\hat{\mathrm{T}}\psi\right\rangle = \hat{\mathrm{T}}\left|\hat{\mathrm{H}}\psi\right\rangle = E\left|\hat{\mathrm{T}}\psi\right\rangle$$

Now, we are faced with two possibilities:

1. If the Hamiltonian has no degenerate eigenstates, there is only one eigenstate with this energy, and $\hat{\mathrm{T}}$ maps it onto itself (with a complex pre-factor). So the non-degenerate eigenstates of $\hat{\mathrm{H}}$ are always also eigenstates of $\hat{\mathrm{T}}$.

2. If the Hamiltonian has degenerate eigenstates, then these are not necessarily also eigenstates of $\hat{\mathrm{T}}$. The set of degenerate eigenstates with a particular eigenvalue forms a subspace. $\hat{\mathrm{H}}$ maps a state within this subspace onto itself, but $\hat{\mathrm{T}}$ maps states within the subspace onto other states within the subspace. Because $\hat{\mathrm{T}}$ is a unitary operator (symmetry operation leaves 'length' of ψ unchanged), we can diagonalise the associated matrix $T_{mn} = \left\langle \psi_m|\hat{\mathrm{T}}|\psi_n \right\rangle$. ThIS means that we can find basis states within the degenerate subspace that are eigenstates of $\hat{\mathrm{T}}$. In other words, we can *choose* a set of states that are simultaneously eigenstates of $\hat{\mathrm{H}}$ and of $\hat{\mathrm{T}}$.

We find, then, that a complete set of eigenstates $\hat{\mathrm{H}}$ can always be found which are at the same time eigenstates of the symmetry operator $\hat{\mathrm{T}}$.

For two commuting operators $\hat{\mathrm{H}}$, $\hat{\mathrm{T}}$, we can always choose simultaneous eigenstates of both $\hat{\mathrm{H}}$ and $\hat{\mathrm{T}}$.

This motivates us to use the eigenvalue of $\hat{\mathrm{T}}$ to give an eigenstate of $\hat{\mathrm{H}}$ a meaningful label. In the lattice, $\hat{\mathrm{H}}$ commutes with translation operator $\hat{\mathrm{T}}_{\mathbf{a}}$, where $\mathbf{a}$ is a Bravais lattice vector. To find the possible eigenvalues of translation operator $\hat{\mathrm{T}}_{\mathbf{a}}$, let it operate on plane wave states $|\mathbf{k}\rangle$, which after all are also eigenstates of $\hat{\mathrm{T}}$ (not necessarily of $\hat{\mathrm{H}}$, though!) and form a complete basis set. This gives us the set of all possible eigenvalues of $\hat{\mathrm{T}}$:

$$\hat{\mathrm{T}}_{\mathbf{a}}\,|\mathbf{k}\rangle = e^{i\mathbf{k}.\mathbf{a}}\,|\mathbf{k}\rangle$$

If we now choose $\hat{\mathrm{H}}$ eigenstates $|\psi\rangle$ which are also eigenstates of $\hat{\mathrm{T}}$: $\hat{\mathrm{H}}\,|\psi\rangle = E\,|\psi\rangle \implies \hat{\mathrm{T}}_{\mathbf{a}}\,|\psi\rangle = c_{\mathbf{a}}\,|\psi\rangle$, where $c_{\mathbf{a}}$ is an eigenvalue of $\hat{\mathrm{T}}_{\mathbf{a}}$, then we know that the $\hat{\mathrm{T}}_{\mathbf{a}}$-eigenvalue $c_{\mathbf{a}}$ must be of the form $e^{i\mathbf{k}.\mathbf{a}}$, because these form a complete set of eigenvalues for $\hat{\mathrm{T}}$.

This realisation leads directly to a form of Bloch's theorem:

$$\hat{\mathrm{T}}_{\mathbf{a}}\,|\psi\rangle = e^{i\mathbf{k}\mathbf{a}}\,|\psi\rangle \tag{4.5}$$

From now on, we use the $\mathbf{k}$ from the exponent in (4.5) to label the energy eigenstate: $\psi \to \psi_{\mathbf{k}}$.

4.2.2 Bloch's theorem

We now join everything up and apply it specifically to electrons subject to a periodic potential. The energy eigenstates of electrons in a lattice:

$$\hat{\mathrm{H}}\psi(\mathbf{r}) = \left[\frac{\hat{\mathbf{p}}^2}{2m} + V(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}) \ ,$$

where $V(\mathbf{r}+\mathbf{R}) = V(\mathbf{r})$ for $\forall\ \mathbf{R}$ in a Bravais lattice.

The $\psi_{\mathbf{k}}^{(n)}$ can be chosen such that

$$\psi_{\mathbf{k}}^{(n)}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r}) \quad \text{, where } \ u_{\mathbf{k}}^{(n)}(\mathbf{r}+\mathbf{R}) = u_{\mathbf{k}}^{(n)}(\mathbf{r}) \tag{4.6}$$

Or, equivalently:

$$\psi_{\mathbf{k}}^{(n)}(\mathbf{r}+\mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{R}}\psi_{\mathbf{k}}^{(n)}(\mathbf{r}) \tag{4.7}$$

Here, n is called the *band index*. It is necessary, because there may be several distinct eigenstates of $\hat{\mathrm{H}}$ with the same symmetry label $\mathbf{k}$. The band index distinguishes between these. Note that whereas the potential is periodic, the wavefunction $\psi(\mathbf{r})$ is not. It is formed by multiplying a plane wave state with a periodic function, which has the same translational symmetry as the lattice.

The two forms of Bloch's theorem (4.6 and 4.7) can be shown to be equivalent – each implies the other. For instance, applying $\hat{\mathrm{T}}_{\mathbf{R}}$ to the product $e^{i\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r})$ in (4.6) will produce the phase shift $e^{i\mathbf{k}\cdot\mathbf{R}}$ required by (4.7). Conversely, substituting a product $e^{i\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r})$ (where $u(\mathbf{r})$ can be any function, not necessarily periodic) into (4.7) will produce (4.6) and demonstrate that $u(\mathbf{r})$ indeed has to be periodic.

The Bloch states (plane wave × periodic function) are similar to eigenstates of free electrons (just plane waves), but the choice of periodic function gives additional freedom in labelling states. Note, for instance that for any reciprocal lattice vector $\mathbf{g}$, $e^{i\mathbf{g}.\mathbf{r}}$ is periodic with same periodicity as the Bravais lattice, which follows from the definition of the reciprocal lattice vectors $\mathbf{g}$. This can be used to relabel a Bloch state $\mathbf{k}$ with a new wavevector $\mathbf{k}-\mathbf{g}$ by introducing a different periodic function $u^{(n)} = e^{i\mathbf{g}\mathbf{r}}u^{(m)}$:

$$\psi_{\mathbf{k}}^{(m)}(\mathbf{r}) = e^{i\mathbf{k}.\mathbf{r}}u_{\mathbf{k}}^{(m)}(\mathbf{r}) = e^{i\mathbf{k}.\mathbf{r}}e^{-i\mathbf{g}.\mathbf{r}}\left[e^{i\mathbf{g}.\mathbf{r}}u_{\mathbf{k}}^{(m)}(\mathbf{r})\right] = e^{i(\mathbf{k}-\mathbf{g}).\mathbf{r}}u_{\mathbf{k}-\mathbf{g}}^{(n)}(\mathbf{r}) = \psi_{\mathbf{k}-\mathbf{g}}^{(n)}(\mathbf{r})$$

In this case, exactly the same function $\psi(\mathbf{r})$ can be labelled by wavevector $\mathbf{k}$, if the periodic function in the Bloch state is $u_{\mathbf{k}}^{(m)}(\mathbf{r})$, or by wavevector $\mathbf{k}-\mathbf{g}$, if the corresponding periodic function is $u_{\mathbf{k}-\mathbf{g}}^{(n)}(\mathbf{r}) = e^{i\mathbf{g}.\mathbf{r}}u_{\mathbf{k}}^{(m)}(\mathbf{r})$. This implies that for every state labelled with a $\mathbf{k}-$ vector outside the first Brillouin zone *we can find an identical state which can be labelled with a vector* $\mathbf{q} = \mathbf{k}-\mathbf{g}$ *inside the first Brillouin zone.* From this, we conclude that:

Any quantity that depends on the wavefunction,
in particular energy, is periodic in wavevector space.

Plane wave expansion of a Bloch state

Knowing that Bloch's theorem follows from general symmetry considerations, we can now rederive (4.4) more quickly, by constructing a state which conforms with Bloch's theorem from the outset:

$$|\psi_{\mathbf{k}}\rangle = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} |\mathbf{k}-\mathbf{G}\rangle; \quad , \tag{4.8}$$

where the sum runs over all reciprocal lattice vectors $\mathbf{G}$. Where does this come from? Bloch's theorem states that $\psi_{\mathbf{k}}(\mathbf{r})$ is a product of a plane wave $e^{i\mathbf{kr}}$ and a function $u_{\mathbf{k}}(\mathbf{r})$ with the periodicity of the lattice. We can Fourier-expand the periodic function as a sum over all reciprocal lattice vectors, $u_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{-i\mathbf{Gr}}$. This gives for $\psi_{\mathbf{k}}(\mathbf{r}) = \langle \mathbf{r}|\psi_{\mathbf{k}}\rangle = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{i(\mathbf{k}-\mathbf{G})\mathbf{r}} = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} \langle \mathbf{r}|\mathbf{k}-\mathbf{G}\rangle$. In this form, the electron wavefunction appears as a superposition of harmonics, whose wavevectors are related by reciprocal lattice vectors $\mathbf{G}$.

Writing the Hamiltonian as $\hat{\mathrm{H}} = \hat{\mathrm{H}}_0 + V$, where $\hat{\mathrm{H}}_0$ gives the kinetic energy and V is the periodic potential of the lattice, we are looking for the eigenvalues $E_{\mathbf{k}}$ in

$$\hat{\mathrm{H}}|\psi_{\mathbf{k}}\rangle = E_{\mathbf{k}}|\psi_{\mathbf{k}}\rangle \quad . \tag{4.9}$$

Left multiply with a plane wave state $\langle \mathbf{k}|$:

$$\langle \mathbf{k}|\hat{\mathrm{H}}|\psi_{\mathbf{k}}\rangle = E_{\mathbf{k}} c_{\mathbf{k}} = \langle \mathbf{k}|\hat{\mathrm{H}}_0|\mathbf{k}\rangle c_{\mathbf{k}} + \sum_{\mathbf{G}} \langle \mathbf{k}|V|\mathbf{k}-\mathbf{G}\rangle c_{\mathbf{k}-\mathbf{G}} \tag{4.10}$$

We can identify $\langle \mathbf{k}|V|\mathbf{k}-\mathbf{G}\rangle$ as the Fourier component $V_{\mathbf{G}}$ of the periodic potential, defined in (4.1). We immediately obtain the key equation:

$$\left(E_{\mathbf{k}}^{(0)} - E_{\mathbf{k}}\right) c_{\mathbf{k}} + \sum_{\mathbf{G}} V_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} = 0 \quad , \tag{4.11}$$

where the kinetic energy $E_{\mathbf{k}}^{(0)} = \frac{\hbar^2}{2m} k^2$. This is the same as (4.3), which we derived from a more general plane wave expansion for $|\psi\rangle$.

It is often convenient to rewrite $\mathbf{q} = \mathbf{k} + \mathbf{G}'$, where $\mathbf{G}'$ is a reciprocal lattice vector chosen so that $\mathbf{q}$ lies in the first Brillouin zone, and to write $\mathbf{G}'' = \mathbf{G} + \mathbf{G}'$ in the second summation. This gives back (4.4):

$$\left[\left(\frac{\hbar^2}{2m}(\mathbf{q}-\mathbf{G}')^2 - E\right) c_{\mathbf{q}-\mathbf{G}'} + \sum_{\mathbf{G}''} U_{\mathbf{G}''-\mathbf{G}'} c_{\mathbf{q}-\mathbf{G}''}\right] = 0 \tag{4.12}$$

4.3 Nearly free electron theory

Although we have, with Eqn. (4.3), reduced the problem of finding the eigenstates of the electronic Hamiltonian to that of solving an eigenvector/eigenvalue problem, this still looks rather intractable: we are stuck with an infinite set of basis functions and therefore with having to diagonalise, in principle, an infinitely-dimensional matrix. Recall that the single-electron state was obtained from the plane wave expansion $|\psi_{\mathbf{k}}\rangle = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} |\mathbf{k}-\mathbf{G}\rangle$, in which we have to fix all the coefficients $c_{-\mathbf{G}}$. However, it should be possible to find approximate eigenstates by reducing the size of the basis set. [1]

[1] There are lengthy descriptions of this approach in all the textbooks. A nice treatment similar to the one given her can be found in the book by Singleton.

4.3.1 Connection to second-order perturbation theory

If the strength of the periodic potential is weak compared to the magnitude of the kinetic energy term, then we would expect that the eigenstates are constructed from a dominant plane wave state $|\mathbf{k}\rangle$, plus an admixture from a small number of 'lattice harmonics' $|\mathbf{k}-\mathbf{G}\rangle$. We can use second-order perturbation theory to find the degree of admixture of the harmonics. Recall that the energy-level shift due to admixing a particular state $|\mathbf{k}-\mathbf{G}\rangle$ in second order perturbation theory is given as:

$$\Delta E_{\mathbf{k}} = \frac{|V_{\mathbf{G}}|^2}{E^{(0)}_{\mathbf{k}} - E^{(0)}_{\mathbf{k}-\mathbf{G}}} \quad .$$

(Remember $V_{\mathbf{G}}$ is Fourier component of the lattice potential at reciprocal lattice wavevector $\mathbf{G}$ so that we can write $V = \sum_{\mathbf{G}} V_{\mathbf{G}} e^{i\mathbf{G}.\mathbf{r}}$.

This energy shift, and the associated admixture of lattice harmonics $|\mathbf{k}-\mathbf{G}\rangle$ into the dominant state $\mathbf{k}$, is most pronounced when $E^{(0)}_{\mathbf{k}} \simeq E^{(0)}_{\mathbf{k}-\mathbf{G}}$, i.e., when there are nearly degenerate states. The approximate Bloch state $\psi_{\mathbf{k}}(\mathbf{r})$ is therefore built from the dominant state, plus an admixture from those states nearly degenerate with it, which form a reduced set of $\mathbf{k}$-states compared to those we started out with. We assume that all the other coefficients $c_{\mathbf{k}-\mathbf{G}}$ can be neglected.

To work out the perturbed energy levels, we apply the general equation obtained earlier:

$$\left(E^{(0)}_{\mathbf{k}} - E^{0}_{\mathbf{k}-\mathbf{G}}\right) c_{\mathbf{k}} + \sum_{\mathbf{G}} V_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} = 0$$

but restrict the choice of $\mathbf{G}$-vectors to those that link together nearly degenerate states: $E^{(0)}_{\mathbf{k}} \simeq E^{(0)}_{\mathbf{k}-\mathbf{G}}$. This is an example of *degenerate perturbation theory*, an approach we have applied before when calculating the energy levels of molecular orbitals (covalent bonding).

4.3.2 Example: one-dimensional chain

As an example, let us consider the problem in one dimension (Fig. 4.1). Start with state $|\mathbf{k}\rangle$. Potential $V_{\mathbf{G}}$ admixes $|\mathbf{k}-\mathbf{G}\rangle$, which is close in energy. It also admixes other states, but their energies are more widely separated from that of $|\mathbf{k}\rangle$, so we concentrate on $|\mathbf{k}-\mathbf{G}\rangle$ for now.

Now, apply $\hat{\mathrm{H}}$ to $|\psi\rangle$:

$$\begin{aligned} |\psi\rangle &= c_{\mathbf{k}} |\mathbf{k}\rangle + c_{\mathbf{k}-\mathbf{G}} |\mathbf{k}-\mathbf{G}\rangle \\ \hat{\mathrm{H}} |\psi\rangle &= E |\psi\rangle = c_{\mathbf{k}} \frac{p^2}{2m} |\mathbf{k}\rangle + c_{\mathbf{k}} V |\mathbf{k}\rangle + c_{\mathbf{k}-\mathbf{G}} \frac{p^2}{2m} |\mathbf{k}-\mathbf{G}\rangle + c_{\mathbf{k}-\mathbf{G}} V |\mathbf{k}-\mathbf{G}\rangle \end{aligned}$$

As usual in all these kind of calculations, we now pick out the two coefficients $c_{\mathbf{k}}$ and $c_{\mathbf{k}-\mathbf{G}}$ one at a time, by left-multiplying with the basis states present in the above equation, (i) $\langle\mathbf{k}|$, and (ii) $\langle\mathbf{k}-\mathbf{G}|$:

$$\begin{aligned} c_{\mathbf{k}} E &= c_{\mathbf{k}} E^{(0)}_{\mathbf{k}} + c_{\mathbf{k}} V_0 + c_{\mathbf{k}-\mathbf{G}} V_{\mathbf{G}} \\ c_{\mathbf{k}-\mathbf{G}} E &= c_{\mathbf{k}} V_{-\mathbf{G}} + c_{\mathbf{k}-\mathbf{G}} V_0 + c_{\mathbf{k}-\mathbf{G}} E^{(0)}_{\mathbf{k}-\mathbf{G}} \end{aligned} \tag{4.13}$$

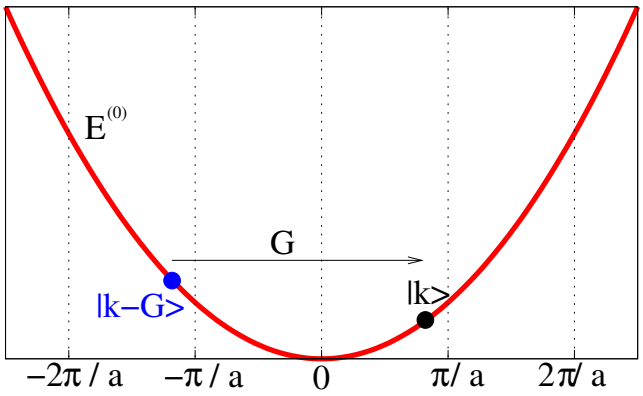


Figure 4.1: The nearly free electron approach illustrated for a one-dimensional solid. The first Brillouin zone boundary is at $\pm\pi/a$. States $|k\rangle$ close to π/a are nearly degenerate with states $|k - 2\pi/a\rangle$, and the matrix element linking those states, the lowest Fourier coefficient of the lattice potential, $V_{2\pi/a}$, is non-zero. Hence, these two states mix (hybridise) to form the Bloch states near the Brillouin zone boundary.

(Note that $E^{(0)}_{\mathbf{k}} = \hbar k^2/2m$, V_0 is set to zero, and $V_{-\mathbf{G}} = V_{\mathbf{G}}{}^*$, because the potential $V(\mathbf{r})$ is always real).

You can see that this is a special case of the general set of equations $\left(E^{(0)}_{\mathbf{k}} - E\right) c_{\mathbf{k}} + \sum_{\mathbf{G}} V_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} = 0$ (4.3), in which we consider just one value of $\mathbf{G}$. If we wanted to consider the effect of more $\mathbf{G}$-vectors, we would have to solve more simultaneous equations. We could have found Eqns. (4.13) directly from (4.3) by setting all the coefficients for $G \neq 0, 2\pi/a$ to zero.

We can solve the set of equations above by finding the roots of a 2×2 determinant, which will give us two perturbed energies E: one will be reduced compared to the unperturbed energy, the other will be increased. We could interpret these energies as the perturbed energies of the $\mathbf{k}$ and $\mathbf{k} - \mathbf{G}$-states, respectively, and call one energy $E_{\mathbf{k}}$ and the other $E_{\mathbf{k}-\mathbf{G}}$. With this convention, we would be following the *extended zone scheme*, in which bands continue beyond the first Brillouin zone. Alternatively, we could interpret these energies as two solutions at the same wavevector $\mathbf{k}$ and call one energy $E^{(1)}_{\mathbf{k}}$ and the other $E^{(2)}_{\mathbf{k}}$. This gives us two energy bands within the first Brillouin zone and is called the reduced zone scheme.

At the Brillouin zone boundary ($k = \pi/a$), the energies of the two solutions are simply $E = E^{(0)}_{\pi/a} \pm |V_{\mathbf{G}}|$. Here, both $|\mathbf{k}\rangle$ and $|\mathbf{k} - 2\pi/a\rangle = |-\pi/a\rangle$ contribute equally to the Bloch states at π/a, which are formed either from the sum or from the difference of the two unperturbed states. Both combinations give rise to standing waves, but with different probability distribution: in one case, the nodes of the probability distribution will be centred on the atomic cores, in the other case the bellies of the probability distribution are centred on the atomic cores.

A complementary and quite instructive approach starts from the alternative form (4.4), which has the periodicity in wavevector space built in:

$$\left[\left(\frac{\hbar^2}{2m}(\mathbf{q} - \mathbf{G}')^2 - E\right) c_{\mathbf{q}-\mathbf{G}'} + \sum_{\mathbf{G}''} V_{\mathbf{G}''-\mathbf{G}'} c_{\mathbf{q}-\mathbf{G}''}\right] = 0 \quad ,$$

where $\mathbf{q}$ is always in the first Brillouin zone, and is obtained from $\mathbf{k}$, which might fall outside the first Brillouin zone, by subtracting $\mathbf{G}'$.

If the lattice potential V were zero, then the sum in the second term of this equation would vanish. We would be left with a set of independent simultaneous equations for E, which have solutions $E = \hbar^2/(2m)q^2$ if $c_{\mathbf{q}} \neq 0$ and all the other $c_{\mathbf{q}-\mathbf{G}'} = 0$, and generally $E = \hbar^2/(2m)(\mathbf{q} - \mathbf{G}')^2$ for a particular $c_{\mathbf{q}-\mathbf{G}'} \neq 0$, when all the other coefficients apart from that one are zero. We obtain a set of parabolic bands (Fig. 4.2) for the unperturbed solutions, which will then hybridise, where the bands cross, when the lattice potential is non-zero.

To make the calculation more specific, we work out the actual dispersion for a one-dimensional chain. We use a simplified atomic potential which just contains the leading Fourier components, i.e.,

$$V(x) = 2V_{2\pi/a} \cos \frac{2\pi x}{a} \tag{4.14}$$

If $V_{2\pi/a}$ is small, we should be able to treat it perturbatively, remembering to take care of degeneracies. If $V_{2\pi/a} = 0$, we get the free electron eigenvalues

$$E_0^{(m)}(k) = \frac{\hbar^2}{2m}(k - 2\pi m/a)^2 \,, \qquad m = ..., -2, -1, 0, 1, 2, ... \tag{4.15}$$

which are repeated, offset parabolas.

Now suppose $V_{2\pi/a}$ is turned on, but is very small. It will be important only for those momenta at which two free electron states are nearly degenerate, for example, m=0,1 are degenerate when $k = \pi/a$. Near that point, we can simplify the band structure to the 2x2 matrix

$$\begin{pmatrix} \frac{\hbar^2}{2m}{}^2 k^2 - E & V_{2\pi/a} \\ V^*_{2\pi/a} & \frac{\hbar^2}{2m}(k - \frac{2\pi}{a})^2 - E \end{pmatrix} \begin{pmatrix} c_k \\ c_{k-2\pi/a} \end{pmatrix} \tag{4.16}$$

The solution of the determinantal leads to a quadratic equation:

$$E^{\pm}(\mathbf{k}) = \frac{\hbar^2}{2m}\frac{1}{2}(k^2 + (k - 2\pi/a)^2) \pm \frac{1}{2}\sqrt{(\frac{\hbar^2}{2m}k^2 - \frac{\hbar^2}{2m}(k - 2\pi/a)^2)^2 + 4V^2_{2\pi/a}} \tag{4.17}$$

Exactly at $k = \pi/a$, the energy levels are

$$E^{\pm}(\pi/a) = E^0_{\pi/a} \pm |V_{2\pi/a}|, \tag{4.18}$$

and if we choose the potential to be attractive $V_{2\pi/a} < 0$, the wavefunctions are (aside from normalisation)

$$\begin{aligned} \psi^-(\pi/a) &= \cos(\pi x/a) \,, \\ \psi^+(\pi/a) &= \sin(\pi x/a) \,. \end{aligned} \tag{4.19}$$

4.3.3 Example calculations for 3D metals

Fig. 4.3 illustrates the results of a three dimensional nearly free electron calculation. You can see that

- Because of Bloch's theorem, for every $\left|\psi^n_{\mathbf{k}+\mathbf{G}}\right\rangle$ there is an identical state $|\psi^m_{\mathbf{k}}\rangle$. Therefore, $E_{\mathbf{k}}$ has the same periodicity as the reciprocal lattice.
- At the Brillouin zone boundary, $|\mathbf{k}| = |\mathbf{k} - \mathbf{G}| \implies E^{(0)}_{\mathbf{k}} = E^{(0)}_{\mathbf{k}-\mathbf{G}}$. Because unperturbed bands cross at the Brillouin zone boundary, this is where hybridisation (admixture of states) and band distortion is strongest.

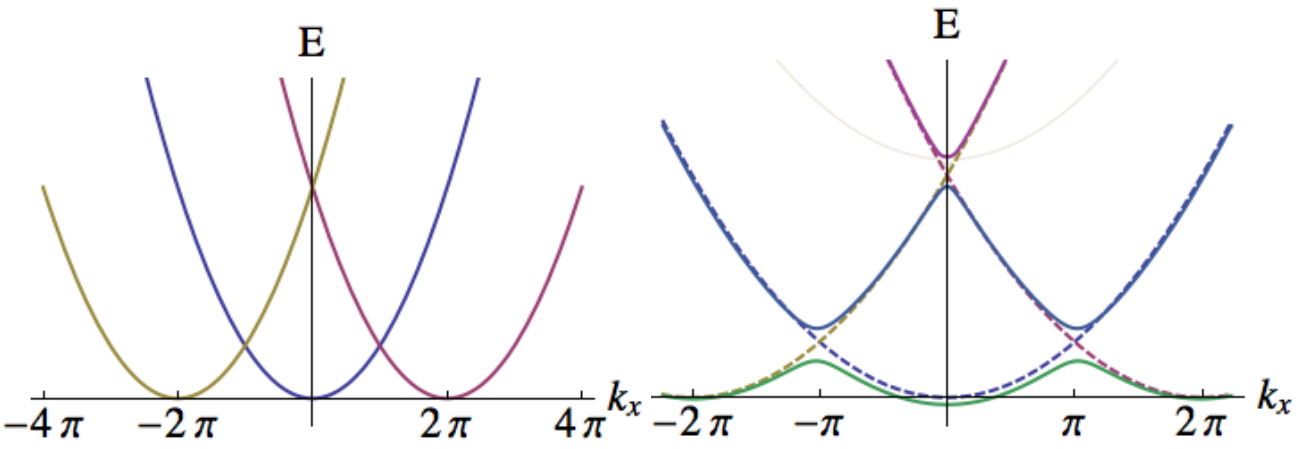


Figure 4.2: Energy dispersion along k_x from a nearly free electron calculation for a three dimensional solid.

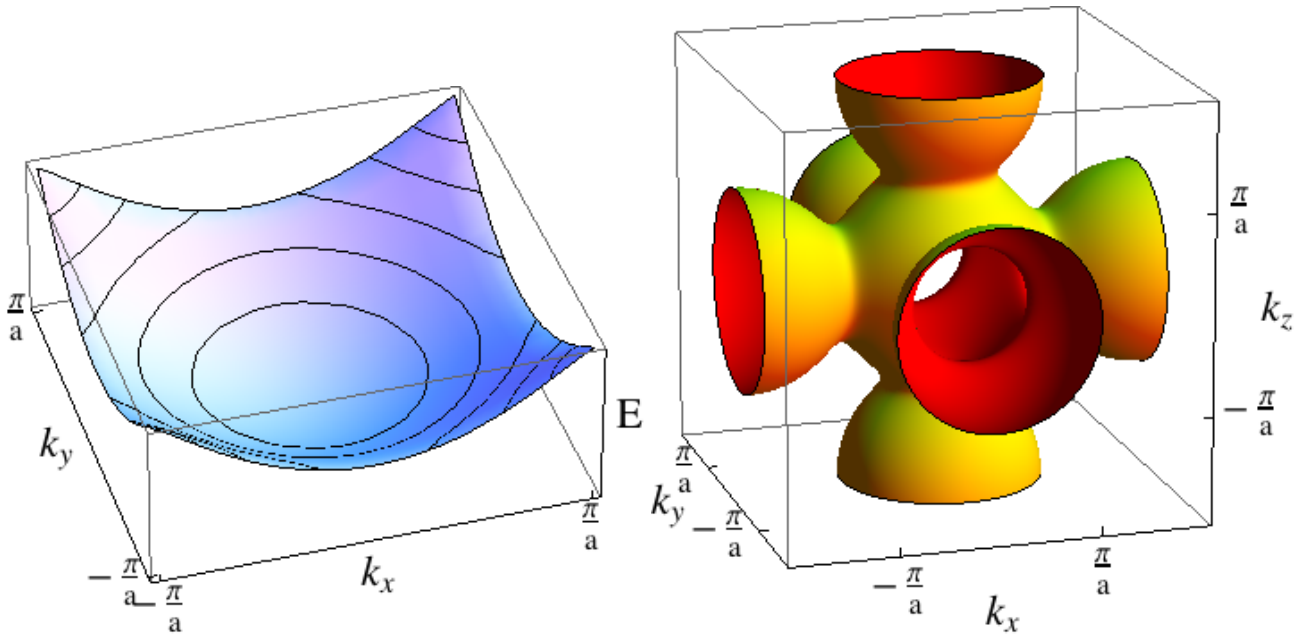


Figure 4.3: Example of a nearly free electron calculation for a three dimensional solid. Left: energy vs. k_x and k_y for $k_z = 0$, within the first Brillouin zone. The energy contours (black lines) are circular near the bottom of the band, but distort as the contours approach the Brillouin zone boundary. Near the top of the band, they enclose the corners of the Brillouin zone. Right: an equal energy surface in reciprocal space. The shape of the surface illustrates that energy is periodic in reciprocal space and that equal energy contours intersect the Brillouin zone boundary at right angles.

4.4 Tight binding: Linear combination of atomic orbitals

Perhaps the most natural view of a solid is to think about it as a collection of interacting atoms, and to build up the wavefunctions in the solid from the wavefunctions of the individual atoms. This is the *linear combination of atomic orbital* (LCAO) or *tight-binding* method.

4.4.1 Diatomic molecule

Remember our modelling of a covalently bonded diatomic molecule, where we worked with a highly restricted basis on one orbital per atom. For identical atoms, the full Hamiltonian consists of

$$H = T + V_a + V_b \tag{4.20}$$

with T the kinetic energy and V_a, V_b the (identical potentials) on the two atoms. The basis set consists of two states $|a>$ and $|b>$ that satisfy

$$T + V_a|a> = E_0|a> \tag{4.21}$$
$$T + V_b|b> = E_0|b> \tag{4.22}$$

so that E_0 is the eigenenergy of the atomic state, and we look for solutions

$$|\psi>= \alpha|a> +\beta|b> \tag{4.23}$$

We solve this in the usual way: Project $H|\psi>= E|\psi>$ onto $\langle a|$ and $\langle b|$ to get the simultaneous equations

$$\begin{pmatrix} \tilde{E}_0 - E & t \\ t^* & \tilde{E}_0 - E \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0 \tag{4.24}$$

neglecting the overlap elements $< a|b >$.

Here

$$\tilde{E}_0 = H_{aa} = \langle a\,|T + V_a + V_b|\,a\rangle = E_a + \langle a\,|V_b|\,a\rangle \tag{4.25}$$

is a shift of the atomic energy by the *crystal field* of the other atom(s).
The more interesting term is the *hopping* matrix element that couples the atomic states together:[2]

$$t = H_{ab} = \langle a\,|T + V_a + V_b|\,b\rangle \tag{4.26}$$

For $t < 0$, the new eigenstates are

$$|\psi\rangle = \frac{1}{\sqrt{2}}\left[|a\rangle \mp |b\rangle\right] \quad E = \tilde{E}_0 \pm |t| \tag{4.27}$$

For the lower energy (bonding) state, the electron density has a maximum between the atoms. For the higher energy (antibonding) state, the electron density has a node between the atoms.

[2]Note the sign of t depends on the symmetry of the orbitals: for s-states, with an attractive potential $V_i < 0$, then t is negative; but for p_x states t is positive for atoms aligned along x.

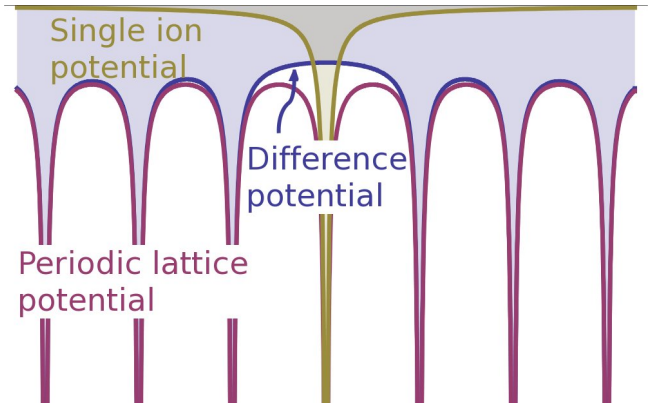


Figure 4.4: Treating the periodic potential as a perturbation on top of the atomic potential caused by a single atomic core

4.4.2 Linear chain

Now let us generalise this approach to a chain of atoms, subject to periodic boundary conditions. Bloch's theorem seriously constrains the possibilities for forming hybridised states from the atomic orbitals. If we want to make up a wave-function using only one-orbital per unit cell we now know that it *must* be of the form

$$|\psi\rangle = \sum_n e^{i\mathbf{k}.\mathbf{R}_n} |n\rangle \quad , \tag{4.28}$$

where $\mathbf{R}_n$ = position of atom n, $|n\rangle$ is an atomic orbital centred on atom n, so that the associated wavefunction is $\langle \mathbf{r}| \, n\rangle = \phi(\mathbf{r} - \mathbf{R}_n)$. The atomic orbitals are eigenfunctions of the single-atom Hamiltonian $\hat{\mathrm{H}}_n^{(0)} = \frac{\hat{\mathrm{p}}^2}{2m} + V_{\mathrm{atom}}(\mathbf{r} - \mathbf{R}_n)$: $\hat{\mathrm{H}}_n^{(0)} |n\rangle = E_0 |n\rangle$, and we are looking for approximate eigenstates to the full Hamiltonian $\hat{\mathrm{H}} = \frac{\hat{\mathrm{p}}^2}{2m} + V_{\mathrm{lattice}}\mathbf{r}$.

To check that $|\psi\rangle = \sum_n e^{i\mathbf{k}\mathbf{R}_n} |n\rangle$ obeys Bloch's theorem, we can apply a translation $\hat{\mathrm{T}}_\mathbf{a}$ to $|\psi\rangle$: $\hat{\mathrm{T}}_\mathbf{a}\psi(\mathbf{r}) = \psi(\mathbf{r} + \mathbf{a})$. This maps an atomic orbital centred on atom n ($\phi(\mathbf{r} - \mathbf{R}_n)$) onto an orbital centred on atom m ($\phi(\mathbf{r} - \mathbf{R}_m)$). $\hat{\mathrm{T}}_\mathbf{a}\phi(\mathbf{r} - \mathbf{R}_n) = \phi(\mathbf{r} - (\mathbf{R}_n - \mathbf{a})) = \phi(\mathbf{r} - \mathbf{R}_m) \implies \mathbf{R}_n - \mathbf{a} = \mathbf{R}_m$

$$\hat{\mathrm{T}}_\mathbf{a} |\psi\rangle = \hat{\mathrm{T}}_\mathbf{a} \sum_n e^{i\mathbf{k}.\mathbf{R}_n} |n\rangle = \sum_m e^{i\mathbf{k}.\mathbf{R}_n} |m\rangle = e^{i\mathbf{k}.\mathbf{a}} \sum_m e^{i\mathbf{k}.\mathbf{R}_m} |m\rangle = e^{i\mathbf{k}.\mathbf{a}} |\psi\rangle$$

Now, we can apply the Hamiltonian to the state $|\psi\rangle$ constructed from the atomic orbitals: $\hat{\mathrm{H}} |\psi\rangle = E |\psi\rangle$ Assuming that the orbitals are orthogonal, we left multiply with one of the basis states. It saves algebra to use the basis state $\langle 0|$, which is centred at the origin. This gives the dispersion $E(\mathbf{k})$:

$$\langle 0| \hat{\mathrm{H}} |\psi\rangle = E = \sum_n e^{i\mathbf{k}.\mathbf{R}_n} \langle 0| \hat{\mathrm{H}} |n\rangle$$

If we neglect matrix elements between atomic orbitals which are not next to one another, i.e., we consider only nearest-neighbour 'hopping', with 'hopping elements' (or 'transfer integrals')

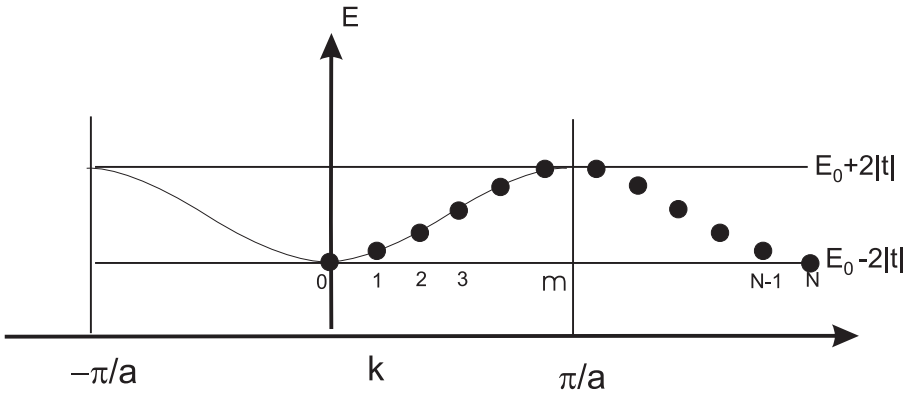


Figure 4.5: Eigenvalues of the 1D chain (4.29) are confined to a band in energy centred on the (shifted) atomic energy level $\tilde{E}_0$. If N is very large, the energies form a continuous band and are periodic in m. Then we replace the index m by the continuous *crystal momentum* $k = 2\pi m/Na$, with a the lattice constant. So we could label the states more symmetrically by keeping a range $-N/2 + 1 < m < N/2$ (or $-\pi/a < k < \pi/a$); this is the first *Brillouin zone*.

$t^* = t = \langle n| \hat{\mathrm{H}} |n+1\rangle$, and define $\tilde{E}_0 = \langle n| \hat{\mathrm{H}} |n\rangle$, then we obtain a particularly simple expression for the energy dispersion:

$$E_k = \tilde{E}_0 + 2t\cos(ka) \tag{4.29}$$

Note that there is some ambiguity in defining t. Some textbooks use a different definition of t, $t := -\langle n| \hat{\mathrm{H}} |n+1\rangle$, in order to obtain $E_k = \tilde{E}_0 - 2t\cos(ka)$. The hopping element $t = \langle n| \hat{\mathrm{H}} |n+1\rangle$ is < 0 between s-orbitals, giving the familiar free-electron like dispersion near $k = 0$, whereas it is < 0 between p_x orbitals along the x-direction, for example.

Because, as usual we apply periodic boundary conditions, the allowed values of k are discrete, but very close together, spaced by $\Delta k = 2\pi/L$, where $L = Na$. The range of k must cover $k_{max} - k_{min} = 2\pi/a$ to give N states. It is convenient to choose the range $-\pi/a < k < \pi/a$, the *first Brillouin zone*.

4.4.3 Generalised LCAO (tight binding) method

We have seen that it is easy to obtain an energy dispersion by combining a single set of atomic orbitals, because Bloch's theorem dictates the precise form of this combination. It is straightforward to extend the calculation presented above to higher dimensions. The Bloch states are written exactly as before, in (4.28), $|\psi\rangle = \sum_n e^{i\mathbf{k}.\mathbf{R}_n} |n\rangle$, but now the sum extends over all the atoms in a three-dimensional solid. Correspondingly, the dispersion given by (4.4.2), $E(\mathbf{k}) = \sum_n e^{i\mathbf{k}.\mathbf{R}_n} \langle 0| \hat{\mathrm{H}} |n\rangle$ now contains contributions from hopping processes in all three directions. For instance, in a simple cubic lattice with nearest-neighbour hopping (matrix elements $\langle 0| \hat{\mathrm{H}} |n\rangle = 0$ if atom n is not a nearest neighbour to atom 0, and $= t$ for if atom n is a nearest neighbour of atom 0), we obtain $E(\mathbf{k}) = E_0 + 2t(\cos(k_x a) + \cos(k_y a) + \cos(k_z a))$. This dispersion is plotted in Fig. 4.6.

There is a major problem with producing Bloch states with only a single orbital per site, this only gives a single energy band. To understand many materials, and in particular semi-

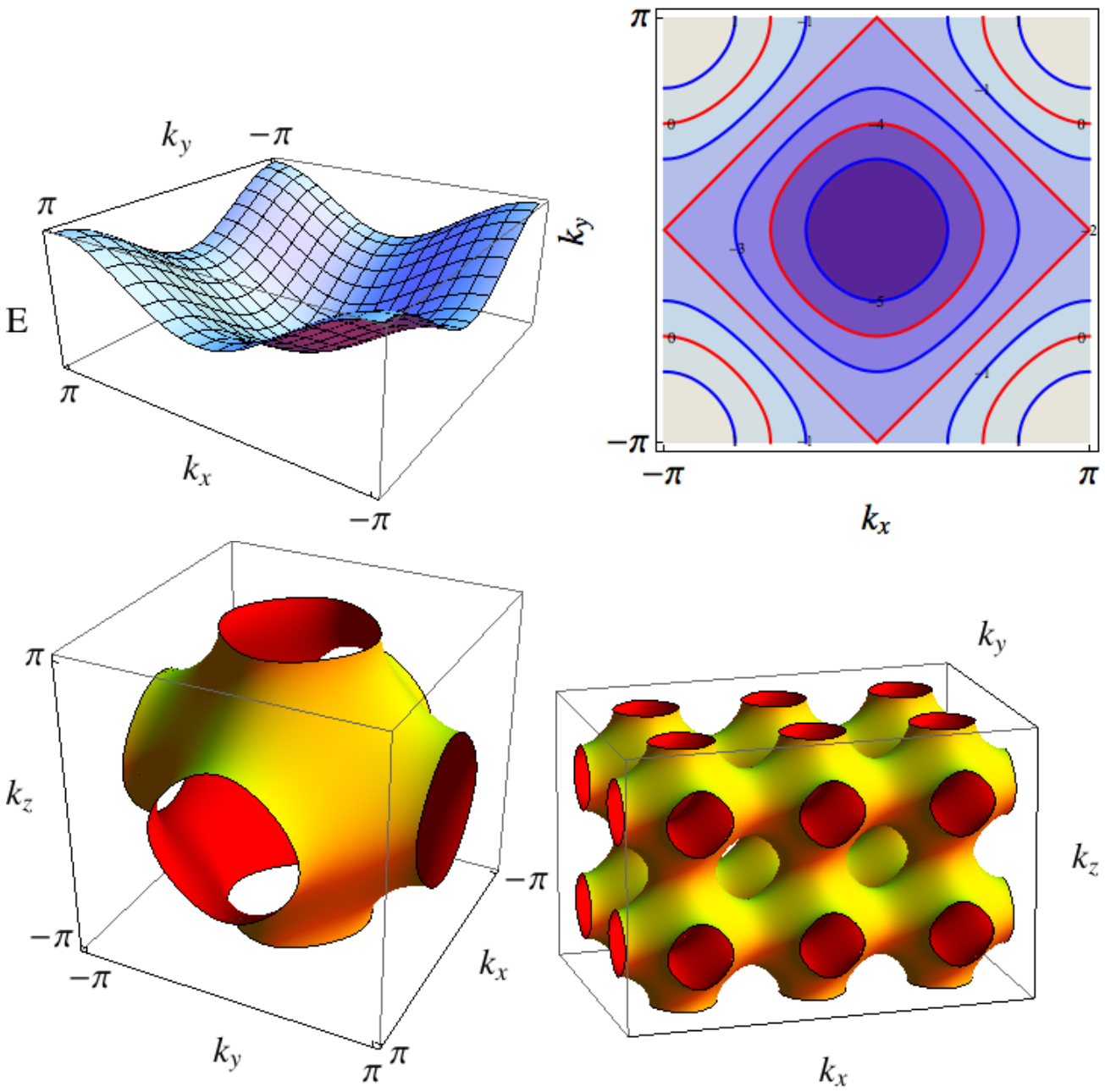


Figure 4.6: Electronic energy dispersion calculated within the tight binding approach for nearest-neighbour hopping, with a single orbital per atom in a simple cubic lattice ($E = E_0 + 2t(\cos(k_x a) + \cos(k_y a) + \cos(k_z a))$).

conductors, we must extend the method so that it can produce several bands. This requires that we build the Bloch states from several orbitals per site.

To see how this works, let us first combine two orbitals per site, namely $|a_n\rangle$ and $|b_n\rangle$, to form a hybridised local orbital, and then we combine these hybridised orbitals to form a Bloch state:

$$|\psi\rangle = \sum_n e^{i\mathbf{k}\mathbf{R}_n} \left(\alpha_{\mathbf{k}} |a_n\rangle + \beta_{\mathbf{k}} |b_n\rangle\right) \tag{4.30}$$

Note that the hybridisation coefficients $\alpha_{\mathbf{k}}$ and $\beta_{\mathbf{k}}$ can in general depend on the wavevector $\mathbf{k}$.

Applying our usual prescription, insert $|\psi\rangle$ into the Schrödinger eqn. $\hat{\mathrm{H}}|\psi\rangle = E|\psi\rangle$, left multiply by basis states $\langle a_0|$ and $\langle b_0|$ and turn the resulting set of simultaneous equations into an eigenvector problem, which gives:

$$\begin{pmatrix} E_a(\mathbf{k}) - E & V_{\mathbf{k}} \\ V_{\mathbf{k}}^* & E_b(\mathbf{k}) - E \end{pmatrix} \begin{pmatrix} \alpha_{\mathbf{k}} \\ \beta_{\mathbf{k}} \end{pmatrix} = 0 \tag{4.31}$$

where $E_a = \sum_n e^{i\mathbf{k}.\mathbf{R}_n} \langle a_0| \hat{\mathrm{H}} |a_n\rangle$ is the dispersion of a band, which would be formed exclusively from the atomic orbitals $|a_n\rangle$, and $E_b = \sum_n e^{i\mathbf{k}.\mathbf{R}_n} \langle b_0| \hat{\mathrm{H}} |b_n\rangle$ is the dispersion of a band, which would be formed exclusively from the atomic orbitals atomic orbitals $|b_n\rangle$. The off-diagonal matrix element $V_{\mathbf{k}}$ is given by

$$V_{\mathbf{k}} = \sum_{\mathbf{R}_n} e^{i\mathbf{k}.\mathbf{R}_n} \langle a_0| \hat{\mathrm{H}} |b_n\rangle$$

It can be shown with a bit of relabelling that

$$V_{\mathbf{k}}^* = \sum_{\mathbf{R}_n} e^{i\mathbf{k}.\mathbf{R}_n} \langle b_0| \hat{\mathrm{H}} |a_n\rangle$$

The eigenvalue problem (4.31) gives rise to two possible energies at each wavevector – we have formed two bands from the two orbitals. The effect of the off-diagonal elements in (4.31) is to hybridise the two bands which would form from each set of atomic orbitals exclusively where they become nearly degenerate.

One could have approached the calculation from another angle by first forming two Bloch states from combining either the local orbitals $|a_n\rangle$ ($|\psi_a\rangle = \sum_n e^{i\mathbf{k}.\mathbf{R}_n} |a_n\rangle$) or the local orbitals $|b_n\rangle$ ($|\psi_b\rangle = \sum_n e^{i\mathbf{k}.\mathbf{R}_n} |b_n\rangle$). Then, we combine the Bloch states:

$$|\psi\rangle = \alpha_{\mathbf{k}} |\psi_a\rangle + \beta_{\mathbf{k}} |\psi_b\rangle$$

If we substitute in our expressions for $|\psi_a\rangle$ and $|\psi_a\rangle$, then we obtain exactly the same form for $|\psi\rangle$ as before, in (4.30). So, whether you consider the bandstructure as arising from hopping between hybridised, molecular orbitals, or attribute it to hybridisation between bands which arise from atomic orbitals makes no difference to the formalism.

It is now clear how to extend this method to multiple orbitals per unit cell. We simply generalise the summation in (4.30):

$$|\psi\rangle = \sum_{n,\nu} e^{i\mathbf{k}.\mathbf{R}_n} c_{\mathbf{k}}^{(\nu)} \left|n^{(\nu)}\right\rangle \quad , \tag{4.32}$$

where the index ν labels the different local orbitals $\left|n^{(\nu)}\right\rangle$, which exist in the n^{th} unit cell, and $c_{\mathbf{k}}^{(\nu)}$ is the associated coefficient, which determines to what level this local orbital is mixed into the final state. This choice of Bloch state will give rise to an eigenvector/eigenvalue problem, in which the number of energy eigenvalues at a particular wavevector $\mathbf{k}$, and thereby the number of bands, is equal to the number of local orbitals chosen per unit cell.

4.4.4 Tight binding versus the Nearly Free Electron approximation

Fundamentally, both the tight binding and the nearly free electron method do the same thing: a Bloch state is constructed from a set of basis functions, and the associated coefficients are determined by solving the eigenvector/eigenvalue equation which arises, when the Hamiltonian is expressed in that basis. If the basis is complete, then the Bloch state is an exact eigenstate of the Hamiltonian. However, a complete basis would give an infinite-dimensional eigenvector problem, so the expansion of the Bloch state is done within a reduced basis set.

In the case of the nearly free electron approximation, the complete basis set required to form $|\psi_{\mathbf{k}}\rangle$ is the set of all plane wave states $|\mathbf{k}-\mathbf{G}\rangle$ (eigenstates of the kinetic energy operator), but we can restrict this by concentrating on those states for which the matrix elements $V_{\mathbf{G}}$ are large and which are nearly degenerate with $|\mathbf{k}\rangle$. If the periodic potential is comparatively smoothly varying and weak (compared to the kinetic energy), then we can disregard Fourier components $V_{\mathbf{G}}$ with high wavevector $\mathbf{G}$, and the set of basis functions is small. In this case, the nearly free electron approximation is computationally efficient. If the periodic potential varies strongly compared to the kinetic energy, then the plane wave expansion has to include a much larger set of states. However, this is not a big problem for modern computers which can handle hundreds of thousands of plane waves. The vast majority of modern electronic structure calculations are performed using large plane wave basis sets.

The tight binding approximation (or linear combination of atomic orbitals) is, however, still useful because it can be used to obtain reasonable answers with rather little computation, and it can also provide important insights into the chemistry which are hidden when large plane wave basis sets are used. The eigenstates of the local Hamiltonians $\hat{\mathrm{H}}_n^{(0)}$, i.e., the set of all atomic orbitals associated with the different unit cells n, form a complete basis in which the Bloch states can be expanded, although in practical calculations only a finite number of orbitals can be included. Section 4.4.3 showed that the number of bands generated in this method is equal to the number of atomic orbitals included per unit cell. If the hopping matrix elements are small compared to the separation between the bands, then the bands generated from different sets of atomic orbitals do not cross. In this case, the strength of the potential – which fixes the energies of the atomic orbitals and also their spacing – is larger than the kinetic energy – which is related to the bandwidth. To get a good description of a particular electronic band only a small set of atomic orbitals needs to be included.

In short, in the nearly-free-electron scheme the kinetic energy appears on the diagonal of energy matrix in the eigenvector equation, and the potential appears in the off-diagonal terms, mixing basis states together. This is a very efficient approach if the periodic potential is a weak perturbation and plane wave schemes are very convenient and work very nicely on large computers. In the tight binding scheme the potential energy appears on the diagonal and the hopping elements, which are the equivalent of the kinetic energy, form the off-diagonal terms. This approach is useful for obtaining chemical insight and for systems in which the potential

is strong compared to the kinetic energy.

4.5 Pseudopotentials

The NFE method and the tight-binding method are not accurate methods of electronic structure determination; nevertheless both of them exhibit the basic principles. They are commonly used to write down simple models for bands, with their parameters fitted to more sophisticated calculations, or to experiment. It turns out that band gaps in semiconductors are usually fairly small, and the true dispersion can be modelled by scattering from a few Fourier components of the lattice potential. The reason is that the relevant scattering potential for valence band electrons can be MUCH smaller than the full atomic potential Ze^2/r of an electron interacting with a nucleus of charge Z. The effective potential for scattering of the valence electrons by the atomic cores is a weak *pseudopotential*.

When we consider the band structure of a typical solid, we are concerned only with the valence electrons, and not with those tightly bound in the core, which remain nearly atomic. If we solve the full Schrödinger equation with the real Coulomb potential, we expect to calculate not just the valence electronic states, but also the atomic like core states. A pseudopotential reproduces the valence states as the *lowest* eigenstates of the problem and removed the core states from the problem.

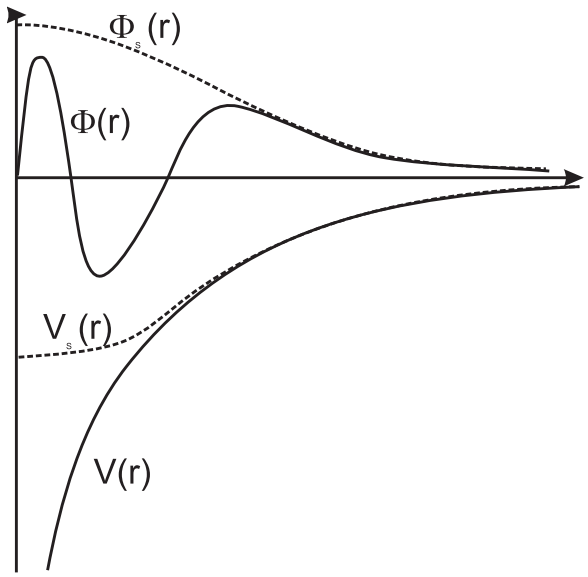


Figure 4.7: Pseudopotential: The true potential $V(r)$ has a wavefunction for the valence electrons that oscillates rapidly near the core. The pseudopotential $V_s(r)$ has a wavefunction $\Phi_s(r)$ that is smooth near the core, but approximates the true wavefunction far from the core region.

A weak pseudopotential acting on a smooth pseudo-wavefunction gives nearly the same energy eigenvalues for the valence electrons as the full atomic potential does acting on real wavefunctions. Away from the atomic cores, the pseudopotential matches the true potential, and the pseudo-wavefunction approximates the true one.

A formal derivation of how this works can be given using the method of orthogonalised plane waves. The atomic states are well described by the Bloch functions $f_{n\mathbf{k}}$ of the LCAO or tight-binding scheme (4.32). Higher states, which extend well beyond the atoms will not necessarily be of this kind, but they

must be *orthogonal* to the core levels. This suggests that we should use as a basis [3]

$$|\chi_{\mathbf{k}}>=|\mathbf{k}>-\sum_n \beta_n |f_{n\mathbf{k}}>, \tag{4.33}$$

where $|\mathbf{k}>$ is a plane wave, and the coefficients $\beta_n(\mathbf{k})$ are chosen to make the states χ orthogonal to the core states $|f_{n\mathbf{k}}>$. The states in (4.33) are *orthogonalised plane waves* (OPW); away from the core, they are plane wave like, but in the vicinity of the core they oscillate rapidly so as to be orthogonal to the core levels.

We can now use the OPW's as basis states for the diagonalisation in the same way that we used plane waves in the NFE, viz

$$|\psi_k>=\sum_{\mathbf{G}} \alpha_{\mathbf{k}-\mathbf{G}} |\chi_{\mathbf{k}-\mathbf{G}}> \ . \tag{4.34}$$

This turns out to converge very rapidly, with very few coefficients, and only a few reciprocal lattice vectors are included in the sum. The following discussion explains why.

Suppose we have solved our problem exactly and determined the coefficients α. Now consider the sum of plane waves familiar from the plane-wave expansion, but using the same coefficients, i.e.,

$$|\phi_k>=\sum_{\mathbf{G}} \alpha_{\mathbf{k}-\mathbf{G}} |\mathbf{k}-\mathbf{G}>, \tag{4.35}$$

and then[4] it is easily shown that

$$|\psi>=|\phi>-\sum_n <f_n|\phi> |f_n> \ . \tag{4.36}$$

Then substitute into the Schrödinger equation, $H|\psi>=E|\psi>$, giving

$$H|\phi>+\sum_n (E-E_n) <f_n|\phi> |f_n>=E|\phi> \tag{4.37}$$

We may look upon this as a new Schrödinger equation with a *pseudopotential* defined by the operator

$$V_s|\phi>=U|\phi>+\sum_n (E-E_n) <f_n|\phi> \ |f_n> \tag{4.38}$$

which may be written as a non-local operator in space

$$(V_s-U)\phi(r)=\int V_R(\mathbf{r},\mathbf{r}')\phi(\mathbf{r}')\,d\mathbf{r}' \ ,$$

where

$$V_R(\mathbf{r},\mathbf{r}')=\sum_n (E-E_n) f_n(\mathbf{r}) f_n^*(\mathbf{r}') \ . \tag{4.39}$$

The pseudopotential acts on the smooth *pseudo-wavefunctions* $|\phi>$, whereas the bare Hamiltonian acts on the highly oscillating wavefunctions $|\psi>$.

One can see in (4.38) that there is strong cancellation between the two terms. The bare potential is large and attractive, especially near the atomic core at $r \approx 0$; the second term V_R is positive, and this cancellation reduces the total value of V_s especially near the core. Away from the core, the pseudopotential approaches the bare potential.

[3] We use Dirac's *bra* and *ket* notation, where $|\mathbf{k}>$ represents the plane wave state $\exp(i\mathbf{k}\cdot\mathbf{r})$, and $<\phi_1|T|\phi_2>$ represents the matrix element $\int d\mathbf{r}\, \phi_1^*(\mathbf{r}) T(\mathbf{r}) \phi_2(\mathbf{r})$ of the operator T.

[4] Saving more notation by dropping the index k.

Chapter 5

Bandstructure of real materials

5.1 Bands and Brillouin zones

In the last chapter, we noticed that we get band gaps within nearly free electron theory by interference of degenerate forward- and backward going plane waves, which then mix to make standing waves.

Brillouin zones.

What is the condition for obtaining a gap in a three-dimensional band structure? A gap will arise from the splitting of a degeneracy due to scattering from some Fourier component of the lattice potential, i.e., we require

$$E_0(\mathbf{k}) = E_0(\mathbf{k} - \mathbf{G}) \tag{5.1}$$

which means (for a given $\mathbf{G}$) that we must find the $\mathbf{k}$ such that $|\mathbf{k}|^2 = |\mathbf{k} - \mathbf{G}|^2$. Equivalently, this is

$$\mathbf{k} \cdot \frac{\mathbf{G}}{2} = \left|\frac{\mathbf{G}}{2}\right|^2 \tag{5.2}$$

which is satisfied by any vector lying in a plane perpendicular to, and bisecting $\mathbf{G}$. This is, by definition, the boundary of a Brillouin zone; it is also the Bragg scattering condition, not at all coincidentally.[1]

Electronic bands.

We found that the energy eigenstates formed discrete bands $E_n(\mathbf{k})$, which are *continuous* functions of the momentum $\mathbf{k}$ and are additionally labelled by a *band index* n. The bandstructure is periodic in the reciprocal lattice $E_n(\mathbf{k}+\mathbf{G}) = E_n(\mathbf{k})$ for any reciprocal lattice vector $\mathbf{G}$. It is sometimes useful to plot the bands in *repeated zones*, but remember that these states are just being relabelled and are not physically different.

[1] Notice that the Bragg condition applies to both the incoming and outgoing waves in the original discussion in Chapter 4, just with a relabelling of $\mathbf{G} \to -\mathbf{G}$.

Bloch's theorem again.

The eigenstates are of the form given by Bloch's theorem

$$\psi_{n\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \tag{5.3}$$

where $u(\mathbf{r})$ is periodic on the lattice. Notice that if we make the substitution $\mathbf{k} \to \mathbf{k} + \mathbf{G}$, (5.3) continues to hold. This tells us that $\mathbf{k}$ can always be chosen inside the first Brillouin zone for convenience, although it is occasionally useful to plot the bands in an extended or repeated zone scheme as in Fig. 4.2.

Crystal momentum.

The quantity $\hbar\mathbf{k}$ is the crystal momentum, and enters conservation laws for scattering processes. For example, if an electron absorbs the momentum of a phonon of wavevector $\mathbf{q}$, the final state will have a Bloch wavevector $\mathbf{k}' = \mathbf{k} + \mathbf{q} + \mathbf{G}$, where $\mathbf{G}$ is whatever reciprocal lattice vector necessary to keep $\mathbf{k}'$ inside the Brillouin zone. Physical momentum can always be transferred to the lattice in arbitrary units of $\hbar\mathbf{G}$. Notice that depending on the energy conservation, processes can thus lead to transitions between bands.

Counting states.

We saw that the spacing between $\mathbf{k}$-points in 1D is $2\pi/L$, where L is the linear dimension of the crystal. This generalises to 3 dimensions: the volume associated with each $\mathbf{k}$ is

$$(\Delta\mathbf{k})^3 = \frac{(2\pi)^3}{V} \tag{5.4}$$

with V the volume of the crystal. Within each *primitive unit cell* or *Brillouin zone* of the reciprocal lattice the number of $\mathbf{k}$ states allowed by the periodic boundary conditions is equal to the number of unit cells in the crystal. In practice N is so big that the bands are continuous functions of $\mathbf{k}$ and we only need to remember density of states to count. Since electrons are fermions, each $\mathbf{k}$-point can now be occupied by two electrons (double degeneracy for spin). So if we have a system which contains one electron per unit cell (e.g., a lattice of hydrogen atoms), half the states will be filled in the first Brillouin zone. From this, we obtain the even number rule:

Even number rule

Allowing for spin, two electrons per real-space unit cell fill a Brillouin zone's worth of $\mathbf{k}$ states.

Periodic boundary conditions and volume per k-point

A formal proof of the number of allowed $\mathbf{k}$-points uses Bloch's theorem, and follows from the imposition of periodic boundary conditions:

$$\psi(\mathbf{r} + N_i\mathbf{a}_i) = \psi(\mathbf{r}) \tag{5.5}$$

where the N_i are integers, and the number of primitive unit cells in the crystal is $N = N_1 N_2 N_3$, and the $\mathbf{a}_i$ are the primitive lattice vectors. Applying Bloch's theorem, we have immediately that

$$e^{iN_i \mathbf{k}\cdot\mathbf{a}_i} = 1, \tag{5.6}$$

so that the general form for the allowed Bloch wavevectors is

$$\mathbf{k} = \sum_i^3 \frac{m_i}{N_i}\mathbf{b}_i, \qquad \text{for } m_i \text{ integral.} \tag{5.7}$$

with $\mathbf{b}_i$ primitive reciprocal lattice vectors. Thus the volume of allowed $\mathbf{k}$-space per allowed $\mathbf{k}$-point is just

$$(\Delta k)^3 = \left|\frac{\mathbf{b}_1}{N_1}\cdot\frac{\mathbf{b}_2}{N_2}\wedge\frac{\mathbf{b}_3}{N_3}\right| = \left|\frac{1}{N}\mathbf{b}_1\cdot\mathbf{b}_2\wedge\mathbf{b}_3\right|. \tag{5.8}$$

Since $|\mathbf{b}_1\cdot\mathbf{b}_2\wedge\mathbf{b}_3| = (2\pi)^3 N/V$ is the volume of the unit cell of the reciprocal lattice (V is the volume of the crystal), (5.8) shows that the number of allowed wavevectors in the primitive unit cell is equal to the number of lattice sites in the crystal. We may thus rewrite Eq. (5.8) as

$$(\Delta k)^3 = \frac{(2\pi)^3}{V} \tag{5.9}$$

.

Metals and insulators in band theory

The last point is critical to the distinction that band theory makes between a metal and an insulator. A (non-magnetic) system with an even number of electrons per unit cell *may* be an insulator. Otherwise, the Fermi energy must lie within a band and the material will be predicted to be a metal. Metallicity may also occur even if the two-electron rule holds, if different bands overlap in energy so that the counting is satisfied by two or more partially filled bands.

Band theory starts to get into trouble when the Coulomb repulsion between electrons is larger than the bandwidth, as is found in *Mott insulators*. Band theory can still be useful in this case, one just has to use a spin-polarized band theory. However, band theory falls flat on its face in describing the metal/insulator transition in such systems.

Notation

The bandstructure $E_n(\mathbf{k})$ defines a function in three-dimensions which is difficult to visualise. Conventionally, what is plotted are cuts through this function along particular directions in k-space. Also, a shorthand is used for directions in k-space and points on the zone boundary, which you will often see in band structures.

- $\Gamma = (0,0,0)$ is the zone centre.
- X is the point on the zone boundary in the (100) direction; Y in the (010) direction; Z in the (001) direction. *Except* if these directions are equivalent by symmetry (e.g., cubic) they are all called "X".
- L is the zone boundary point in the (111) direction.

- K in the (110) direction.
- You will also often see particular bands labelled either along lines or at points by greek or latin capital letters with a subscript. These notations label the group representation of the state (symmetry) and we won't discuss them further here.

Density of states

We have dealt earlier with the density of states of a free electron band in 2. The maxima E_{max} and minima E_{min} of all bands must have a locally quadratic dispersion with respect to momenta measured from the minima or maxima. Hence the density of states (in 3D) near the minima will be the same

$$g(E \gtrsim E_{min}) = \frac{V}{\pi^2}\frac{m^*}{\hbar^2}\left(\frac{2m^*(E-E_{min})}{\hbar^2}\right)^{\frac{1}{2}} \quad . \tag{5.10}$$

as before, with now however the replacement of the bare mass by an effective mass $m^* = (m_x^* m_y^* m_z^*)^{1/3}$ averaging the curvature of the bands in the three directions [2]. A similar form must apply near the band maxima, but with now $g(E) \propto (E_{max} - E)^{\frac{1}{2}}$. Notice that the flatter the band, the larger the effective mass, and the larger the density of states[3].

Since every band is a surface it will have saddle points (in two dimensions or greater) which are points where the bands are flat but the curvature is of opposite signs in different directions. Examples of the generic behaviour of the density of states in one, two and three dimensions are shown in Fig. 5.1. The saddle points give rise to cusps in the density of states in 3D, and a logarithmic singularity in 2D.

For any form of $E(k)$, the density of states is

$$g(E) = \sum_n g_n(E) = \sum_n \int \frac{d\mathbf{k}}{4\pi^3}\delta(E - E_n(\mathbf{k})) \, , \tag{5.11}$$

Because of the δ-function in (5.11), the momentum integral is actually over a surface in k-space S_n which depends on the energy E; $S_n(E_F)$ is the Fermi surface. We can separate the integral in $\mathbf{k}$ into a two-dimensional surface integral along a contour of constant energy, and an integral perpendicular to this surface $dk_\perp$ (see Fig. 5.2). Thus

$$\begin{aligned} g_n(E) &= \int_{S_n(E)} \frac{dS}{4\pi^3} \int dk_\perp(\mathbf{k})\, \delta(E - E_n(\mathbf{k})) \\ &= \int_{S_n(E)} \frac{dS}{4\pi^3} \frac{1}{|\nabla_\perp E_n(\mathbf{k})|}, \end{aligned} \tag{5.12}$$

where $\nabla_\perp E_n(\mathbf{k})$ is the derivative of the energy in the normal direction.[4]

Notice the appearance of the gradient term in the denominator of (5.12), which must vanish at the edges of the band, and also at saddle points, which exist generically in two and three dimensional bands.

[2] Since the energy $E(\mathbf{k})$ is a quadratic form about the minimum, the effective masses are defined by $\frac{\hbar^2}{m_\alpha^*} = \frac{\partial^2 E(\mathbf{k})}{\partial k_\alpha^2}\Big|_{\mathbf{k}_{min}}$ along the principal axes α of the ellipsoid of energy.

[3] The functional forms are different in one and two dimensions.

[4] We are making use of the standard relation $\delta(f(x) - f(x_0)) = \delta(x - x_0)/|f'(x_0)|$

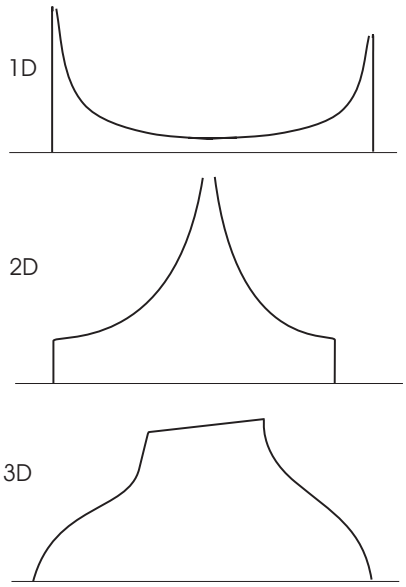


Figure 5.1: Density of states in one (top curve), two (middle curve) and three (lower curve) dimensions

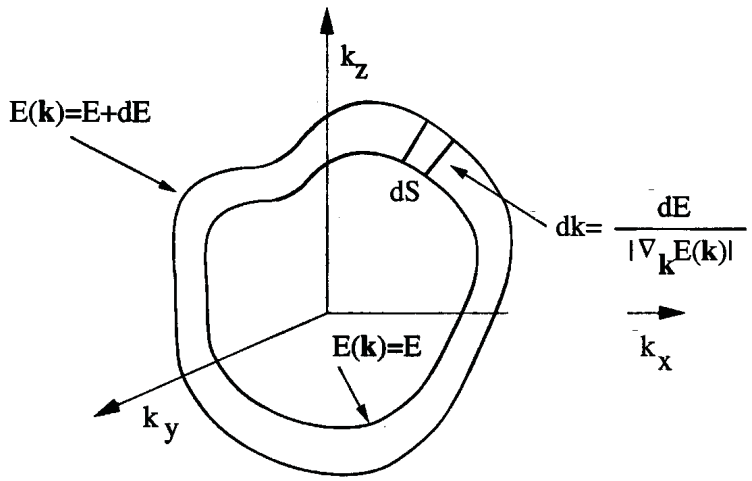


Figure 5.2: Surface of constant energy

Maxima, minima, and saddle points are all generically described by dispersion (measured relative to the stationary point) of

$$E(\mathbf{k}) = E_0 \pm \frac{\hbar^2}{2m_x}k_x^2 \pm \frac{\hbar^2}{2m_y}k_y^2 \pm \frac{\hbar^2}{2m_z}k_z^2 \tag{5.13}$$

If all the signs in (5.13) are positive, this is a band minimum; if all negative, this is a band maximum; when the signs are mixed there is a saddle point. In the vicinity of each of these critical points, also called van Hove singularities, the density of states (or its derivative) is singular. In two dimensions, a saddle point gives rise to a logarithmically singular density of states, whereas in three dimensions there is a discontinuity in the derivative.

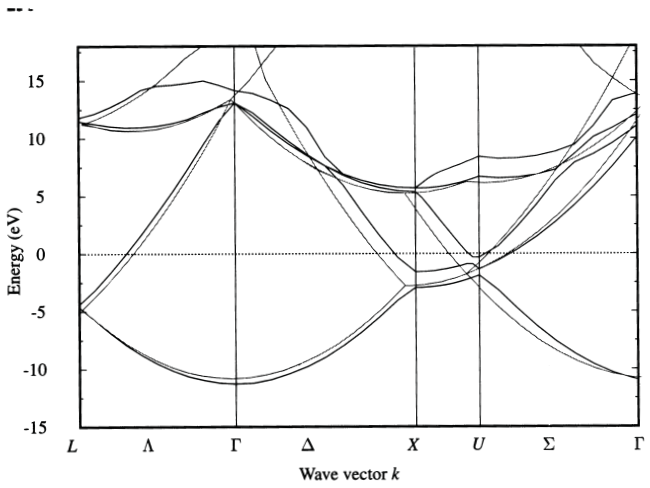


Figure 5.3: Band structure of Al (solid line) compared to the free electron parabolas (dotted line). Calculations from Stumpf and Scheffler, cited by Marder.

5.2 Examples of band structures

Metals

If there is an odd number of electrons per primitive unit cell, the chemical potential must lie within a band, and there will be no energy gap (here we are ignoring the possibility of magnetism). Because there are low-lying electronic excitations, the system is a metal. The *Fermi surface* is the surface in momentum space that separates the filled from the empty states. In a simple metal such as Na ($3s^1$ with 1 valence electron) or Al ($3s^2p^1$ with 3 valence electrons) this is close to a free-electron sphere. In other cases (e.g., Cu, $4s3d^{10}$) the sphere extends in some directions to meet the Brillouin zone boundary surface. There can be situations where several bands are cut by the Fermi energy, and the topology of Fermi surfaces is sometimes complicated.

Semimetals

Even if there is the right number of electrons to fill bands and make a semiconductor, the bands may still overlap. Consequently, the chemical potential will intersect more than one band, making a pocket of electrons in one band and removing a pocket of electrons from the band below (which as we shall see later, are sometimes called holes). This accounts for the metallicity of Ca and Mg (which have two electrons per unit cell), and also As, Sb and Bi. The latter, despite being group V elements, have crystal structures that contain 2 atoms per unit cell and therefore 10 valence electrons. We have previously alluded to graphite, which is a special kind of semimetal. We noted that a graphene sheet has conduction and valence bands that touch at special points on the Brillouin zone boundary. Over all except these points, the band structure has a gap - thus graphene is more correctly described as a *zero-gap semiconductor*.

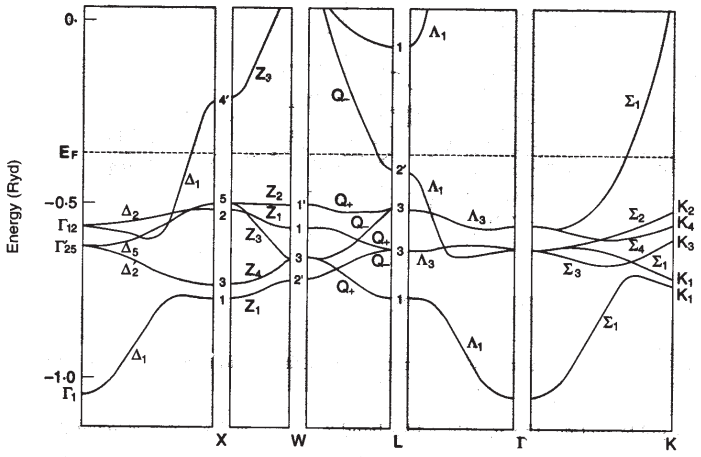


Figure 5.4: Band structure of Cu metal [from G.A. Burdick, *Phys. Rev.***129**,138 (1963)], cited by Grosso and Parravicini.

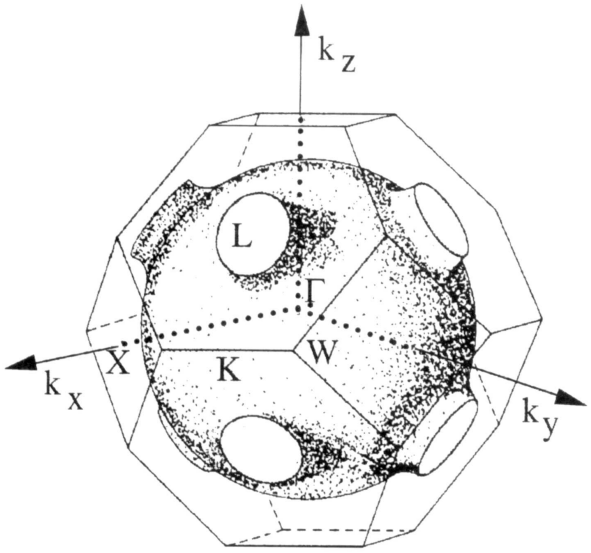


Figure 5.5: Fermi surface of Cu

Semiconductors and insulators

If there is an even number of electrons per unit cell, then it is possible (if the bands do not overlap) for all of the occupied states to lie in a set of filled bands, with an energy gap to the empty states. In this case the system will be a semiconductor or insulator. Such is the case for the group IV elements C, Si and Ge, as well as important III-V compounds such as $GaAs$ and $AlAs$. These elements and compounds in fact have 2 atoms per unit cell (diamond or zincblende structure) and have a total of 8 valence electrons per unit cell — 4 filled bands.

The band structures of Si, Ge, and GaAs are shown in Fig. 5.6 and Fig. 5.7. The maximum of the valence bands of all the materials is at Γ. Si and Ge are both *indirect* gap materials,

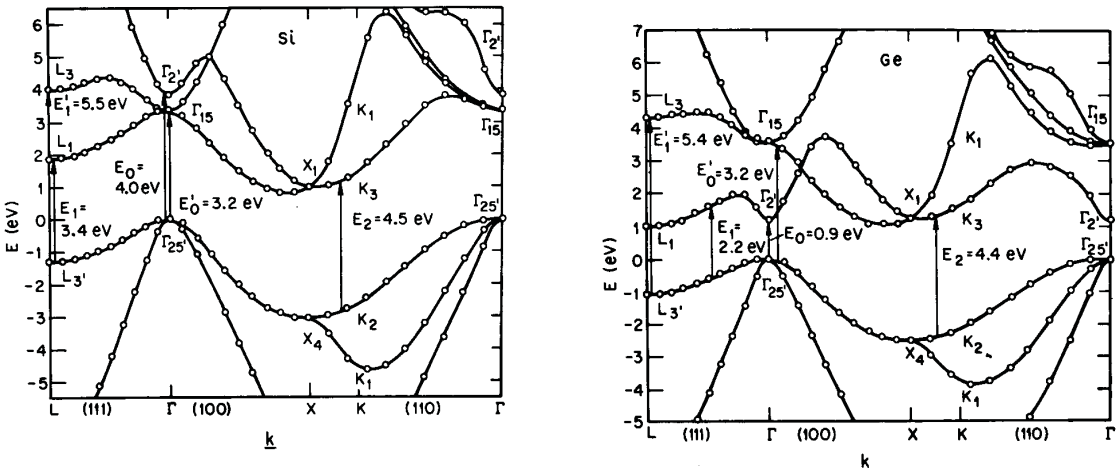


Figure 5.6: Pseudopotential band structure of Si and Ge [M.L. Cohen and T.K. Bergstresser *Phys. Rev.* **141**, 789 (1966)]. The energies of the optical transitions are taken from experiment.

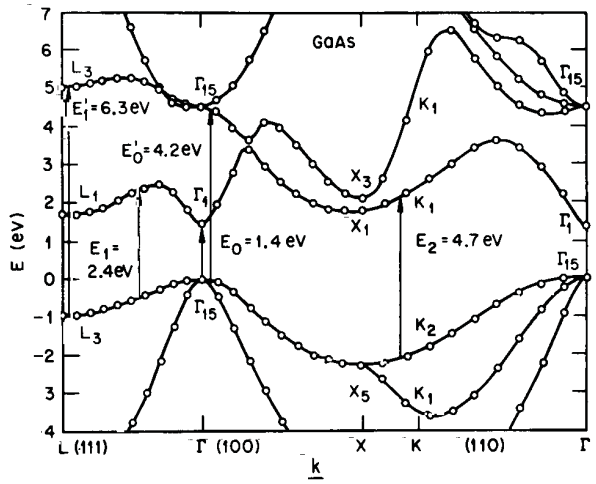


Figure 5.7: Band structure of GaAs [M.L.Cohen and T.K. Bergstresser *Phys. Rev.* **141**, 789 (1966)]

because the conduction bands have minima either in the (100) direction (Si) or the (111) direction (Ge).

5.3 Semiclassical model of electron dynamics

5.3.1 Wavepackets and equations of motion

We now want to discuss the dynamics of electrons in energy bands. Because the band structure is dispersive, we should treat particles as wave-packets. The band energy $\epsilon(\mathbf{k})$ is the frequency associated with the phase rotation of the wavefunction, $\psi_{\mathbf{k}} e^{-i\epsilon(\mathbf{k})t/\hbar}$, but for the motion of a wave packet in a dispersive band, we should use the group velocity, $d\omega/dk$, or as a vector

$$\dot{\mathbf{r}} = \mathbf{v}_g = \hbar^{-1} \nabla_{\mathbf{k}} \epsilon(\mathbf{k}) \quad , \tag{5.14}$$

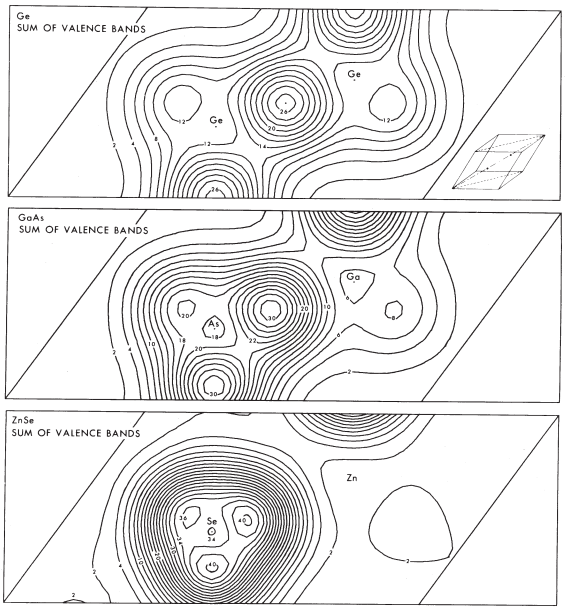


Figure 5.8: The valence charge density for Ge, GaAs, and ZnSe from an early pseudopotential calculation, plotted along a surface in a 110 plane that contains the two atoms of the unit cell. Note the (pseudo-)charge density shifting from the centre of the bond in Ge to be almost entirely ionic in ZnSe. [M.L. Cohen, *Science* **179**, 1189 (1973).]

where $\mathbf{r}$ is the position of the wavepacket. All the effects of the interaction with the lattice are contained in the dispersion $\epsilon(\mathbf{k})$.

If a force F is applied to a particle, the rate of doing work on the particle is

$$\frac{d\epsilon_{\mathbf{k}}}{dt} = \frac{d\epsilon}{dk}\frac{dk}{dt} = Fv_g \tag{5.15}$$

which leads to the key relation

$$\hbar\frac{d\mathbf{k}}{dt} = \mathbf{F} = -e(\mathbf{E} + \mathbf{v} \wedge \mathbf{B}) = -e(\mathbf{E} + \hbar^{-1}\nabla_{\mathbf{k}}\epsilon(\mathbf{k}) \wedge \mathbf{B}) \tag{5.16}$$

where we have introduced electric E and magnetic B fields.

The effect of an electric field is to shift the crystal momentum in the direction of the field, whereas the effect of a magnetic field is conservative - the motion in $\mathbf{k}$-space is normal to the gradient of the energy. Thus a magnetic field causes an electron to move on a line of constant energy, in a plane perpendicular to the magnetic field. This property is the basis of magnetic techniques for measuring Fermi surfaces of metals.

Bloch oscillations

Suppose we have a one-dimensional electron band, such as shown in Fig. 5.9. The group velocity is also shown — note that it reaches maximum size about half way to the zone boundary, and then decreases to zero at the zone boundary. If an electron in this band is subject to a constant

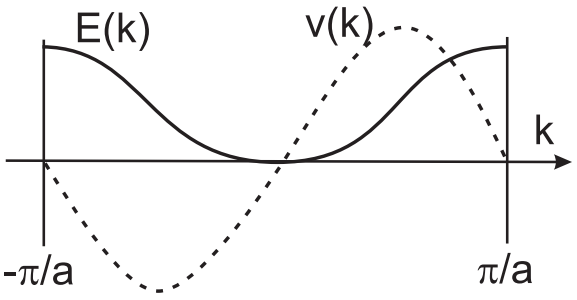


Figure 5.9: Band energy $E(k)$ (solid line) and group velocity $v(k)$ dashed line in a simple 1D band. A wavepacket progressing its crystal momentum according to (5.17) accelerates as k increases from zero, and then slows and reverses direction as k approaches the zone boundary.

electric field, we obtain

$$k(t) = k(0) - \frac{eEt}{\hbar} \quad , \tag{5.17}$$

so that the wave packet of electrons oscillates up and down the energy surface. It we start from the minimum of the band, then the group velocity grows linearly in time as for a free electron accelerating (though with a mass different from the free electron mass). However, on approaching the zone boundary, the group velocity slows - the acceleration of the particle is *opposite* to the applied force. What is actually happening is buried within the semiclassical model via the dispersion $\epsilon(k)$: as the wavepacket approaches the Brillouin zone boundary, real momentum (not crystal momentum k) is transferred to the lattice, so that on reaching the zone boundary the particle is Bragg-reflected.

Thus a DC electric field may be used - in principle - to generate an AC electrical current. All attempts to observe these *Bloch oscillations* in conventional solids has so far failed. The reason is that in practice it is impossible to have wavepackets reach such large values of momentum as π/a due to scattering from impurities and phonons in the solid. We will incorporate scattering processes in the theory in a moment.

It turns out however, that one can make *artificial* periodic potentials in a semiconductor superlattice. The details of this process will be discussed later, but for our purposes the net effect is to produce a square well potential that is periodic with a periodicity that can be much longer than the atom spacing. The corresponding momentum at the zone boundary is now much smaller, so the wavepacket does not have to be excited to such high velocities. The signature of the Bloch oscillations is microwave radiation produced by the oscillating charge - at a frequency that is proportional to the DC electrical field.

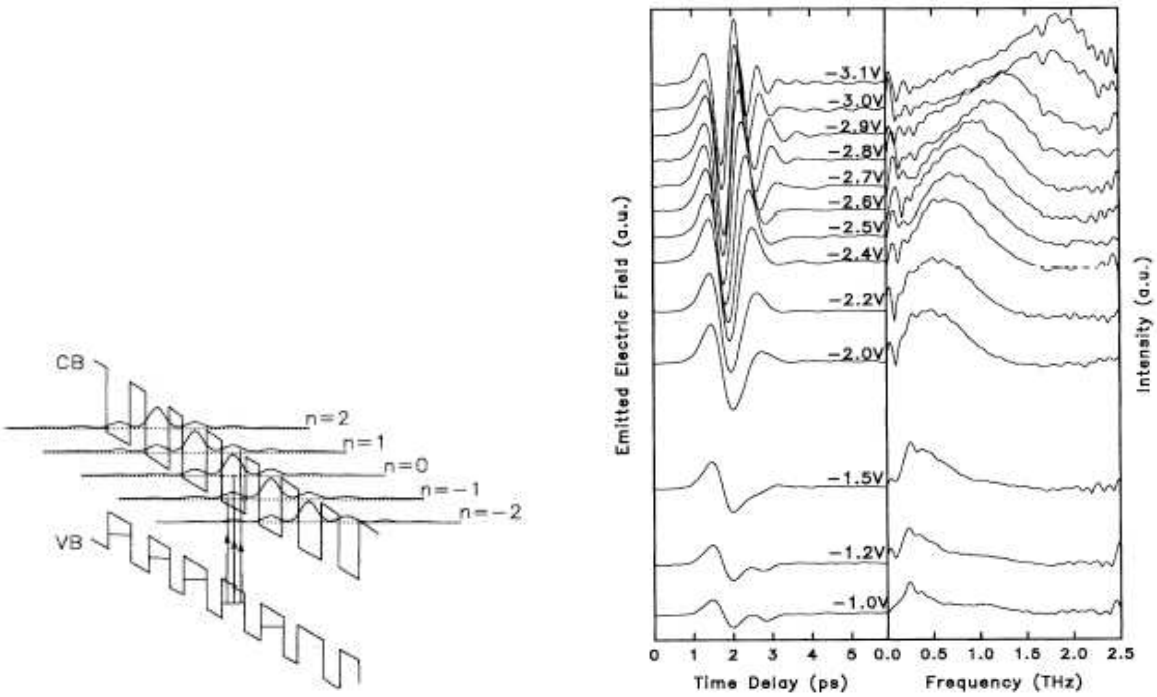


Figure 5.10: Schematic diagram of energy versus position of the conduction and electron bands in a periodic heterostructure lattice. The tilting is produced by the applied electric field. The levels shown form what is called a *Wannier-Stark* ladder for electron wavepackets made by excitation from the valence band in one quantum well, either vertically (n=0) or to neighbouring ($n = \pm 1$) or next-neighbouring $n = \pm 2$ wells of the electron lattice. In the experiment, electrons (and holes) are excited optically by a short-pulse laser whose frequency is just above the band gap of the semiconductor (i.e., a few $\times 10^{15}$ Hz). The electrical radiation (on a time scale of picoseconds) is monitored as a function of time and for different DC electrical biases, shown on the left panel. The spectral content is then determined by taking a Fourier transform of the wavepackets (right panel); at large negative voltages one sees a peak at a frequency that increases with increasing bias. The device is not symmetric, and therefore has an offset voltage of about -2.4 V before the Bloch oscillation regime is reached. [From Waschke et al., Physical Review Letters **70**, 3319 (1993).]

Approximations and justification for the semiclassical model

A full justification of the semiclassical model is not straightforward and we will not go into that here. [See Kittel, Appendix E, and for a more formal treatment J. Zak, Physical Review **168** 686 (1968)]

- Note that at least the semiclassical picture takes note of the fact that the Bloch states are stationary eigenstates of the full periodic potential of the lattice, and so there are no collisions with the ions.

- We must be actually describing the motion of a wavepacket

$$\psi_n(\mathbf{r},t) = \sum_{\mathbf{k}'} g(\mathbf{k}-\mathbf{k}')\psi_{n\mathbf{k}'}(\mathbf{r},t)\,\exp\left[-i\epsilon_n(\mathbf{k}')t/\hbar\right] \text{ where } g(k)\to 0 \text{ if } |\mathbf{k}| > \Delta k \tag{5.18}$$

  The wavepacket is described by a function $g(\mathbf{k})$ that is sharply peaked, of width Δk, say. Clearly $\Delta k \ll 1/a$, with a the lattice constant (otherwise the packet will disperse strongly).

- The size of the packet in real space is therefore $\Delta R \approx 1/\Delta k$. Consequently the semiclassical model can only be used to describe the response to fields that vary *slowly* in space, on a scale much larger than the lattice constant.
- The band index n is assumed to be a good quantum number. Clearly if the lattice potential were tiny, we would expect to return to free electrons, and be able to accelerate particles to high energies and make transitions between bands. Rather naturally the constraint is that the characteristic field energies should be small in comparison to the band gap E_{gap}: they are in fact

$$eEa \ll E_{gap}^2/E_F, \tag{5.19}$$

  with E_F the characteristic Fermi energy, or overall bandwidth. The electric fields in a metal rarely exceed 1 V m^{-1}, when the left hand side of this inequality is about 10^{-10} eV; not in danger.
- The corresponding constraint on magnetic fields is

$$\hbar\omega_c \ll E_{gap}^2/E_F \tag{5.20}$$

  with $\omega_c = eB/mc$ the cyclotron frequency. This corresponds to about 10^{-2} eV in a field of 1 Tesla, so that strong magnetic fields indeed may cause transitions between bands, a process of magnetic breakdown.
- The last condition is that, of course, the frequency of the fields must be much smaller than the transition energies between levels, i.e., $\hbar\omega \ll E_{gap}$.

5.3.2 Electrons and holes in semiconductors

An immediate consequence of this picture is that filled bands are inert. If all the electrons states in a zone are occupied, then the total current is obtained by integrating the group velocity over the whole zone; but the group velocity is the gradient of a periodic function; so this integral yields zero. Indeed all insulating solid elements have either even valence, or a lattice containing an even number of atoms in the basis, and therefore filled bands.

It is of interest to consider what happens to a filled band with one electron removed. This can be created by absorption of a photon whose energy exceeds the energy gap of a semiconductor, to make a transition of an electron from the valence band into the conduction band [See Fig. 5.11]. The removal of an electron from a filled band leaves a *hole*, which in fact can be viewed as a fermionic particle with distinct properties.

Hole momentum.

$$\mathbf{k}_h = -\mathbf{k}_e \tag{5.21}$$

This can be seen from the optical absorption experiment. The light produces a (nearly) vertical transition and gives no momentum to the electron hole pair. Since the initial state is a filled band with total momentum zero, (5.21) follows.

Hole energy.

$$\epsilon_h(\mathbf{k}_h) = -\epsilon_e(\mathbf{k}_e) \tag{5.22}$$

This sign is needed because (measuring energies from the top of the band) removing an electron of *lower* energy requires more work.

Hole velocity. A combination of the first two rules then gives

$$\mathbf{v_h} = \hbar^{-1}\nabla_{\mathbf{k}_h}\epsilon_h(\mathbf{k}_h) = \hbar^{-1}\nabla_{\mathbf{k}_e}\epsilon_e(\mathbf{k}_e) = \mathbf{v}_e \tag{5.23}$$

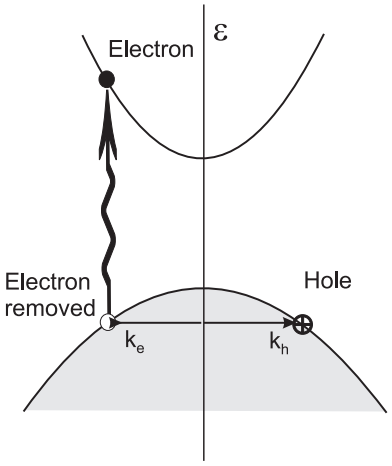


Figure 5.11: Absorption of a photon creates an electron-hole pair, with an energy $\epsilon_e+\epsilon_h+2E_{gap}$ but adds negligible momentum to the system. Hence the hole momentum is the negative of the momentum of the empty electronic state, and its energy is positive (measured conventionally from the top of the band).

Effective mass. The dispersion at the bottom (top) of the bands is parabolic, and therefore can be approximated as

$$\epsilon = \epsilon_0 + \frac{\hbar^2 k^2}{2m^*} \tag{5.24}$$

defining an effective mass m^*. We have

$$m_h^* = -m_e^* \tag{5.25}$$

so the hole mass is positive at the top of the electron band.

Hole charge. The effective charge of a hole is positive, as can be seen by taking the equation of motion for the electron

$$\hbar \frac{d\mathbf{k}_e}{dt} = -e(\mathbf{E} + \mathbf{v_e} \wedge \mathbf{B}) \tag{5.26}$$

and making the replacement $k_e \to -k_h$ and $v_e \to v_h$, giving

$$\hbar \frac{d\mathbf{k}_h}{dt} = e(\mathbf{E} + \mathbf{v_h} \wedge \mathbf{B}). \tag{5.27}$$

The same result comes from noticing that the current carried by the hole ev_h must be the same as the (missing) current (not) carried by the empty electron state.

Chapter 6

Experimental probes of the band structure

6.1 Optical transitions

The band structure provides the excitation spectrum of the solid. The ground state of the system involves filling states up to the Fermi energy, but we can also excite the system in different ways. One of the simplest is the absorption of a photon, which can be visualised as an excitation of an electron from an occupied state into an empty state, leaving behind a "hole" in the valence band. See Fig. 6.1.

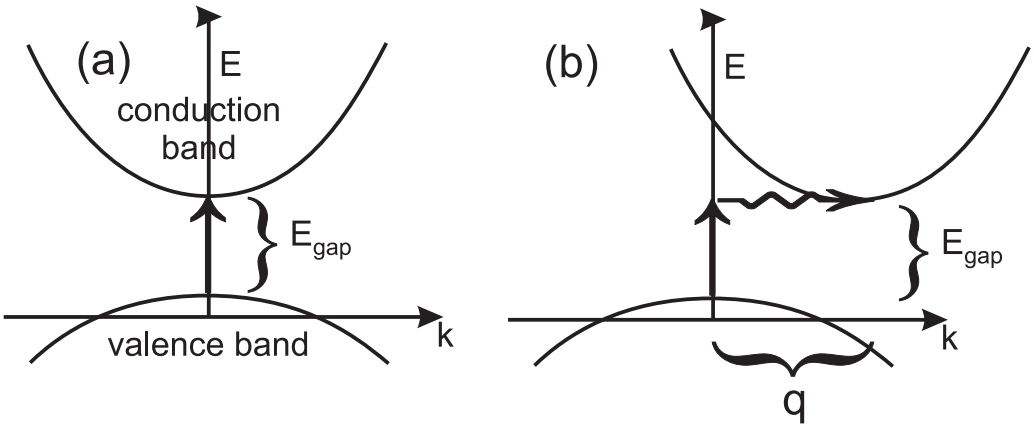


Figure 6.1: Direct absorption by light is a nearly vertical transition since the wavevector of a photon with energy of order a semiconductor gap is much smaller than the typical momentum of an electron. (a) In a *direct gap* semiconductor, such as GaAs, the lowest energy available states for hole and electron are at the same momentum, and the optical threshold is at the vertical energy gap. (b) In an indirect gap material (e.g., Si or Ge), the minimum energy excitation of electron and hole pair connects state of different momenta - and a phonon of momentum q must be excited concurrently with the photon.

The minimum gap in a semiconductor is the energy difference between the highest occupied state and the lowest unoccupied state, and this is the threshold for optical absorption (neglecting

excitonic physics, see later). In some semiconductors, the maximum valence band state and the minimum in the conduction band occur at the same momentum - in such a *direct gap* system, direct optical excitation is allowed at the minimum gap, and an important example is $GaAs$.

Si and Ge are example of *indirect gap* materials, because the conduction band minimum is toward the edge of the zone boundary. The minimum energy transition is at large momentum, and therefore cannot be accomplished by direct absorption of a photon. The lowest energy transition is instead a *phonon-mediated* transition where the energy is provided by the photon and the momentum provided by the phonon. This is much less efficient than direct optical absorption.

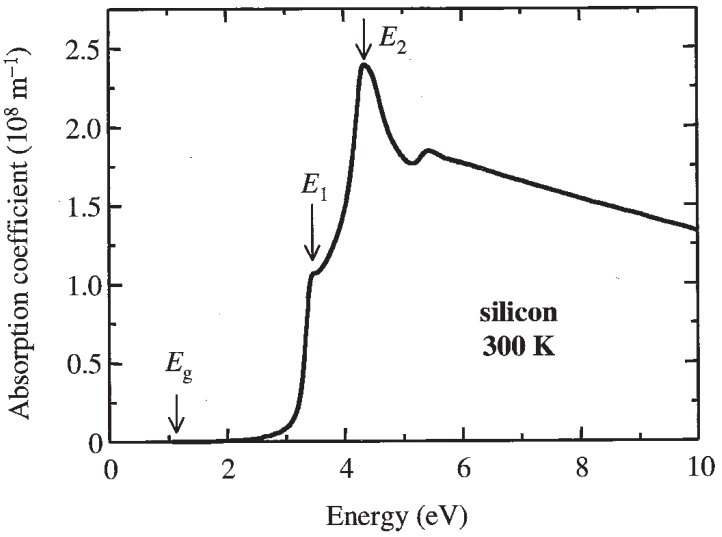


Figure 6.2: The interband absorption spectrum of Si has a threshold at the indirect gap $E_g \approx 1.1\ eV$ which involves a phonon and is very weak. The energies E_1 and E_2 correspond to critical points where the conduction and valence bands are vertically parallel to one another; absorption is direct (more efficient) and also enhanced by the enhanced joint density of electron and hole states. [E.D. Palik, *Handbook of the optical constants of solids*, AP, 1985]
.

Luminescence is the inverse process of recombination of an electron-hole pair to emit light. Luminescence appears if electrons and holes are injected into a semiconductor (perhaps electrically, as in a light-emitting diode). Obviously, this process will not be efficient in an indirect gap semiconductor but it is more efficient in a direct gap material. This simple fact explains why $GaAs$ and other III-V compounds are the basis of most practical opto-electronics in use today, whereas Si is the workhorse of electrical devices.

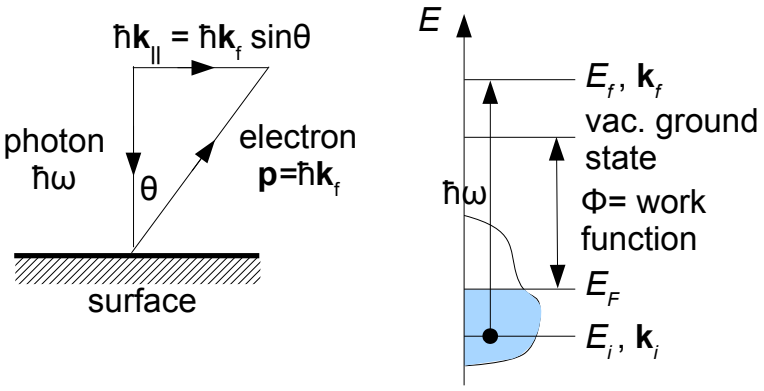


Figure 6.3: Schematics of an Angular Resolved Photoemission Spectroscopy (ARPES) experiment. The incoming photons transfer negligible momentum, so the electrons are excited from the valence bands to high energy excited states (above the vacuum energy necessary to escape from the crystal) with the same crystal momentum. When the excited electrons escape through the surface of the crystal, their momentum perpendicular to the surface will be changed. If the surface is smooth enough, the momentum of the electron parallel to the surface is conserved, so the angle of the detector can be used to scan $k_{\|}$.

6.2 Photoemission

The most direct way to measure the electron spectral function directly is by photoemission, although this is a difficult experiment to do with high resolution. In a photoemission experiment, photons are incident on a solid, and cause transitions from occupied states to plane wave-like states well above the vacuum energy; the excited electron leaves the crystal and is collected in a detector that analyses both its energy and momentum.[1] The photon carries very little momentum, so the momentum of the final electron parallel to the surface is the same as the initial state in the solid, while of course the perpendicular component of the momentum is not conserved.

We can relate the energy of the outgoing electrons, E_f, to the energy of the incoming photons, $\hbar\omega$, the work function ϕ and the initial energy of the electron in the solid before it is ejected, E_i (Fig. 6.3). Conservation of energy and of the momentum component parallel to the surface then give:

$$E_f = \frac{\hbar^2 k_f^2}{2m} = E_i + \hbar\omega - \phi \qquad k_{f||} = k_{i||}$$

Here, E_i is referenced to the Fermi energy E_F, whereas E_f is referenced to the vacuum ground state energy. The detector angle θ is used to extract $k_{\|}$.

Photoemission data is therefore easiest to interpret when there is little dispersion of the electronic bands perpendicular to the surface, as occurs in anisotropic layered materials. It is fortunate that there are many interesting materials (including high-temperature superconductors) in this class.

[1] For a detailed discussion of photoemission experiments, see Z.X. Shen and D.S. Dessau, *Physics Reports*, **253**, 1-162 (1995)

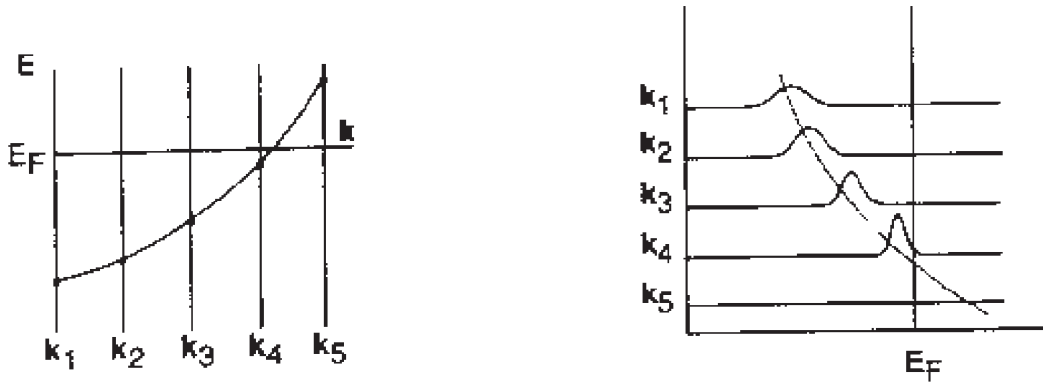


Figure 6.4: Idealised results from a photoemission experiment. A peak is observed at the band energy in each spectrum, but disappears when the band crosses the Fermi energy, where there are no more electrons to excite.

If one analyses both the energy and the momentum of the outgoing electron, (this is Angle Resolved Photo-Emission Spectroscopy, or ARPES) one can determine the band structure directly. Integrating over all angles gives a spectrum that is proportional to the total density of states.

Fig. 6.4 illustrates schematically how such an analysis might proceed, and Fig. 6.5 shows some results derived from ARPES in the oxide superconductor Sr_2RuO_4.

Photoemission can give information only about occupied states. The technique of *inverse photoemission* involves inserting an electron of known energy into a sample and measuring the ejected photon. Since the added electron must go into unoccupied state, this spectroscopy allows one to map out unoccupied bands, providing information complementary to photoemission.

6.3 Quantum oscillations – de Haas van Alphen effect

In high magnetic fields $B > 1$ T and in pure samples, many material properties have been found to oscillate as a function of applied magnetic field. The form of these *quantum oscillations*, in particular the frequency of oscillation, can be used to infer the shape of the Fermi surface and other key electronic properties.

6.3.1 Size of cyclotron orbits

A full quantum mechanical treatment of the motion of electrons in a strong magnetic field is problematic. When the lattice potential can be neglected, for free electrons, the Schrödinger equation can be solved directly. For real materials, however, the lattice potential is essential to the band structure and cannot be neglected. In this case, progress can be made with a semiclassical treatment which makes use of the Bohr-Sommerfeld quantisation condition (see,

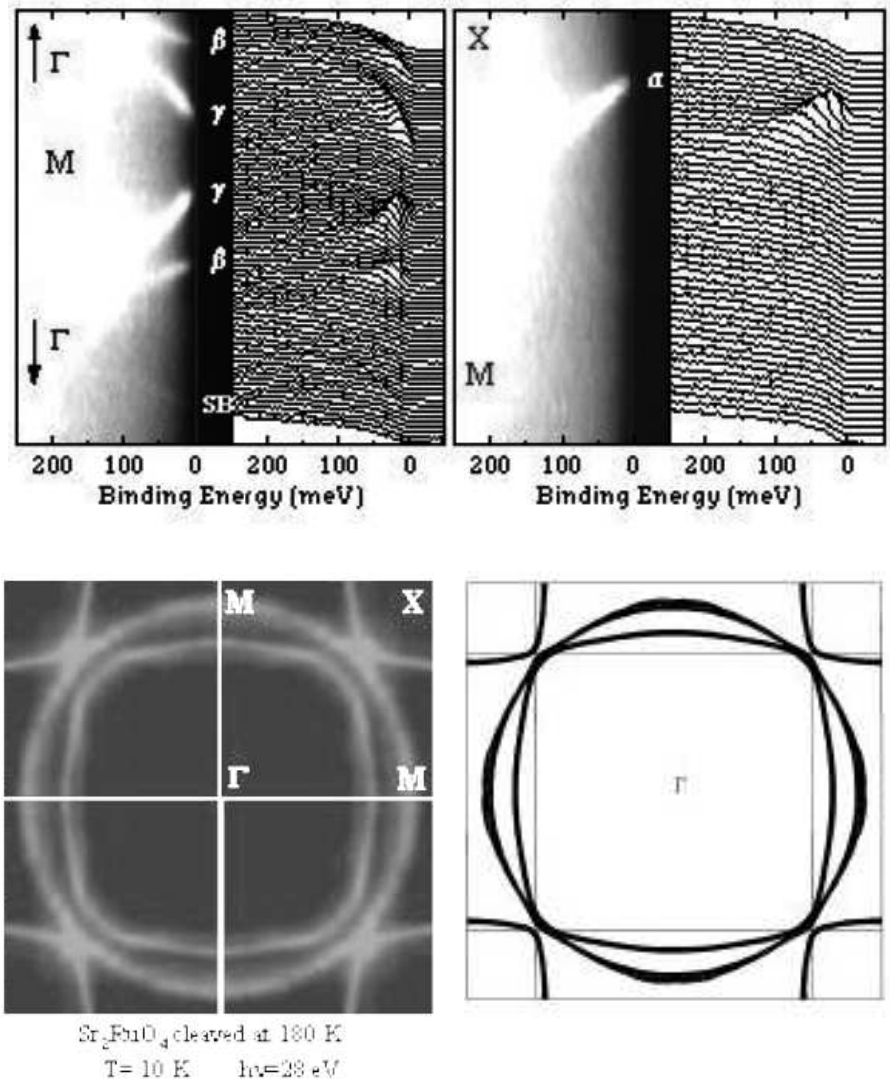


Figure 6.5: Photoemission spectra on the two dimensional layered metal Sr_2RuO_4. The bands are nearly two-dimensional in character, so the interpretation of the photoemission data is straightforward – different angles (see Fig. 6.3) correspond to different in-plane momenta. The upper panels show energy scans for different angles that correspond to changing the in-plane momentum in the direction from the centre of the Brillouin zone Γ towards the centre of the zone face M and the corner X. Several bands cross the Fermi energy, with different velocities, and sharpen as their energies approach E_F. The left hand lower panel plots the positions of the peaks as a function of momentum at the Fermi energy, to be compared with the band structure calculation of the Fermi surface(s) on the lower right. [Experiment from Damascelli et al., PRL; theory from Mazin et al. PRL 79, 733 (1997)]

e.g., Kittel ch. 9):

$$\oint \mathbf{p} d\mathbf{r} = \left(n + \frac{1}{2}\right) h \tag{6.1}$$

Here, $\mathbf{p}$ is the *canonical momentum*, conjugate to the position $\mathbf{r}$. The canonical momentum can be written as the sum of the kinetic (or mv-) momentum, $m\mathbf{v}$, and the field momentum, $q\mathbf{A}$.

Particles with charge q moving in a strong magnetic field $\mathbf{B}$ are forced into an orbit by the Lorentz force: $m\dot{\mathbf{v}} = q\dot{\mathbf{r}} \times \mathbf{B}$. This relation connects the components of velocity and position of the particle in the plane perpendicular to $\mathbf{B}$ and can be integrated: $m\mathbf{v}_\perp = q\mathbf{r} \times \mathbf{B}$, where $\mathbf{r}$ is measured from the centre of the orbit.

This allows us to write $\mathbf{p} = m\mathbf{v} + q\mathbf{A} = q(\mathbf{r} \times \mathbf{B} + \mathbf{A})$, and using $\oint \mathbf{A} d\mathbf{r} = \Phi$ (the magnetic flux), we obtain:

$$\oint \mathbf{p} d\mathbf{r} = q \oint \mathbf{r} \times \mathbf{B} d\mathbf{r} + q\Phi = -q\Phi \ , \tag{6.2}$$

because $\oint \mathbf{r} \times \mathbf{B} d\mathbf{r} = -\mathbf{B} \oint \mathbf{r} \times d\mathbf{r} = -2\mathbf{B}A_r$, where A_r is the real space area enclosed by the orbit's projection onto the plane perpendicular to $\mathbf{B}$.

We arrive at the conclusion that the flux threading the real space orbit is quantised:

$$\Phi_n = \mathbf{A}_r^{(n)} \mathbf{B}_n = \left(n + \frac{1}{2}\right) \frac{h}{e} \tag{6.3}$$

Can we relate the motion of the electron in real space to the accompanying motion in $\mathbf{k}$-space? From our earlier result for the relation between momentum and position, $m\mathbf{v}_\perp = \hbar\mathbf{k}_\perp = q\mathbf{r} \times \mathbf{B}$, we find that the $\mathbf{k}$-space orbit has the same shape as the real space orbit, but is turned by 90 degrees and stretched by $\frac{Bq}{\hbar}$. This means that the area enclosed by the $\mathbf{k}$-space orbit A_k is

$$A_k = \left(\frac{e}{\hbar}\right)^2 B^2 A_r \ , \tag{6.4}$$

where q has been replaced by the electron charge e has. Combining this result with Eqn. 6.3, we find

$$A_k = \frac{2\pi e}{\hbar} B(n + \frac{1}{2}) \tag{6.5}$$

6.3.2 Density of states oscillations

In a magnetic field, the allowed $\mathbf{k}$-states no longer form a regular lattice in reciprocal space, as $\mathbf{k}$ is no longer a good quantum number. All the $\mathbf{k}$-states in the vicinity of a $\mathbf{k}$-orbit superimpose to form the orbital motion of the electrons. The electrons now 'live' on a set of cylinders, the *Landau tubes*, with quantised cross-sectional areas.

These cylinders, whose cross-sectional area expands with increasing field B, cut through the zero-field Fermi surface of the metal. What effect will this have on the B-dependence of

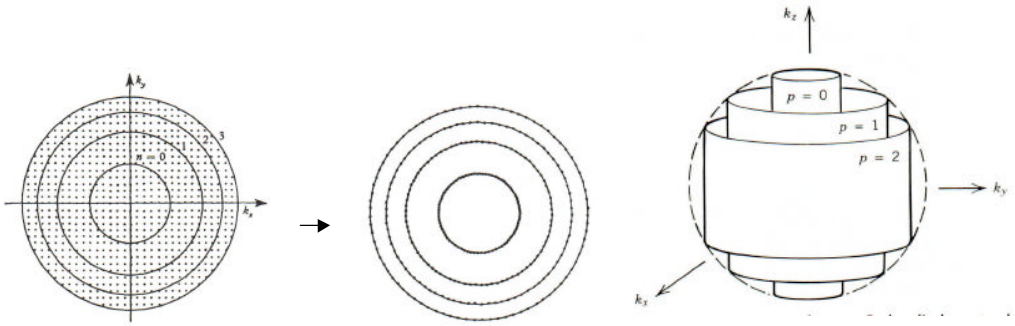


Figure 6.6: Quantisation of **k**-space orbits in high magnetic fields. In 2D, the grid of allowed states in zero field collapses onto rings spaced according to the Onsager relation (left and middle panel). In 3D, these rings extrude to cylinders - the *Landau tubes* (right panel). Quantum oscillations will be detected for extremal Fermi surface cross-sections, as successive Landau tubes push through the Fermi surface with increasing magnetic field **B**.

the density of states at the Fermi level, $g(E_F)$? Considering a particular slice $\perp$ **B** through the Fermi surface with area A_k, this will now only contribute to $g(E_F)$ if its area coincides with the area of one of the Landau tubes. As B increases, one Landau tube after the other will satisfy this condition, at field values $\frac{1}{B_n} = \frac{2\pi e}{\hbar A_k}(n + \frac{1}{2})$. Consequently, the contribution of this slice to $g(E_F)$ oscillates with a period

$$\Delta\left(\frac{1}{B}\right) = \frac{1}{B_{n+1}} - \frac{1}{B_n} = \frac{2\pi e}{\hbar}\frac{1}{A_k} \tag{6.6}$$

This is the *Onsager relation*, which links the period of quantum oscillations to the cross-sectional area of the Fermi surface.

There remains one important consideration: in reality, we can only measure quantum oscillations associated with *extremal* orbits. These arise, where a Landau tube can touch, rather than cut through, the Fermi surface. At such regions of the Fermi surface, there are many close-lying orbits with nearly identical cross-section, which causes the corresponding density of states oscillations to add coherently. For the rest of the Fermi surface, the oscillations attributed to each orbit have different period and they add incoherently, which wipes out the effect.

6.3.3 Experimental observation of quantum oscillations

Many observable properties depend directly on the density of states at the Fermi level, and many of these have been used to detect quantum oscillations. The classic example is the magnetic susceptibility χ, which according to simple theory is proportional to $g(E_F)$. Measurements of $\chi(B)$ at low temperature exhibit oscillations which, when plotted versus $(1/B)$ allow the determination of extremal Fermi surface cross-sections. This is called the *de Haas-van Alphen effect*.

Similar oscillations can be observed in measurements of electrical resistivity ('Shubnikov-de Haas'), of the magnetisation, of the sample length and of the entropy – which can be picked up by measuring the temperature oscillations of a thermally isolated sample. Generally, these

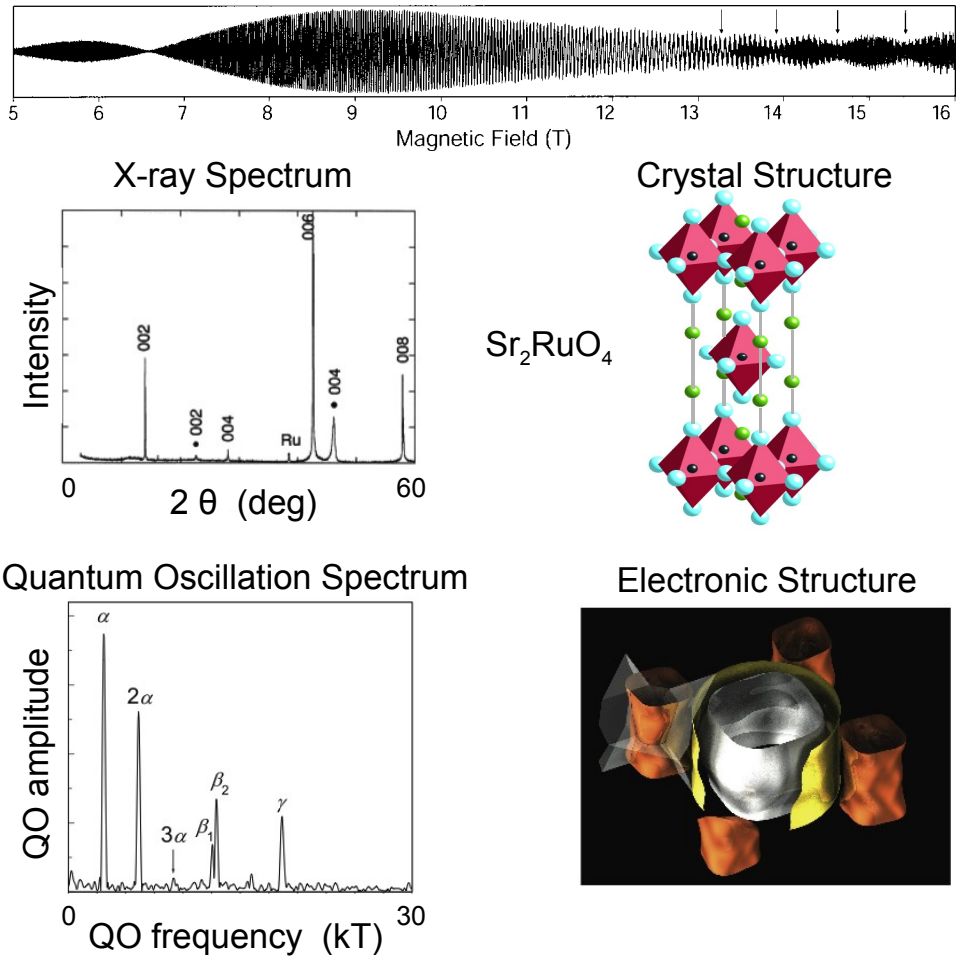


Figure 6.7: **Analogy between quantum oscillations and x-ray diffraction in Sr_2RuO_4:** Obtaining the lattice structure from an x-ray diffraction experiment involves solving an inverse problem, in which the spatial frequencies produced by plausible lattice structures are compared against those measured as peaks in the diffraction pattern. Similarly, by analysing the periodicities in bulk properties such as magnetisation (top panel) or electrical resistivity in magnetic field sweeps and matching them against those expected from numerical calculations, we can determine the *electronic structure* of metals. Advances in crystal growth and the availability of very high magnetic fields have allowed this technique to be applied in some of the most complex materials currently studied in condensed matter research, including the high temperature superconducting cuprates and ferro-pnictides.

experiments require:

- High purity samples: the electronic mean free path must be long enough to allow the electrons to complete roughly one cyclotron orbit before scattering.

- High magnetic field: high magnetic fields make the cyclotron orbits tighter, which equally helps to fulfil the mean free path condition.

- Low temperature: The density of states oscillations are smeared out, when the Fermi surface itself is smeared by thermal broadening of the Fermi-Dirac distribution. Typically,

experiments are carried out below 1 K for transition metal compounds, and below 100 mK for heavy fermion compounds (see Appendix).

6.4 Tunnelling

Tunnelling spectroscopies (injecting or removing electrons) through a barrier have now evolved to be very important probes of materials. The principle here is that a potential barrier allows one to maintain a probe (usually a simple metal) at an electrical bias different from the chemical potential of the material. Thus the current passed through the barrier comes from a non-equilibrium injection (tunnelling) through the barrier.

A model for a simple metal tunnelling into a more complex material is shown in Fig. 6.8. With the metal and sample maintained at different electrical potentials separated by a bias voltage eV, then the current through the junction can be estimated to be of the form

$$I \propto \int_{\mu+eV}^{\mu} g_L(\omega) g_R(\omega) T(\omega) \tag{6.7}$$

where T is the transmission through the barrier for an electron of energy ω and g_L and g_R are the densities of states.[2] If the barrier is very high so that T is not a strong function of energy, and if the density of states in the contact/probe is approximately constant, then the energy-dependence comes entirely from the density of states inside the material. Notice then that the differential conductivity is proportional to the density of states (see Fig. 6.8):

$$dI/dV \propto g(\mu + eV) \quad . \tag{6.8}$$

It is difficult to maintain very large biases, so most experiments are limited to probing electronic structure within a volt or so of the Fermi energy.

Tunnel junctions are sometimes fabricated by deposition of a thin insulating layer followed by a metal contact.

The technique of *scanning tunnelling microscopy* (STM) uses a small tip, with vacuum as the surface barrier. Because the tunnel probability is an exponential function of the barrier thickness, this scheme provides high (close to atomic, in some cases) spatial resolution, even though the tip radius will be nm or larger. By hooking this up to a piezoelectric drive in a feedback loop, it has proved possible to provide not only $I-V$ characteristics at a single point, but also spatial maps of the surface.

Scanned probe spectroscopies have advanced to become extraordinary tools at the nanoscale. As well as *STM*, it is possible to measure forces near a surface (*atomic force microscopy, AFM*), which is particularly useful for insulating samples. It has proved possible to manipulate individual atoms, to measure the magnetism of a single spin, and with small single-electron transistors to study to motion of single electron charges in the material.

[2]Strictly this formula applies when the tunnelling process does not conserve momentum parallel to the interface, i.e., if the surface is rough or disordered.

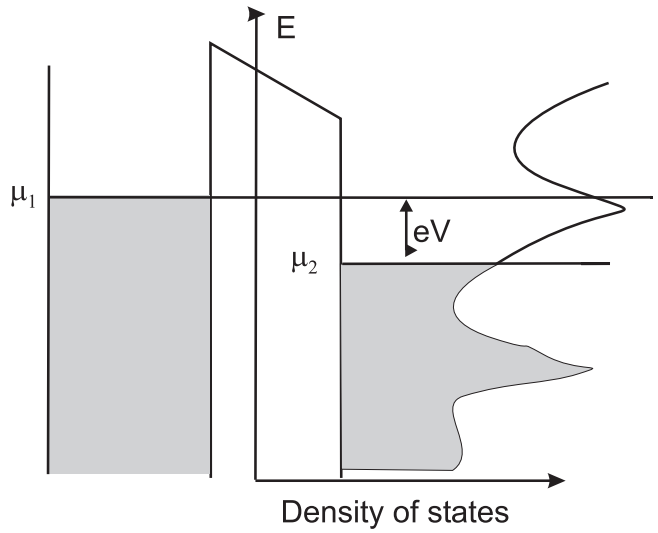


Figure 6.8: Schematic description of tunnelling between two materials maintained at a relative bias eV. The current is approximately given by the integrated area between the two chemical potentials (provided the matrix element for tunnelling is taken as constant.) If the density of states of the contact (or probe, labelled 1 in the figure) is also slowly varying, then the differential conductance dI/dV is proportional to the density of states of the material itself, at the bias eV above the chemical potential μ_2.

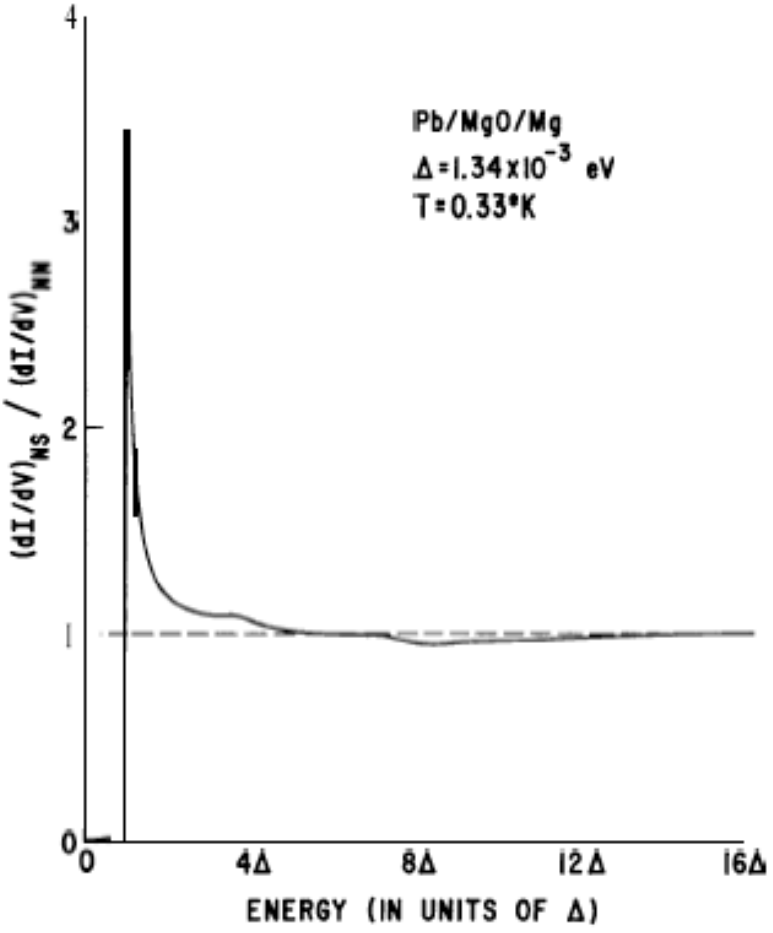


Figure 6.9: Differential conductance of a tunnel junction between superconducting Pb and metallic Mg reveals the gap in the density of states of superconducting lead. [I. Giaever, Nobel Prize Lecture, 1973]

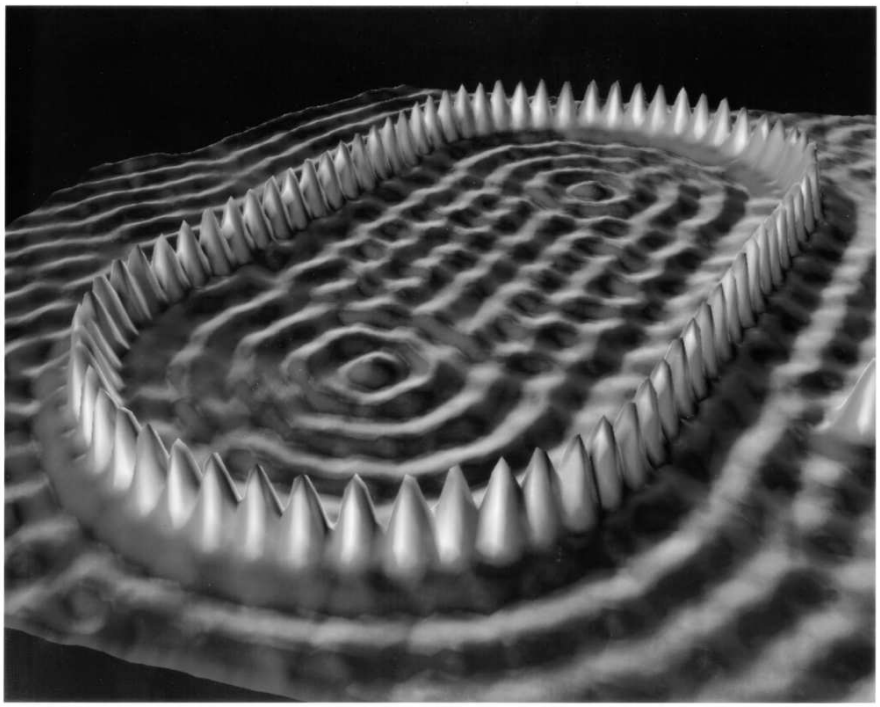


Figure 6.10: An array of Fe atoms arranged in a corral on the surface of Cu traps a surface electron state whose density can be imaged by STM. M.F. Crommie, C.P. Lutz, D.M. Eigler, E.J. Heller. Surface Review and Letters 2 (1), 127-137 (1995).

Chapter 7

Semiconductors

7.1 Semiconductor band structure

Direct gap semiconductors

The band structure near $\mathbf{k} = 0$ of a diamond-structure (Si, Ge) or zincblende-structure ($GaAs$) semiconductor is shown in Fig. 7.1. The conduction band is a simple parabola, but the valence bands are more complex. The complexity arises because the symmetry of the valence bands is p-like and there are three degenerate bands (in cubic symmetry) at $\mathbf{k} = 0$. At finite $\mathbf{k}$ they split into *light hole* and *heavy hole* bands, so called because of the difference in the electron masses. Additionally, there is a deeper lying band, split off by spin-orbit interactions from the others. This is usually not important for thermally excited carriers.

The band masses are quite different from free electron masses, for example, in $GaAs$ $m_e^* = 0.066$, $m_{lh}^* = 0.082$, $m_{hh}^* = 0.17$ (in units of the free electron mass). The cubic symmetry of the crystal means that the bands are isotropic (to order k^2).

Indirect gap semiconductors

As we remarked earlier, while there is a local minimum at the origin (the Γ-point), the conduction bands of Si and Ge do not have their global minima at the Γ-point, but far out in the zone.

The conduction band minima of Ge are at the eight equivalent L-points $2\pi/a(0.5\ 0.5\ 0.5)$, on the surface of the Brillouin zone. Here the band edges have a spheroidal energy surface, and are not isotropic as near the centre of the zone. In Ge, the longitudinal mass – along (111) – is $m_l = 1.59\,m$, much larger than the tranverse mass $m_t = 0.082\,m$.

In Si the conduction band minima are along the six (100) directions, close to the zone boundary at X $[2\pi/a(100)]$. The constant energy surfaces are ellipsoids, $m_l = 0.92\,m$, and $m_t = 0.19\,m$.

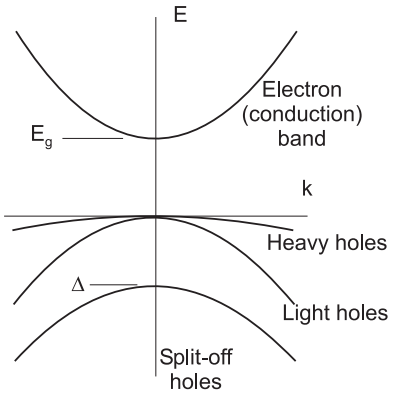


Figure 7.1: Sketch of the valence bands of diamond or zincblende structure semiconductors near the Γ-point ($\mathbf{k} = (000)$). The lowest hole band - the spin-orbit split-off band - is lower by an energy Δ that is a few tenths of an eV and therefore not relevant for thermally excited carriers at room temperature and below. In III-V semiconductors the absolute minimum in the conduction band is at Γ; in *Si* and *Ge* the absolute minimum of the conduction band is elsewhere in the zone.

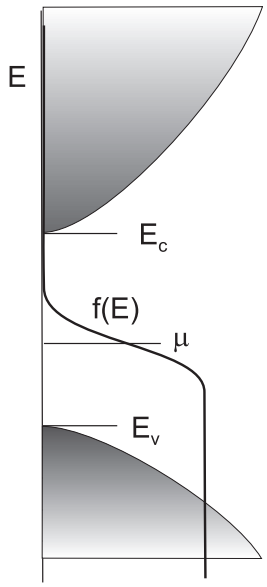


Figure 7.2: Density of states for electrons and the Fermi function determining the occupancy of the thermally excited states in an intrinsic semiconductor. The chemical potential lies mid-gap, and the temperature is assumed small in comparison to the gap.

7.2 Intrinsic carrier concentration

Semiconductors are materials where the energy gap is small enough that thermal excitation of carriers across the gap is important. Here we calculate the thermal *intrinsic* carrier concentration in a model semiconductor with parabolic electron and hole bands. The conduction and

valence band dispersions are therefore (see Fig. 7.2)

$$E_c(k) = E_c + \frac{\hbar^2 k^2}{2m_e^*} \;\; ; \;\; E_v(k) = E_v - \frac{\hbar^2 k^2}{2m_h^*} \tag{7.1}$$

We shall need the densities of states for the conduction band

$$g_e(E) = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{3/2}(E - E_c)^{1/2} \tag{7.2}$$

and for the valence band

$$g_h(E) = \frac{1}{2\pi^2}\left(\frac{2m_h^*}{\hbar^2}\right)^{3/2}(E_v - E)^{1/2} \;\; . \tag{7.3}$$

We can calculate the carrier density once the chemical potential μ is known. For electrons in the conduction band

$$n = \int_{E_c}^{\infty} dE\, g_e(E) f(E) \tag{7.4}$$

with f the Fermi function

$$f(E) = \frac{1}{e^{(E-\mu)/(k_B T)} + 1} \approx e^{-(E-\mu)/(k_B T)} \tag{7.5}$$

with the latter approximation valid when $E - \mu >> k_B T$ (non-degenerate Fermi gas). This gives

$$n \approx 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-\frac{E_c-\mu}{k_B T}} \tag{7.6}$$

A similar calculation determines the concentration of holes

$$p \approx 2\left(\frac{m_h^* k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-\frac{\mu-E_v}{k_B T}} \tag{7.7}$$

Note that the prefactors to the Boltzmann factors $e^{-\frac{E_c-\mu}{k_B T}}$ and $e^{-\frac{\mu-E_v}{k_B T}}$ can conveniently be absorbed into temperature-dependent concentrations

$$n_c(T) = 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2} \tag{7.8}$$

$$n_v(T) = 2\left(\frac{m_h^* k_B T}{2\pi\hbar^2}\right)^{3/2} \tag{7.9}$$

These functions express the (temperature dependent) number of states within range $k_B T$ of the band edge for the conduction and valence band, respectively. The resulting expression for the number of electrons in the conduction band

$$n = n_c(T) e^{-\frac{E_c-\mu}{k_B T}} \tag{7.10}$$

is formally identical to that obtained for a set of $n_c(T)$ degenerate energy levels at energy E_c, i.e., bunched up at the band edge. For the number of holes in the valence band, we obtain, equivalently:

$$p = n_v(T)e^{-\frac{\mu - E_v}{k_B T}} \tag{7.11}$$

Eq. (7.10) and (7.11) give the concentration of electrons and holes at a temperature T, in terms of the chemical potential μ, as yet unknown. It is useful to notice that the product

$$np = n_c(T)n_v(T)e^{-\frac{E_g}{k_B T}} \tag{7.12}$$

is independent of μ. Here, $E_g = E_c - E_v \simeq 1eV$ is the size of the energy gap. This result is also called the *law of mass action*. We will be able to use it when carriers are introduced by doping.

For an intrinsic semiconductor, the electron and hole densities are equal, and can be obtained my taking the square root of (7.12)

$$n_i = p_i = \left(n_c(T)p_v(T)\right)^{1/2} e^{-\frac{E_g}{2k_B T}} \tag{7.13}$$

and substituting back into either the equation for n (7.10) or p (7.11) yields the chemical potential

$$\mu = \frac{1}{2}E_g + \frac{3}{4}k_B T \log(m_h^*/m_e^*) \ . \tag{7.14}$$

The chemical potential thus sits mid gap at zero temperature, and shifts slightly away from that position if the carrier masses are different. Note that the activation energy to create intrinsic carriers (either electrons or holes) is always exactly *half* the optical energy gap.

7.3 Doped semiconductors

What differentiates semiconductors from insulators is the fact that the energy gap E_g is sufficiently small in semiconductors to allow significant carrier concentrations at room temperatures by thermal activation alone. However, carriers can also be created in semiconductors by adding impurity atoms in a process called *doping*.

Donor levels. Consider the effect in a Si crystal of replacing a single Si atom by an As atom. As is a group V element and therefore provides 5 electrons instead the of the 4 of the Si it replaced. Formally, it appears like a Si atom with one extra electron, and one extra positive charge in the nucleus. We now ask whether the added electron stays tightly bound to the extra positive charge.

Suppose the electron wanders away from the impurity site. It will of course see an attractive Coulomb force from the charged As impurity. Because the As atom carries a single positive charge, the energy levels are calculated in the same way as those of the Hydrogen atom. We take into account the influence of the surrounding material, in which the extra electron moves, by making two corrections: (i) the Coulomb potential is screened by the dielectric constant of Si ($\epsilon \approx 12$), so it is much weaker than in free space; and (ii) the band mass of the electron is

smaller than the free electron mass, so the kinetic energy of an electron in a given momentum state is larger. The net effect is that the binding energy of the $1s$ impurity state is now

$$\Delta_d = \frac{e^4 m_c^*}{2(4\pi\epsilon\epsilon_o\hbar)^2} = \frac{m_c^*/m_e}{\epsilon^2} \times 13.6 \text{ eV} \quad , \tag{7.15}$$

which can be very small compared to the band gap, and often comparable or smaller than thermal energies. Such *donor* impurities readily donate electrons to the conduction band.

The binding energy of the electronic states of the hydrogen atom express the energy difference between the lowest vacuum state and the respective bound states. The hydrogen-like bound impurity states we calculated are referenced to the bottom of the conduction band, because the electron unbinds from the impurity by occupying a conduction band state, just as an electron unbinds from the Hydrogen atom by assuming a plane-wave vacuum state.

As donors in Si have an ionisation energy of 50 meV; donors in $GaAs$ have an ionisation energy of about 6 meV, which is approximately 50 Kelvin. When a donor atom is ionised, it releases its formerly bound electron into the conduction band.

Acceptor levels. A trivalent impurity (e.g., B in Si) appears like a Si atom with an added *negative* charge, and with a missing electron. It is the mirror image of the case of the donor impurity, and corresponds to a positive charge (a 'hole') circling a negatively charged nucleus. As in the case of donor atoms, the binding energy of the hole is reduced by the combined effects of high permittivity and low band mass in semiconductors. When the hole unbinds, the impurity *accepts* an electron from the valence band. The accepted electron is used to complete the covalent bond with the neighbouring atoms, and renders the site *negatively* charged. So, while ionising a donor atom releases an electron into the conduction band, ionising an acceptor atom absorbs an electron from the valence band, which leaves a hole in the valence band.

n- and p-type materials. Even for very low densities of impurities, since the donor or acceptor energies are much smaller than the gap, impurities in semiconductors are often the principal source of electrically active carriers. If donor atoms predominate, the carriers are predominantly electrons, and the material is said to be *n-type*. If holes are the dominant carrier type, the material is called *p-type*. Experimentally, these regimes may be distinguished by measuring the Hall effect, whose sign depends on the carrier type.

Impurity ionisation. Here we quote here the results for thermal ionisation of the carriers in simple limits. If there are no acceptors, the carrier concentration at low temperatures is

$$n = (n_c N_d)^{1/2} e^{-\frac{\Delta_d}{2k_B T}} \tag{7.16}$$

where N_d is the donor density and the factor $n_c = 2(m_e^* k_B T/2\pi\hbar^2)^{3/2}$ is the effective density of electron states within an energy $k_B T$ of the band edge. Notice again that the activation energy is *half* the binding energy.

Since Δ_d is small and $n_c(T)$ is usually large compared to N_d, donor atoms are fully ionised down to very low temperatures, and $n \approx N_d$. This is called the *extrinsic* regime. At a still higher temperature, the *intrinsic* carrier generation by thermal activation from the valence band into the conduction band takes over.

If there are only acceptors and no donors, then a similar formula can be obtained for holes by inspection. When both donors and acceptors exist, the problem is in general more

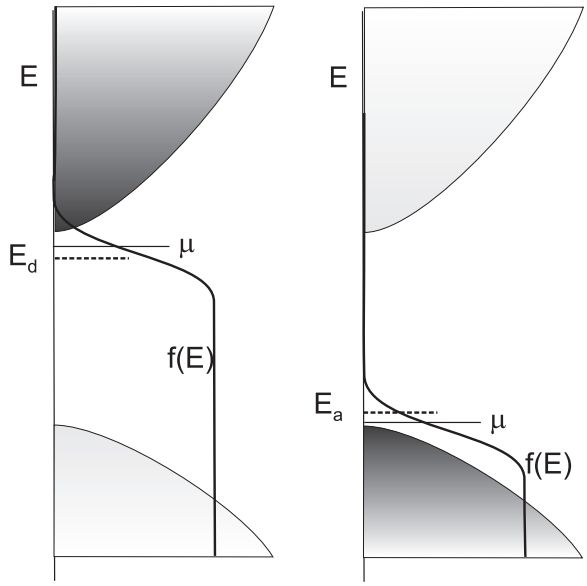


Figure 7.3: The left hand figure shows the effect of donor levels at an energy $E_d = E_c - \Delta_d$. The chemical potential will shift to near the conduction band edge, increasing the electron density and decreasing the hole density in comparison to the intrinsic case, while the product of the two is nevertheless constant. The right hand figure shows the corresponding picture for acceptor levels at an energy $E_a = E_v + \Delta_a$.

complicated. However, in many practical cases, both donor and acceptor states can be assumed to be fully ionised, so that the resulting carrier concentration in the extrinsic regime is given by the difference $N_d - N_a$.

Note that by the law of mass action, np is a constant at fixed temperature, given by (7.12). Doping with donor or acceptor atoms modifies the chemical potential and thereby shifts the balance between n and p, but the product remains constant. Even in the presence of strong donor doping, when the majority of the carriers are electrons, there will still be a small (minority) population of holes, given by $p = n_i p_i / N_d$.

Chapter 8

Semiconductor devices

We now consider the properties of inhomogeneous systems and devices. In this section we discuss the general properties of surfaces and interfaces between materials, and then the basic devices of bulk semiconductor physics. For bulk devices we will use the semiclassical approximation, treating electrons as classical particles governed by the Hamiltonian[1]

$$H = E_n(\mathbf{k}) - e\phi(\mathbf{r}) \tag{8.1}$$

with the momentum $\mathbf{p} = \hbar\mathbf{k}$ and a spatially varying electrostatic potential $\phi(\mathbf{r})$. The potential will arise from externally applied fields, from charges induced by doping, and from changes in the material composition. When we discuss narrow quantum wells, we shall need to modify this approximation to quantise the levels.

For an isolated solid in equilibrium, the energy difference between the chemical potential μ and the vacuum level is the work function Φ. This is the energy required to remove an electron from the Fermi level and place it in a state of zero kinetic energy in free space.

Two different isolated materials with different work functions will then have different chemical potentials. If these two materials are placed in contact, their chemical potentials must equalise, which is accomplished by electron flow to the more electronegative material; this material becomes charged, its potential ϕ changes, and an overall balance will be established. But in general there will be as a result internal inhomogeneous electric fields.

8.1 Metal - semiconductor contact

Fig. 8.1 is a schematic of this process for an ideal metal placed in contact with a semiconductor. We consider the more interesting case where the chemical potential of the (doped) semiconductor is above that of the metal.

8.1.1 Rectification by a metal contact

The barrier set up between the metal and insulator inhibits current flow. An electron from the metal must either tunnel through the barrier (at low temperatures) or may be thermally

[1] We shall keep to the convention that e is a positive number, and therefore $-e\phi$ is the potential energy of electrons in an electrostatic potential ϕ.

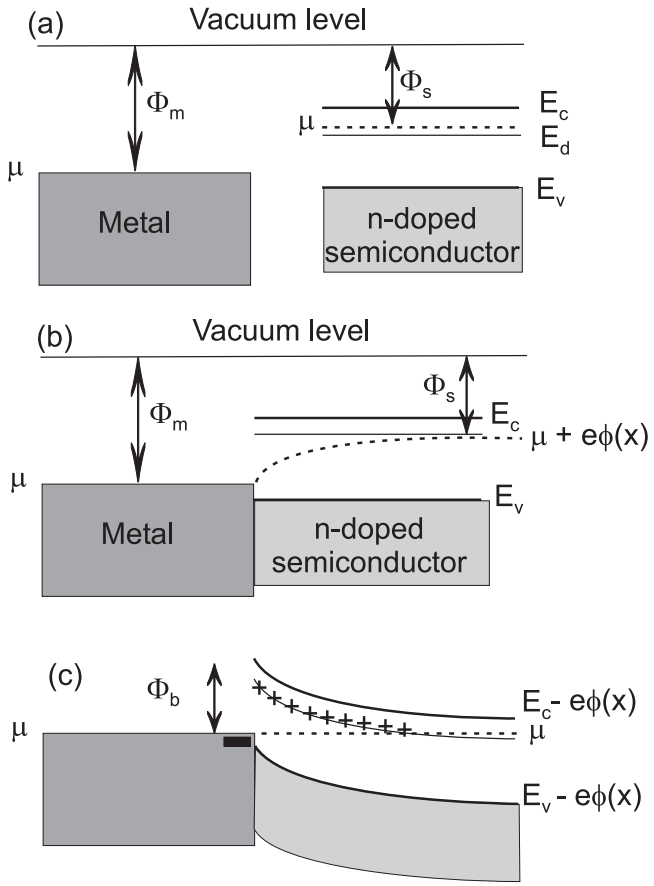


Figure 8.1: (a) When metal and semiconductor are not in contact they are in equilibrium with the vacuum level. We consider an n-type semiconductor, with the chemical potential lying close to the conduction band edge. (b). When the two are brought into contact, electrons leave the semiconductor and are transferred to the metal. This produces an electrical potential $\phi(x)$ which will eventually equilibrate so that the chemical potential is constant over the whole system. (The combined function $\mu+e\phi(x)$ is sometimes called the *electrochemical potential*.) (c) Shows the energy level diagram relative to the constant chemical potential. The semiconductor bands bend upwards, so that the donor levels near the interface are emptied of electrons - leaving a positively charged *depletion region*, and a *Schottky barrier* ϕ_b.

excited over it (thermionic emission). However, when a large enough external bias is applied, the junction may act as a rectifier Fig. 8.2. We will not analyse this in detail, as the more important case of a $p-n$ diode is similar, and will follow soon.

8.2 p-n junction

A p-n junction is formed by inhomogeneous doping: a layer of n-type material (containing donors) is placed next to p-type material (with acceptors). A schematic layout is shown in Fig. 8.3. The behaviour can be understood by an extension of the discussion for the metal-semiconductor contact.

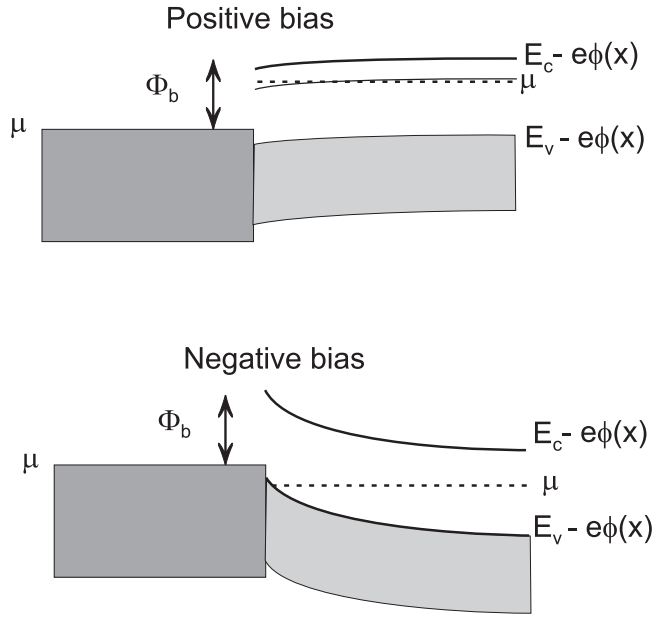


Figure 8.2: Schematic picture of a Schottky diode. In the upper panel, applying a positive bias across the junction lowers the barrier for electrons to enter the metal, and can eventually tilt the electron bands so much that the barrier disappears. The current grows rapidly with positive bias. However, if the bias is negative, the depletion width grows and the current is little changed.

- Deep inside the n-doped (p-doped) regimes, the chemical potential must lie close to the donor (acceptor) levels, and thereby also close just below the edge of the conduction band (just above the edge of the valence band).

- If we were to place the n-type and p-type regions in contact, charge would flow because of the different chemical potentials.

- In doing so, the interface region becomes depleted of carriers, and the ionised donors (acceptors) now have positive (negative) charge (see Fig. 8.4).

- The electrostatic potential so generated shifts the energy levels of the donors (acceptors) down (up) and the chemical potential is equalised

The typical extent of the depletion region is between 10 nm and 1 μm. See Fig. 8.5 for a summary of the physics of a p-n junction in equilibrium.

8.2.1 Rectification by a p-n junction

A p-n junction behaves as a *diode*, allowing current to flow much more readily in one direction that the other. A simple picture can be given as follows, with reference to the diagram in Fig. 8.6. Our sign convention is to apply an electrical bias where *positive* voltage V is applied to the *p-type* side of the junction.

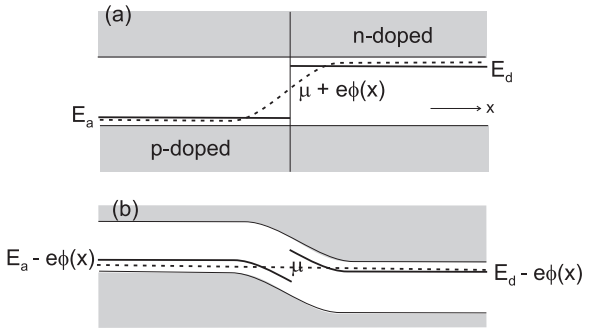


Figure 8.3: Two equivalent ways of representing the energy levels in a p-n junction. (a) shows the energy levels, and includes the electrostatic potential in the electrochemical potential $\mu + e\phi(x)$. In (b) we recognise that the chemical potential is constant, and the effect of the potential ϕ is a shift in the position of the energy levels $E_d(x) = E_d - e\phi(x)$, $E_a(x) = E_a - e\phi(x)$. When the shifted donor or acceptor levels pass through the chemical potential, these levels are *ionised*, and the carriers pass from one side of the barrier to the other, and annihilate. The impurity levels within the depletion region are now charged.

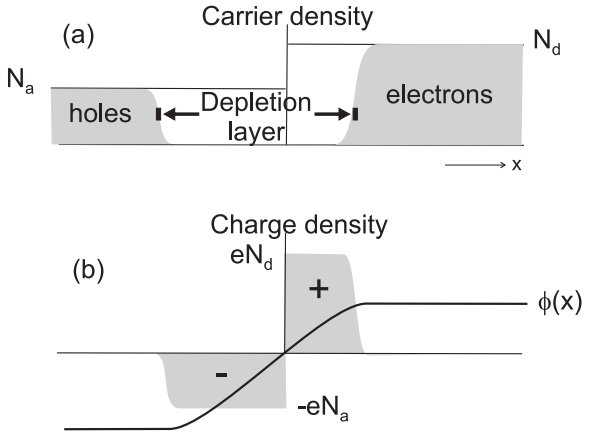


Figure 8.4: (a) Carrier densities and (b) charge densities near the depletion region of a p-n junction. When the temperature is low, the carrier density changes abruptly at the point where the chemical potential passes through the donor or acceptor level. Close to the barrier, the carriers are depleted, and here the system is now physically charged, with a charge density of $+eN_d$ on the n-type side, and $-eN_a$ on the p-type side. This dipole layer produces a potential $\phi(x)$ shown in (b). The potential itself self-consistently determines the charge flow and the width of the depletion region.

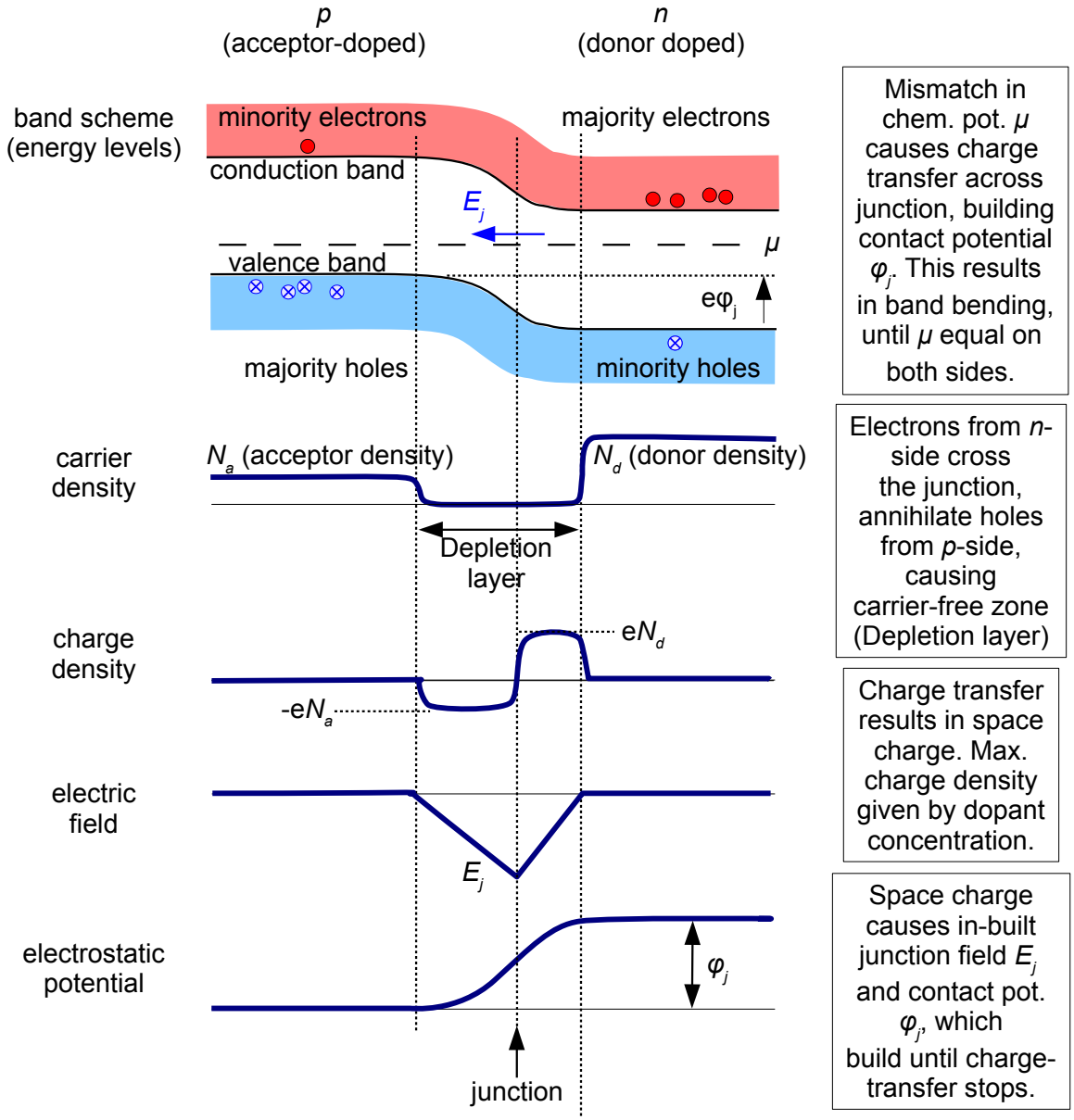


Figure 8.5: Overview of a p-n junction in equilibrium. Far away from the junction, the chemical potential μ must lie close to the bottom of the conduction band in the n-doped material, and close to the top of the valence band in the p-doped material. This is achieved by building up a contact potential ϕ, which shifts the energy levels as $E(z) = E_0 - e\phi(z)$. The change in potential across the junction is ϕ_j. It gives rise to an in-built field $\mathbf{E}_j$.

Potential barrier: The depletion regime of the junction is a high-resistance in comparison to the n- or p-type doped semiconductors. Any potential across the device is dropped almost entirely across the depletion layer. The overall potential seen by a (positively charged) hole is therefore $\phi_j - V$, where ϕ_j is the junction potential at equilibrium.

Balance of currents: In equilibrium with no external voltage bias, there is no net current flowing across the junction. We can, however, distinguish mechanisms which would drive currents across the barrier in both directions. In equilibrium, these currents cancel. We consider

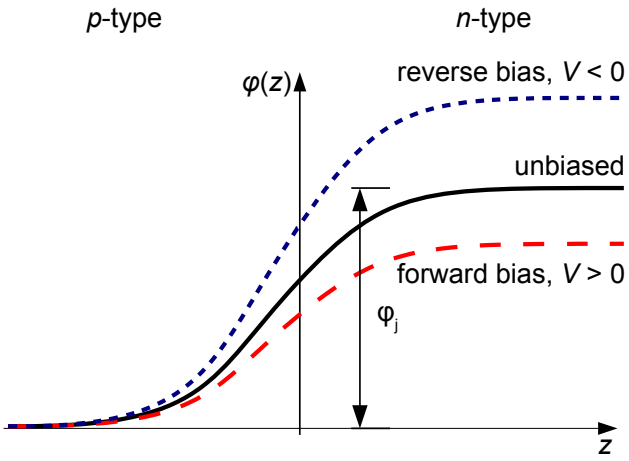


Figure 8.6: The effect of applying a bias voltage across a diode. The potential across the junction is decreased in forward bias (corresponding to a positive voltage V applied to the p-type side) and increased in reverse bias.

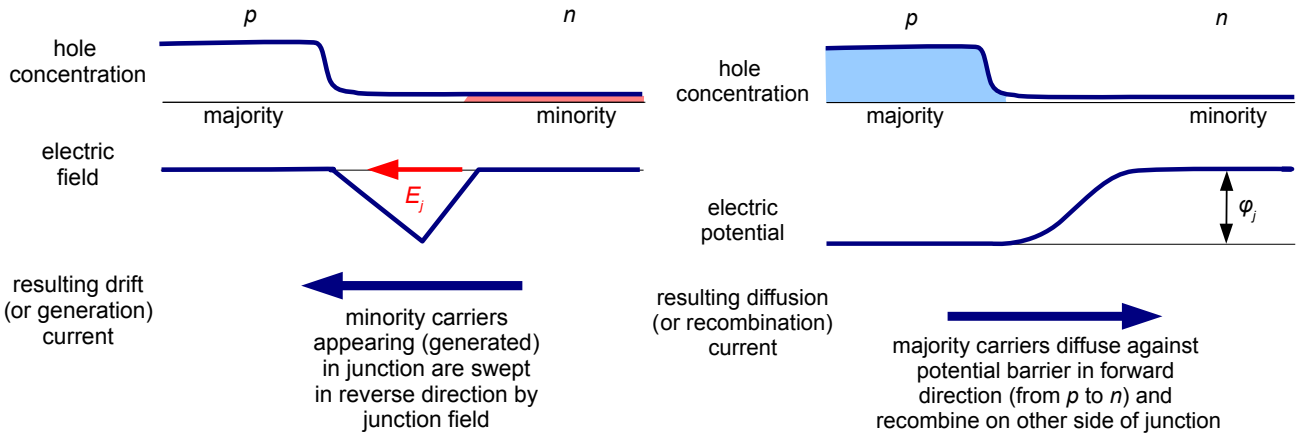


Figure 8.7: Left panel: Drift or generation current, illustrated with the example of holes: minority carriers, which are always present just outside of the depletion zone, because they are thermally *generated* to satisfy the law of mass action, are swept across the junction by the in-built field $\mathbf{E}_j$. The motion of charges in an applied field is called drift, which gives this current its other name. This current flows in the 'reverse' direction, from n to p, and depends only weakly on the bias voltage. Right panel: Diffusion or recombination current, here illustrated with the example of holes: majority carriers (holes on p-side, electrons on n-side) cross the junction to recombine with the oppositely charged majority carriers on the other side. This current flows in the 'forward' direction, from p to n. In forward bias, the recombination current (which is thermally activated) grows exponentially with bias.

them separately for holes and electrons, focussing first on the holes, because their positive charge simplifies matters slightly.

Drift or generation current: On the n-type side of the depletion region, the majority carriers are electrons, but detailed balance ((7.12)) means that there will always be some small density of *minority* holes. Any minority carrier wandering into the depletion regime will be swept into the p-type region by the in-built junction field (Fig. 8.7). This generates a current (from n to p, and therefore in the reverse direction)[2]

$$-J_h^{gen} \tag{8.2}$$

It does not depend strongly on the external bias V, because of the large inbuilt potential drop in the depletion regime.

Diffusion or recombination current: The holes in the p-type region, which are the majority carriers there, have a small probability of being thermally excited up the potential hill into the n-type region (Fig. 8.7). More strictly speaking, we need the number of holes with energy at least $e\phi_j$ from the band edge, because these will find states with equal energy on the other side of the junction. Since the temperature is low compared to the height of the junction potential, this current is activated, and in the presence of a bias voltage V takes the form

$$J_h^{rec} \propto e^{-e(\phi_b - V)/k_B T} \quad . \tag{8.3}$$

Net current: We know that in equilibrium at zero bias the hole recombination current and generation currents must cancel. The total hole current then takes the form

$$h = J_h^{gen}(e^{eV/k_B T} - 1) \quad . \tag{8.4}$$

Electrons. The same analysis applies to electrons, except that the corresponding electron generation and recombination (number) currents flow in the opposite directions to their hole counterparts. But electrons are oppositely charged, so the electrical current density has the *same* form as (8.4).

Diode IV characteristic. The sum of the contributions of electrons and holes gives an asymmmetrical form

$$I = I_{sat}\left(e^{eV/k_B T} - 1\right) \tag{8.5}$$

where the *saturation current* I_{sat} is proportional to n_i^2 and therefore of the Arrhenius form $e^{-E_g/k_B T}$ (see footnote 2).

8.3 Solar cell

If light shines on a p-n junction, without an external bias voltage, then each absorbed photon will create an electron-hole pair (Fig. 8.9). If these carriers reach the junction, the built-in field will separate them - the potential gradient pulls electrons and holes in opposite directions. The

[2] The magnitude can be estimated to be $(n_i^2/N_d)(L_p/\tau_p)$, where the first factor in brackets is the minority hole density in the n-type region (7.13) τ_p is the recombination time of a carrier, and L_p is the length that the hole will diffuse before it recombines with an electron.

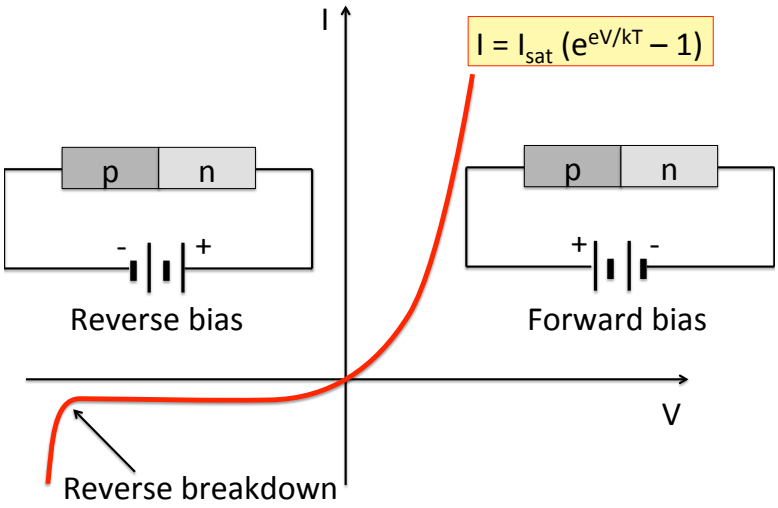


Figure 8.8: Theoretical I-V characteristic from (8.5).

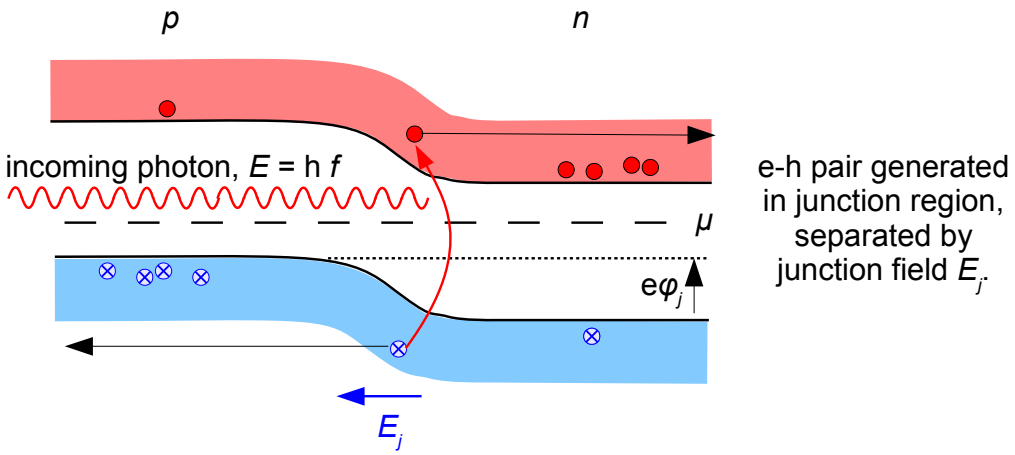


Figure 8.9: Operation of a p-n junction based solar cell. When illuminated, electron-hole pairs are generated. Pairs generated away from the junction will recombine rapidly, but those electrons and holes generated near the junction will be separated by the in-built electric field. Electrons flow towards the n-side, holes towards the p-side, which is equivalent to increasing the generation current, which flows in the reverse direction (n to p).

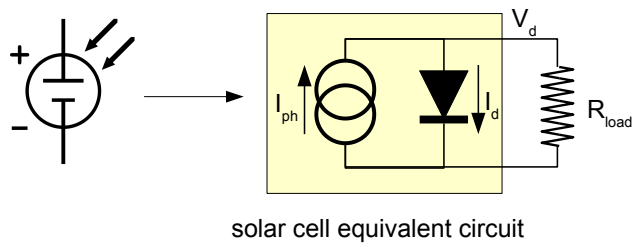


Figure 8.10: Equivalent circuit for a solar cell. We can model a solar cell as a p-n diode with a current source in parallel, which produces the photocurrent.

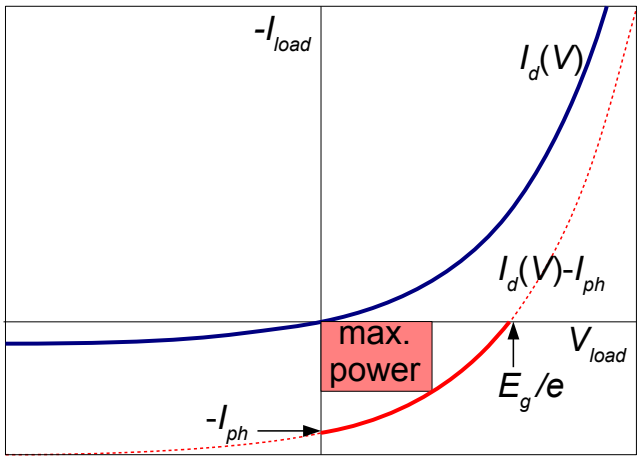


Figure 8.11: Current-voltage characteristic of a solar cell. The amount of power that can be extracted is given by the product of the voltage and current. These are limited by the size of the band gap and by the photocurrent.

resulting current is in the same direction as the generation current mentioned above, namely from n to p, or in the reverse direction. The separation of the charges across the depletion layer adds an extra internal dipole to the system - like charging a capacitor - and therefore generates an overall electrical bias. The induced voltage is in the forward direction - because it is opposite in sign to the built-in potential.

This is the *photovoltaic effect*, which can deliver power to an external circuit. Large arrays of p-n junctions of Si are used to make solar panels, converting solar radiation to electrical power. We can model the effect of this process as a current source added in parallel to the diode normally associated with a p-n junction (Fig. 8.10). The current delivered by this current source, the photocurrent I_{ph} depends on the amount of light falling onto the junction area. Fig. 8.11 illustrates how much power can be extracted from a solar cell, by considering the $I-V$ characteristic of the full device – current source plus diode. Note that for zero load resistance (short circuit, $I_{load} = I_{ph}$ but $V = 0$) and for infinite load resistance ($I_{load} = 0$), no power is extracted. What is the open circuit V_d? Its upper limit is given by the band gap E_g, because if V_d exceeds ϕ_j ($\sim E_g/e$), then the in-built junction field vanishes and photo-generated carriers are no longer swept out of the junction area. The maximum power extracted for the ideally chosen load resistance is therefore determined, up to a factor less than but of order unity, by the product $I_{ph}E_g/e$.

Shockley-Queisser limit: How far is it possible to optimise solar cells by tuning the gap energy E_g? Shockley and Queisser made a careful analysis of the maximum efficiency of solar cells, which considered, among other influences, the matching between the semiconductor band gap and the intensity spectrum of sunlight.

Their considerations are outlined in Fig. 8.12. The key point is that photons can only be captured by the solar cell, if the band gap is lower than the photon energy. On the other hand, the power extracted is given by the size of the band gap, not by the photon energy. This is evident from the preceding argument that the open circuit voltage is related to the band gap, but also from the fact that electrons and holes excited into states far away from their respective band edges will rapidly decay to lower-lying states near the band edge, well before they can

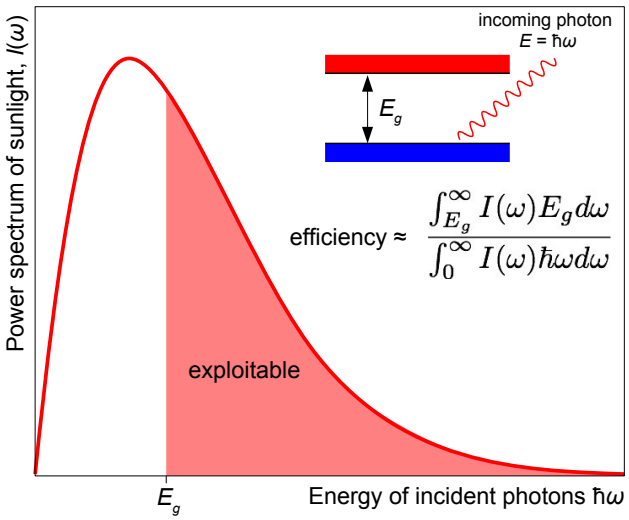


Figure 8.12: Fundamental limit to the efficiency of solar cells (Shockley-Queisser limit). The ratio of extracted power over total incident solar power is determined by the band gap of the semiconductor used to build the solar cell. Large band gaps lead to large open circuit voltages, but reduce the fraction of the sun's spectrum that can be exploited. Small band gaps allow more sun-light to be used, but limit the open circuit voltage and thereby reduce the extracted power.

leave the device. There is therefore a trade-off between (i) having a small band gap in order to capture as much as possible of the light, but at the penalty of only achieving a small open circuit voltage and thereby harvesting little energy per photon, and (ii) having a large band gap in order to extract as much energy as possible from each captured photon, but at the penalty of capturing only few photons.

The ratio between the energy extracted from the sun-light, $\int_{E_g}^{\infty} I(\omega)E_g d\omega$, and the total energy incident on the device, $\int_0^{\infty} I(\omega)\hbar\omega d\omega$, where $I(\omega)$ denotes the spectral intensity at angular frequency ω can be optimised as a function of the gap energy E_g. When combined with other fundamental limitations, the optimum efficiency of a single-junction solar cell is near 33% for a band gap of around 1.2 eV. Modern solar cells achieve efficiencies of about 22%, which is already quite close to this theoretical limit. This limit may be overcome, at least in principle, by more advanced designs which combine several junctions with different band gaps.

8.3.1 Light-emitting diode

The inverse process to the photovoltaic effect powers light-emitting diodes or LEDs Fig. 8.13. Here, a current is injected into a p-n diode in a non-equilibrium situation where the electron and hole chemical potentials differ by a large bias potential eV. Electrons are injected from the n-side to the p-side of the junction, and holes in the reverse direction. Recombination of an electron-hole pair occurs with the emission of a photon, whose energy will be close to the band gap of the semiconductor.

This process is not efficient for an indirect band-gap semiconductor such as Si or Ge, and so direct gap III-V or II-VI materials are commonly employed. Using materials with wider

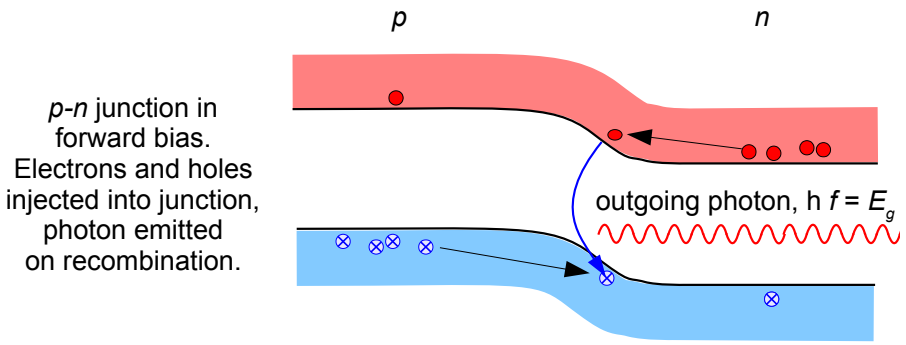


Figure 8.13: Principle of operation of a light emitting diode (LED).

band gaps gives higher frequency light. In recent years, efficient LEDs have been developed across the visible spectrum, and are now more efficient than incandescent bulbs. Recent LED developments include not only wide-gap inorganic materials but also organic materials. These can be processed in different and simpler ways from inorganic compound semiconductors, have a larger intrinsic radiative coupling, and are of course flexible.

8.4 Field effect transistor

Field effect transistors (FET) are the mainstay of the semiconductor industry. Their principle of operation is based on our ability to manipulate the carrier density in a channel between two electrodes via a controlling voltage applied to a third electrode. This controlling electrode is called the *gate*, and of the other two electrodes, the one at which the mobile carriers (usually the electrons) originates is called the *source*, whereas the one towards which the carriers move is called the *drain*.

A very readable account of the operation of FETs can be found at `http://www.freescale.com/files/rf_if/doc/app_note/AN211A.pdf`. We distinguish junction based FETs (JFET), which use the principles of p-n junctions to control the width of the conducting channel, and FETs in which the gate is separated from the rest of the device by an insulating layer, the metal-oxide semiconductor FET (MOSFET).

8.4.1 Junction field effect transistor: JFET

The operation of a JFET is based on the existence of depletion regions near the gate electrodes (Fig. 8.14). This makes it possible to vary the current between source and drain by changing the size of the conducting channel:

(a) Between two contacts on an n-type (donor-doped) semiconductor, called source and drain, the electrical conductivity would be high because of the high carrier density in the n-doped region.

(b) Adding p-type regions between the source and drain contacts and connecting them to gate electrodes (here, two gate electrodes, one is sufficient in principle), allows control

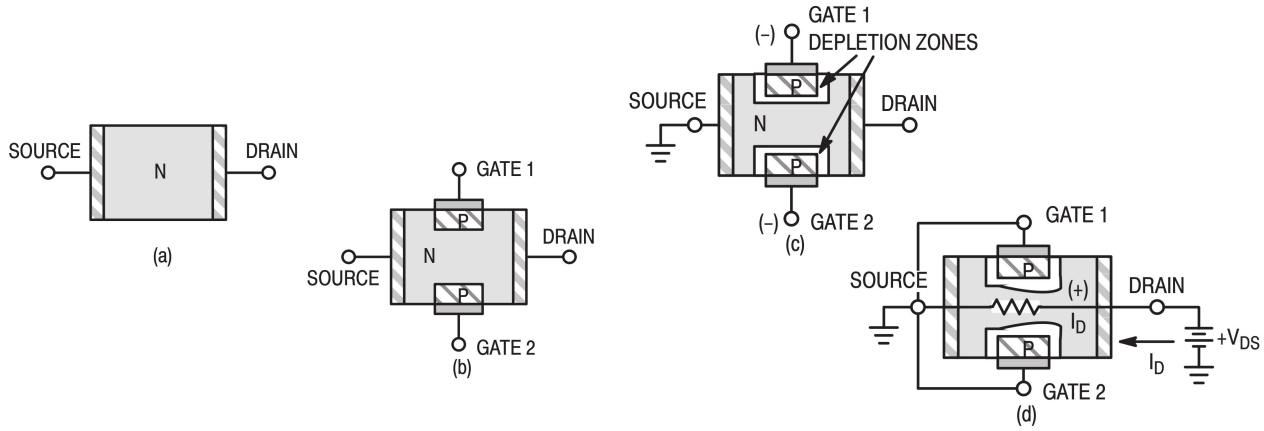


Figure 8.14: Principle of operation of a junction field effect transistor, or JFET: a) Metallic current contacts on n-doped semiconductor (source, drain). b) p-doped gate contacts in between source and drain. c) Depletion zones surround the p-n junctions near the gate electrodes. d) Biasing the gates changes the width of the depletion zones. This changes the width of the remaining current-carrying channel in between the gates. The figure also illustrates the phenomenon of **pinch-off** near the drain electrode (see text).

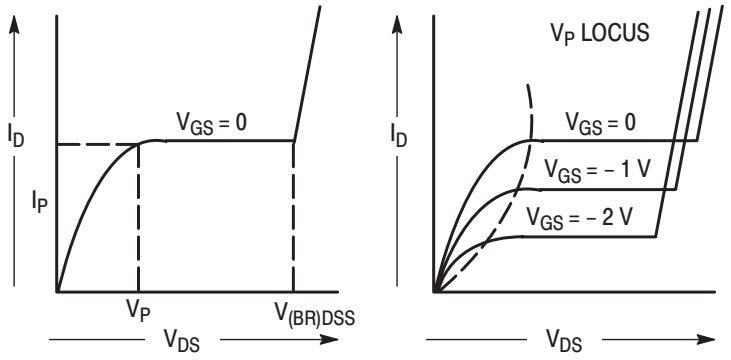


Figure 8.15: Current voltage characteristics of an n-type JFET with p-type gate. With increasing drain-source voltage V_{DS}, at constant gate-source voltage V_{GS}, the drain-source current I_D first rises roughly linearly. Pinch-off causes it to saturate at a level (I_P), which depends on V_{GS}.

over the current flow between the source and drain.

(c) At the junction between the p-type and n-type regions of the device, depletion zones form, as they would in a p-n junction diode (previous section). There are very few carriers in the depletion zone, and thereby the conducting cross-section of the channel between the source and the drain is reduced.

(d) By applying a voltage to the gate electrodes, the width of the depletion zone can be controlled, thereby altering the width of the conducting channel: a positive gate voltage reduces the size of the depletion zone, and increases the current in the conducting channel, whereas a negative voltage would widen the depletion zone and reduce the current further.

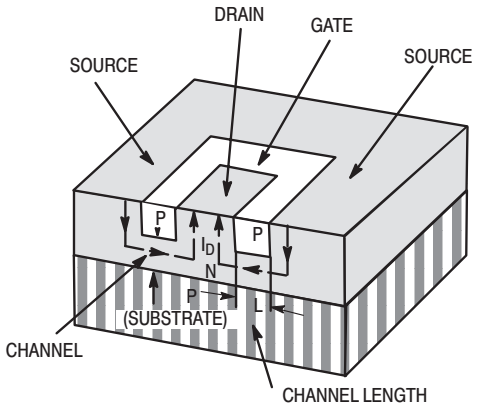


Figure 8.16: Fabrication of a JFET: doping an annular p-type region into an otherwise n-type bulk semiconductor crystal allows comparatively simple manufacture of a JFET.

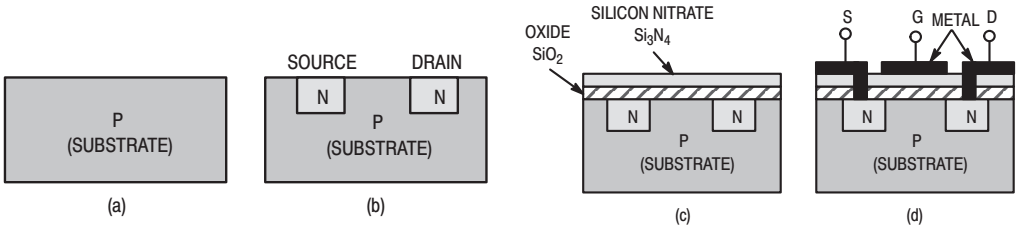


Figure 8.17: Manufacture of an enhancement-mode n-channel MOSFET: a) p-doped substrate, b) n-doped source and drain contacts. The depletion zones around these contacts produces very high intrinsic source-drain resistance. c) Insulating oxide layer (silicon nitride guards against sodium diffusing in). d) Metallic contacts to the source and drain are made through holes etched into the insulating layer, the metallic gate electrode is insulated from the substrate. Applying a positive voltage to the gate pulls electrons into the depletion zones and establishes a conducting channel between the source and drain.

Pinch-off: At finite drain-source current, the potential in the channel changes along the channel – it drops from drain to source. This causes the width of the depletion zone to change along the channel. The depletion zone is widest at the drain end, because there the potential of the n-type channel is highest, and so the voltage between a positively charged gate and the channel is lowest there. If we increase the drain-source voltage while keeping the gate potential constant, then the depletion zone will widen near the drain electrode, until it eventually covers most of the width of the semiconductor at that point, pinching off the conducting channel. Plotting the $I-V$ characteristic, drain-source current I_D as a function of drain-source voltage V_{DS}, at constant gate-source voltage V_{GS} (Fig. 8.15) therefore shows saturation of I_D at high V_{DS}.

Amplifier: The saturation of the drain-source current at a level depending on V_{GS} is the basis for operating the JFET as an amplifier: we can control the current through the device by changing the gate voltage, without having to worry about keeping the drain-source voltage exactly constant. Basically, the JFET acts like a voltage-controlled current source.

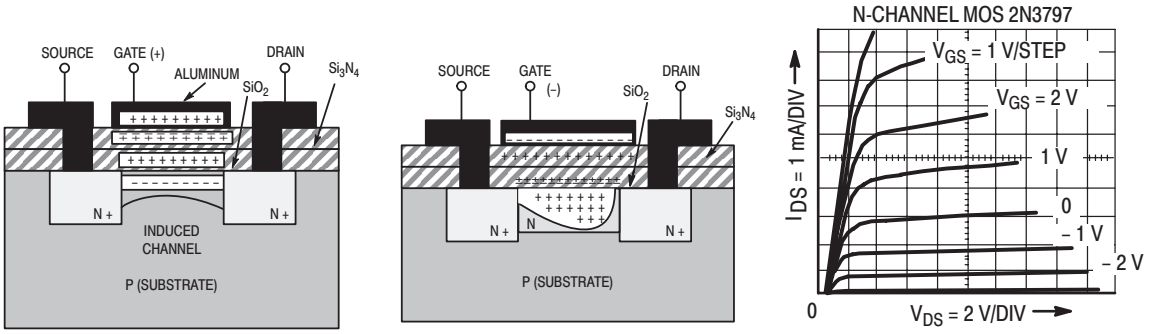


Figure 8.18: Principle of operation of MOSFET devices: applying a gate voltage redistributes minority carriers in the source-drain channel. Left: enhancement mode MOSFET: positive voltage pulls minority carriers (electrons) towards the surface, forming a high-conductivity channel, also called inversion layer. Centre: depletion-enhancement mode MOSFET: negative voltage depletes conducting channel, increasing the resistance, positive voltage reduces the resistance Right: typical I-V characteristic for a depletion-enhancement mode MOSFET. Note the pinch-off at high drain-source voltage.

8.4.2 Metal-oxide-semiconductor field effect transistor: MOSFET

In a MOSFET, the workhorse of modern electronics, the width of the conducting channel between the drain and source electrode is controlled by applying electric fields, using a nearby gate electrode which is electrically insulated from the rest of the device. This has the advantage that there is no current flow from the gate electrode (in the JFET, there will always be some small current across the depletion region), giving rise to extremely high input impedances. There are many ways for achieving this aim, one of which is shown in Fig. 8.17.

MOSFET operation: Although a variety of different MOSFET designs can be distinguished, their operation relies on the same two fundamental principles (Fig. 8.18). Firstly, by changing the gate voltage, depleted regions between the source and drain electrodes can be filled with carriers or, alternatively, conducting channels can be depleted. This allows us to vary the resistance of the drain-source channel. Secondly, as for the JFET, pinch-off occurs near the drain electrode, causing the drain-source current to saturate. This makes the device useful as an amplifier.

Inversion layer: The above discussion is still somewhat qualitative. It is at the same intuitive level as explaining the operation of a p-n junction diode by discussing the effect of forward or reverse bias on the width of the depletion zone. A more detailed explanation would consider the effect of an external potential on the electronic energy levels of the semiconducting material between source and drain (Fig. 8.19).

We see that a bias potential on the gate bends the band edges of the semiconductor beside the insulating barrier. If the bending is large enough, a p-type semiconductor — as shown here — can pull the conduction band below the chemical potential. This creates a narrow channel called an inversion layer, whose width can be controlled by the gate potential eV.

The width of the potential well formed at the oxide/semiconductor interface is often narrow enough that the levels within it are quantised. New and potentially useful effects arise from

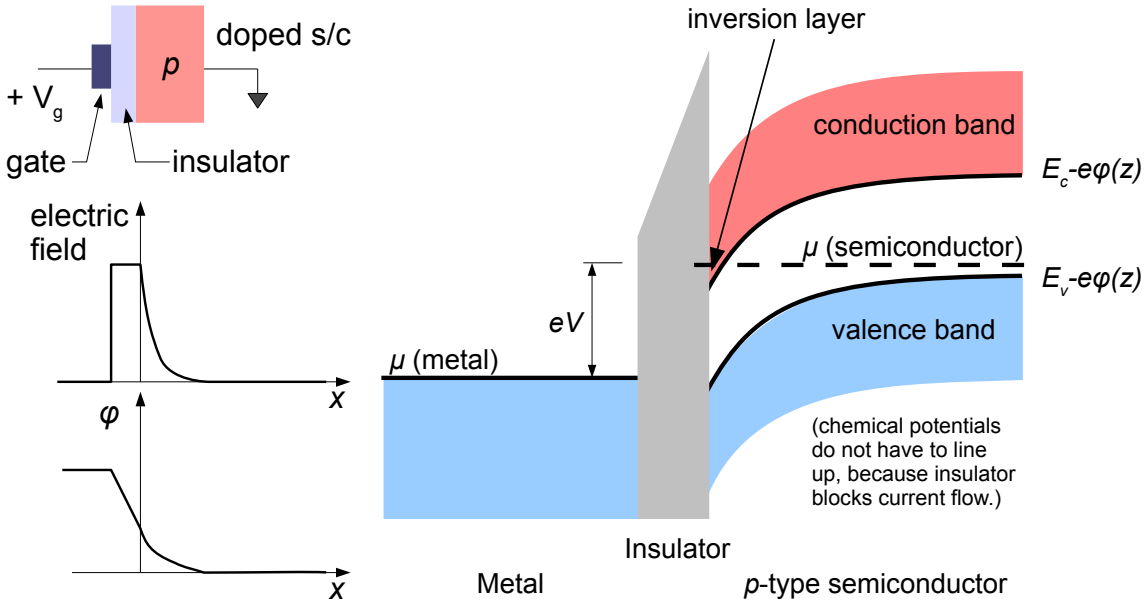


Figure 8.19: Band bending induces an inversion layer in a MOSFET: applying a positive voltage to the gate electrode creates an electric field across the insulating oxide layer, which penetrates some distance into the semiconductor. This causes a varying potential $\phi(x)$ close to the surface of the semiconductor. If the resulting band-bending at the semiconductor/oxide interface becomes larger than the energy gap E_g, then the conduction band edge falls below the chemical potential at the surface, causing an inversion layer to form.

the quantisation of energy levels in such *quantum well* structures (see below). Because the semiconductor in a MOSFET has to be doped in the region were the inversion layer can form, the electronic mean free path, and thereby the mobility, in the quantum well structure is small, which restricts its usefulness.

8.5 Compound semiconductor heterostructures

8.5.1 Bandstructure engineering

Another way to make an inversion layer is to change the semiconductor chemistry in a discontinuous fashion within the same crystal structure. Epitaxial, atomic layer-by-layer growth allows the chemical composition and doping to be manipulated in fine detail. Such devices made from compound semiconductors are used, for example, in semiconductor lasers for optical discs, in high speed electronics (e.g., cellphones) and high-speed lasers in telecommunications. This technology has also enabled fundamental science, by preparing very high mobility electron systems (e.g., for the quantum Hall effects), making "quantum wires" that are so thin as to have quantised levels, and for studies of the neutral electron-hole plasma as a possible superfluid.

Alloys of compound semiconductors, e.g., $Al_{1-x}Ga_xAs$, allow one to continuously vary the optical gap and the position of the band edges by varying the composition x. Two different semiconductors will - when referred to the vacuum potential at infinity - have bands that will in general line up in an offset fashion. We consider here only the case (like $(Al, Ga)As$) when

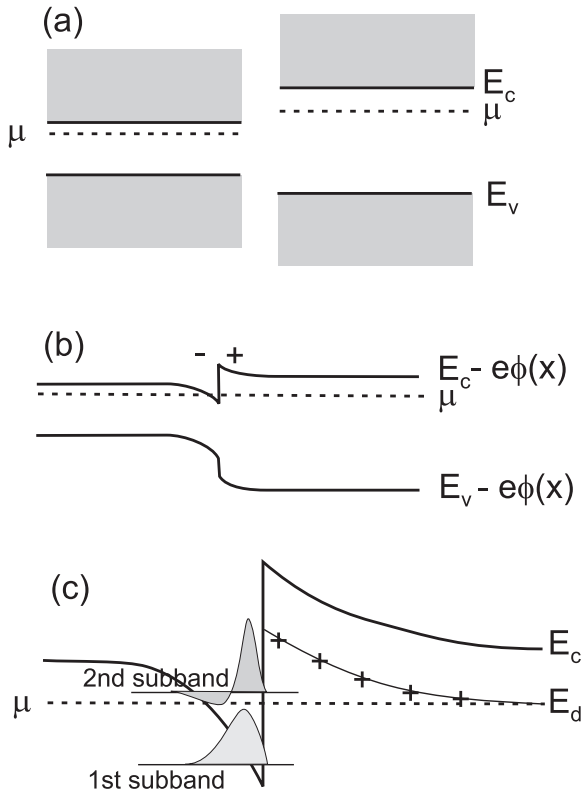


Figure 8.20: Formation of a 2D electron system by modulation doping. (a) Shows two semiconductors not in contact, with different chemical potentials, determined by the doping level. (b) is the band scheme that results when they are placed in contact. If the doping level is high enough - as shown here - the band edge on the left may fall below the chemical potential, and a layer of electrons is formed at the interface. (c) Shows the modulation doping scheme in more detail. Donors are placed to the right of the interface, so that the electron layer is pristine and free of impurities. The well width may be narrow enough that electron levels are quantised in a direction perpendicular to the barrier, forming sub-bands.

the band edges of one semiconductor lie entirely within the band gap of the other, though staggered overlaps do occur. When the materials are placed in contact, their Fermi energies must equalise, which is accomplished by charge transfer across the boundary. This lowers the conduction band edge on one side of the interface, and if doped sufficiently the band edge falls below the chemical potential, so that an equilibrium electron gas forms at the interface.

8.5.2 Inversion layers

Fig. 8.20 shows an outline of a scheme called modulation doping, where the donor levels are placed on the side of the interface away from the electron layers (and often at some distance from the interface). This has the advantage of creating an electron gas in a region where the crystal structure is nearly perfect, and mobilities greater than $10^3\ m^2V^{-1}s^{-1}$ have been achieved at low temperature. By addition of metal gates to the surface of the structures, electrical potential gradients can be applied to continuously vary the electron density in the layer, to pattern one dimensional wires, and to construct other interesting spatial structures.

8.5.3 Quantum wells

One of the most widespread applications of semiconductor multilayers is to make a *quantum well* — a thin region of a narrow gap material sandwiched inside a wide-gap one. Because the wells can be made very narrow, quantisation of the levels is important. In general, the eigenstates will be of the form $\Phi(\mathbf{r}, z) = \phi_n(z)e^{i\mathbf{k}\cdot\mathbf{r}}$ where $\mathbf{r}$ and $\mathbf{k}$ are here *two*-dimensional vectors, describing position and momentum in the plane. The situation for holes is more complex, because the degeneracy of the light and heavy hole states in bulk is broken by the 2D geometry. The details are important in practice.

8.5.4 Quantum well laser

The operation of a laser requires an efficient mechanism for luminescent electron-hole recombination, which rules out indirect gap semiconductors in practice. Lasing requires high densities of electrons and holes so that the probability of stimulated emission overcomes that of absorption. This latter condition requires *inversion*, meaning that the average electron (hole) occupancy in the luminescing states exceeds 1/2.

A double heterojunction laser is designed to achieve high densities, by using a quantum well — designed to trap both electrons and holes — with the source of carriers being a p-doped region on one side of the well, and an n-doped region on the other (see Fig. 8.21). This is indeed a diode (because holes can flow in from the p-side and electrons from the n side, but not vice versa), but it is not operated in the same regime as a conventional diode. Instead, a rapid rate of recombination in the lasing region maintains different chemical potentials for electron and hole systems.

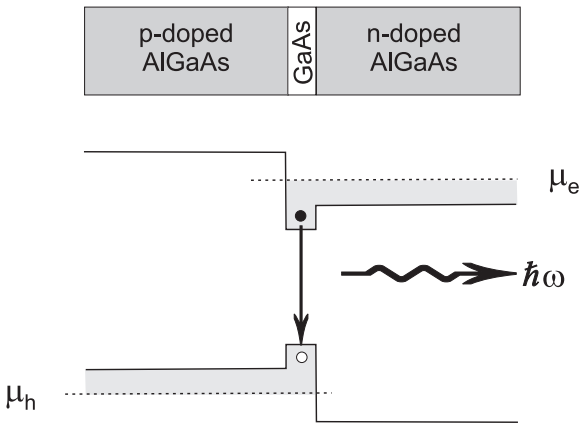


Figure 8.21: Operation of a double heterojunction laser. Notice the quasi-equilibrium condition, with separate electron and hole chemical potentials.

Chapter 9

Electronic instabilities

General remarks on theories and models in condensed matter physics

Solid state physics is concerned with the abundance of properties that arise when atoms are amalgamated together. Much of what we think of as "core physics" is deliberately reductionist; we look for the very simplest unified description of a basic phenomenon, and the progress of much of basic physics has always been a progress toward grander unified theories, each of which is simpler (at least in concept) than the previous generation.

Condensed matter physics is not like this. The Hamiltonian is not in doubt - it is the Schrödinger equation for the many particle system:

$$H_{elec} = -\sum_i \frac{\hbar^2}{2m}\nabla_i^2 + \sum_I \frac{P_I^2}{2M_I} + \frac{1}{4\pi\epsilon_0}\left(-\sum_{i,I}\frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{R}_I|} + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} + \frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J e^2}{|\mathbf{R}_i - \mathbf{R}_j|}\right), \tag{9.1}$$

where the $\mathbf{r}_i, \mathbf{R}_I$ label the coordinates of the electrons and the ions respectively, and Z_I, M_I are the nuclear charge and mass. The terms in (9.1) represent, in order, the kinetic energy of the electrons, the kinetic energy of the nuclei, and the Coulomb interaction between electron and nucleus, electron and electron, and between nucleus and nucleus. In some sense, a complete theory of solids would be to solve the Schrödinger equation and then apply all the standard methods of statistical physics to determine thermodynamic and physical properties. From this point of view, there is no "fundamental" theory to be done, although the calculations may indeed be complex (and in fact, impossible to perform accurately for solids with macroscopic numbers of atoms). Because an accurate solution for a macroscopic number of atoms is impossible, we have to treat (9.1) using a sequence of approximations (for example, perhaps fixing the ions in place, or neglecting electron-electron interactions) that will make the problem tractable.

This view of condensed matter physics as a series of approximations is widely held, but severely incomplete. Suppose for a moment that we could solve the full Hamiltonian, and we would then have a wavefunction describing some 10^{23} particles that contained all of the physics of solids.

Writing the solution down would be hard enough, but comprehending its meaning would be beyond us. Condensed matter physics is about phenomena, from the mundane (why is glass

transparent), to the exotic (why does ^{3}He become a superfluid). There are a host of physical phenomena to be understood, and their explanation must involve more than just detailed calculation.

Understanding a phenomenon involves building the simplest possible model that explains it, but the models are more than just approximations to (9.1). Models, and the theories that they give rise to, elucidate paradigms and develop concepts that are obscured by the complexity of the full Hamiltonian. The surprise about condensed matter physics is that there are so many *different* theories that can arise from the Hamiltonian (9.1).

"The Properties of Matter"

A venerable route to condensed matter physics, and one followed by almost all textbooks, is to find ways of performing approximate calculations based on the full Schrödinger equation for the solid. Performing approximate, but quantitative calculations of the physical properties of solids has been one of the enduring agendas of condensed matter physics and the methods have acquired increasing sophistication over the years. We would like to understand the cohesion of solids – why it is, for example that mercury is a liquid at room temperature, while tungsten is refractory. We wish to understand electrical and optical properties – why graphite is a soft semi-metal but diamond a hard insulator, and why GaAs is suitable for making a semiconductor laser, but Si is not. Why is it that some materials are ferromagnetic, and indeed why is it that transition metals are often magnetic but simple *sp* bonded metals never? We would like to understand chemical trends in different classes of materials – how properties vary smoothly (or not) across the periodic table.

Nowadays we can use sophisticated computational techniques to calculate physical and chemical properties of systems, sometimes with quite high accuracy. The computations are, however, complicated and produce many numbers. Such computations provide invaluable assistance in modelling condensed matter, but they provide limited insights into the types of behaviour that we can expect and the actual behaviour that is observed. Perhaps surprisingly, the types of behaviour that might be observed may often be understood in terms of simplified models. These models must incorporate the basic machinery of the quantum mechanics of periodic structures, especially the concept of electronic bandstructure describing the dispersion relation between the electron energy and momentum. We also need to understand how the effects of strong interactions between electrons can be subsumed into averaged effective interactions between independent quasiparticles and the background medium. The aim is to generate a landscape upon which models and theories can be built and understood.

Collective phenomena and emergent properties

There is a complementary view of condensed matter physics which we shall explore, that is less concerned with calculation and more concerned with phenomena per se. The distinguishing characteristic of solid state systems is that they exhibit *collective* phenomena. These are properties of macroscopic systems that exist because the systems have many interacting degrees of freedom. A familiar example is a phase transition (between liquid and solid, say) which is a concept that can only apply to a macroscopic ensemble. We are so used to phase transitions that we rarely wonder why when water is cooled down it does not just get "thicker" and more

viscous (which is actually what happens to a glass).

Condensed matter systems have collective modes that are a consequence of their order; both a solid and a liquid support longitudinal sound waves, but a solid that has a nonzero shear stiffness also has transverse sound modes. In fact the existence of shear waves could be defined as the characteristic feature distinguishing a solid from a liquid or gas. We can say that solidity is a *broken symmetry* (with the broken symmetry being that of translational invariance); because of the broken symmetry, there is a new collective mode (the shear wave). Because of quantum mechanics, the waves are necessarily quantised as phonons, and they are true quantum particles, with Bose statistics, that interact with each other (due to anharmonicity) and also with other excitations in the solid. This idea, that a broken symmetry can generate new particles, is one of the central notions of condensed matter physics – and of course of particle physics too.

A different example is the behaviour of electrons in a semiconductor. If one adds an electron to the conduction band of a semiconductor it behaves like a particle of charge $-|e|$, but a mass different from the free electron mass due to the interactions with the lattice of positively charged ions as well as all the other electrons in the solid. But we know that if we remove an electron from the valence band of the semiconductor it acts as a *hole* of charge $+|e|$; the hole is in fact a collective excitation of the remaining 10^{23} or so electrons in the valence band, but it is a much more convenient and description to think of it as a new fermionic *quasi*-particle as an excitation about the ground state of the solid. The electrons and holes, being oppositely charged, can bind together to form an exciton - the analog of the hydrogen atom (or more directly positronium), which however has a binding energy considerably lower than hydrogen, because the Coulomb interaction is screened by the dielectric constant of the solid, and the electron and hole masses are different from the electron and proton masses in free space.

The solid is a new "vacuum", inhabited by quantum particles with properties which may be renormalised from those in free space (e.g., photons, electrons) or may be entirely new, as in the case of phonons, plasmons (longitudinal charge oscillations), magnons (waves of spin excitation in a magnet), etc. In contrast to the physical vacuum, there are different classes of condensed matter systems that have different kinds of vacua, and different kinds of excitations. Many of these new excitations arise because of some "broken" symmetry , for example, magnetism implies the existence of spin waves, and solidity implies the existence of shear waves. Some of the phenomena that come to mind; superconductivity, superfluidity, and the quantum Hall effect, are remarkable and hardly intuitive. They were discovered by experiment; it seems unlikely that they would ever have been uncovered by an exercise of pure cerebration starting with the Schrödinger equation for 10^{23} particles.

Solid state systems consist of a hierarchy of processes, moving from high energy to low; an energy scale of electron volts per atom determines the cohesive energy of a solid, the crystal structure, whether the material is transparent or not to visible light, whether the electrons are (locally) magnetically polarised, and so on. But after this basic landscape is determined, many further phenomena develop on energy scales measured in meV corresponding to thermal energies at room temperature and below. The energy scales that determine magnetism, superconductivity, etc., are usually several orders of magnitude smaller than cohesive energies, and the accuracy required of an *ab initio* calculation which described such phenomena cannot be achieved. Although all condensed matter phenomena are to be found within the Schrödinger equation, they are not transparently derived from it, and it is better to start with specific models that incorporate the key physics; we shall see some of them in this course. The mod-

els will mostly be of interactions between excitations of the solid, with accompanying sets of parameters parameters that are usually estimated, or derived from experiment.

9.1 Charge Density Waves

The crystal structures of solids are much more complex than one might have expected. Even if you take the elements, rather few form simple close-packed structures. For example Ga metal has a complicated structure with 5 nearest neighbours, Se crystallises in a structure that can be thought of as an array of spiral chains with three atoms per unit cell, As, Sb and Bi have puckered sheets where each atom has three near neighbours.

Of course, all of this reflects the production of chemical bonds inside the solid, and a complicated balance of forces. But the fundamental principle of bonding is that by placing the chemical potential in a gap, the occupied states are lowered in energy (and the unoccupied states go up in energy). Getting the chemical potential to lie in a gap involves making sure that the Brillouin zone boundary lies "in the right place", i.e. at a momentum that will contain exactly the correct number of states to account for all of the electrons in the solid.

9.1.1 The Peierls transition

As a concrete example we consider a one-dimensional chain of atoms, with lattice constant a, and an electron density chosen such that the Fermi wave-vector k_F falls somewhere in the middle of the band. It is a metal.

Notice that we could turn this metal into an insulator by applying an *external* potential with a periodicity of $2\pi/Q$ where $Q = 2k_F$: following the earlier lectures, we know that a periodic potential $V_o \cos(Qx)$ produces Bragg scattering at a wavevector $Q/2$ (hence a new Brillouin zone boundary). If $Q/2 = k_F$ then there is a gap induced on the fermi surface.

Rather than applying an external potential, we could get the same effect by making a periodic lattice distortion (PLD) with the same periodicity: namely move the n^{th} atom in the chain to a new position

$$R_n = na + u_o cos(Qna) \quad . \tag{9.2}$$

We assume that the amplitude of the PLD is small, $u_o \ll a$. [We have already met this phenomenon in the diatomic chain, studied earlier, but where we considered the case of a half-filled band — in that case $k_F = \pi/2a$ and $Q = \pi/a$, and the periodicity of the distorted lattice is twice that of the undistorted one.]

If the atoms have a PLD with period $2\pi/Q$, they will produce a new potential seen by the electrons with the same period. It is also evident that the amplitude of the Fourier component $V_Q \propto u_o$ is linearly proportional to the displacement (for small displacements). We may write $V_Q = g_Q u_o$, where the coefficient g_Q is the electron-phonon coupling constant.

Now remember that the energy gap on the zone boundary is $|V_Q|$. That means that an energy level at a momentum just below k_F is lowered by an energy proportional to the atomic displacement $|u_o|$, (and the unoccupied one just above k_F is raised in energy by the same amount). So overall there is an energy lowering as a result of the PLD. The magnitude of this can be computed (see the problem on sheet 4) by adding up all the energy changes of all occupied states: the answer can be written as

$$E_{electronic} = A(u_o/a)^2 \ln|u_o/a| \tag{9.3}$$

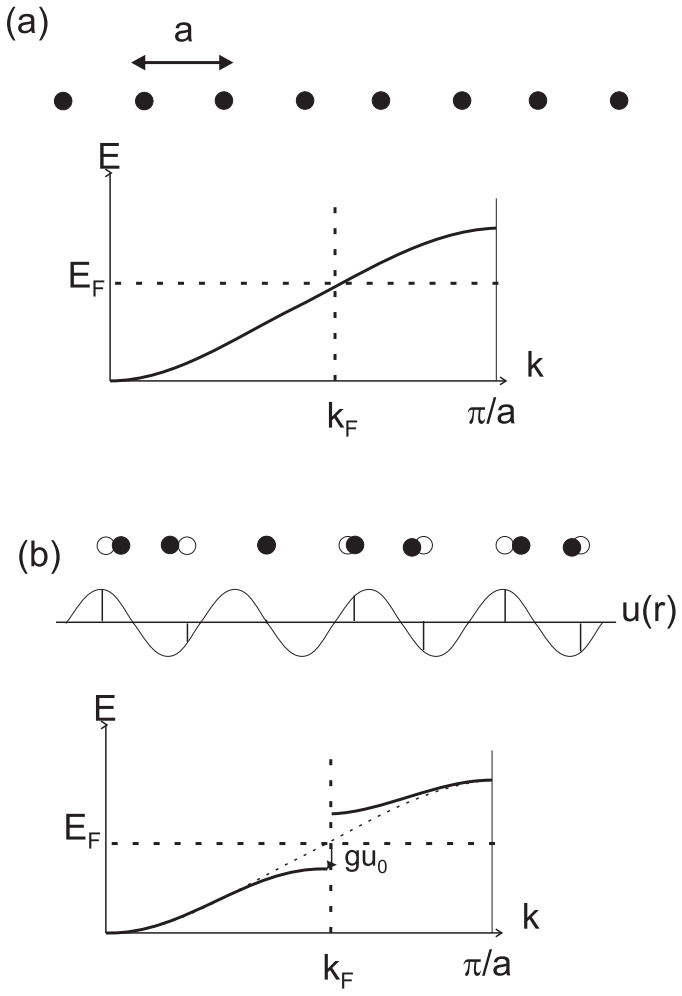


Figure 9.1: The Peierls transition. The upper figure shows the familiar one-dimensional chain with lattice constant a and the corresponding lowest electronic band, plotted for momenta between 0 and π/a. In the lower figure (b) a periodic lattice modulation is introduced, with u(r) of the form of (9.2). The period is cunningly chosen to be exactly $2\pi/2k_F$, so that a band gap of amplitude $2g_Q u_0$ is introduced exactly at the chemical potential.

in the limit $u_o/a \ll 1$, and A is a constant (depending on g_Q). Note the logarithm — this varies faster than quadratically (just). It is negative - the energy goes down with the distortion.

By an extension of the standard band structure result, it should be clear that there is an electronic charge modulation accompanying the periodic lattice distortion - this is usually called a charge density wave (CDW).

(9.3) is just the electronic contribution to the energy from those states very close to the fermi surface. But as we have argued before, it is sensible to model the other interactions between atoms just as springs, in which case we should add an elastic energy that is of the form

$$E_{elastic} = K(u_o/a)^2 \tag{9.4}$$

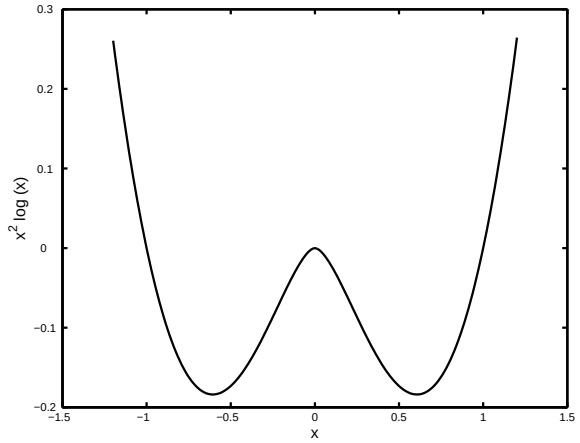


Figure 9.2: Sketch of (9.5) showing the double minimum of energy

Adding the two terms together gives a potential of the form

$$E(x) = Ax^2 \ln|x| + Bx^2 \tag{9.5}$$

which *always* has a minimum at non-zero displacement. The system lowers its energy by distorting to produce a PLD and accompanying CDW, with a period that is determined by the fermi wave-vector, viz. $2\pi/2k_F$. Such a spontaneous lattice distortion is a broken symmetry phase transition (see Fig. 9.2), that goes by the name of its discoverer, Peierls. It tells us that a one-dimensional metal is always unstable to the formation of a CDW, even if the electron-phonon coupling is weak. [1]

Materials that are strongly anisotropic in their electronic structure are thus prone to a spontaneous lattice distortion and accompanying charge density wave. (The logarithmic singularity does not appear in dimensions greater than one — although CDW's indeed happen in higher dimension, they don't necessarily occur in weak coupling).

Commonly there will be a phase transition on lowering the temperature that corresponds to the onset of order — one can monitor this by the appearance of new Bragg peaks in the crystal structure, seen by electron, neutron, or X-ray scattering (see Fig. 9.3).

More subtly, the onset of a CDW can be seen in the phonon spectrum. Notice that by calculating the energy change as a result of a small lattice displacement, we have in the coefficient of the quadratic term in the energy as a function of displacement, the phonon stiffness for a mode of the wavevector $2k_F$. Consequently, the onset of a CDW is when the stiffness becomes *zero* (and negative below the transition), so there is no restoring force associated with the displacement. Then the phonon spectrum $\omega(q)$ (even in the high temperature undistorted phase) will be expected to show a sharp minimum in the vicinity of $q = 2k_F$, as seen in Fig. 9.4.

[1] There are of course other periodicities produced by beating of the spatial frequencies Q with $2\pi/a$. These need not concern us if the amplitude is small, because they will generally occur at momenta different from k_F, so the gap will lower and raise the energy of pairs of states that are either both unoccupied or both occupied, cancelling in the total energy.

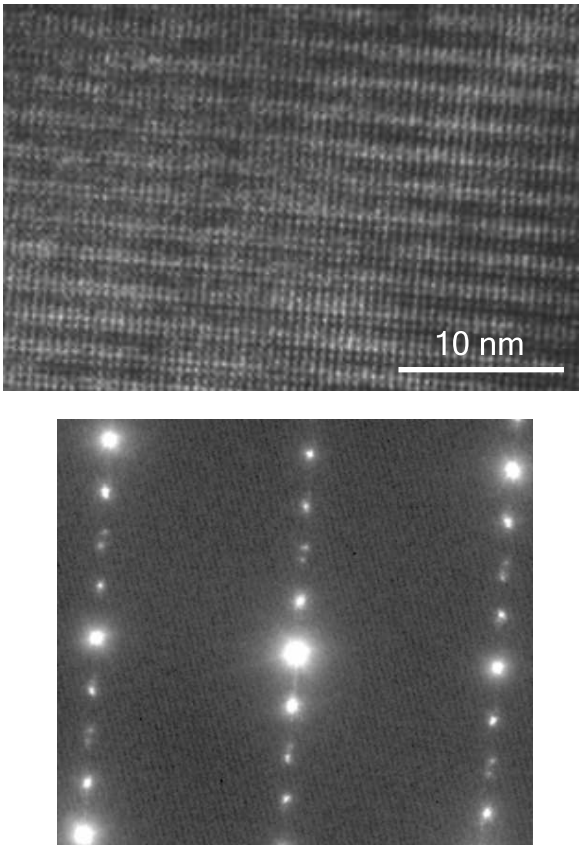


Figure 9.3: Electron diffraction image shows Bragg scattering from a CDW formed in the compound $La_{0.29}Ca_{0.71}MnO_3$. The real space image is at the top, showing a short scale checkerboard that is the atomic lattice, with a periodic modulation. The bottom figure is the fourier transform, so that the widely spaced bright peaks come from the small unit cell (this compound is based on a cubic perovskite, where the Mn atoms are on a simple cubic lattice), and the less intense peaks in between are the CDW. The two periods (lattice and CDW) are not related, because the CDW period is determined by the fermi surface size and shape, which depends on the electron concentration. Here the presence of an incommensurate ratio of trivalent La to divalent Ca means that the Mn d-bands are only partially filled. [Image courtesy of J. Loudon, P.A. Midgley, N.D. Mathur]

9.1.2 Polyacetylene and solitons

One of the celebrated cases of such a CDW occurs in the polymer $(CH)^n$, *trans*-polyacetylene (Fig. 9.5). Of the 4 valence electrons contributed by each carbon atom, one is involved in a bonding band (non-dispersive) with the H, leaving 3 electrons per atom to be accommodated along the $-C-C-C-$ chain. If the C atoms were equally spaced, then there would be one full and one *half-filled* band. This half-filled band is unstable to dimerisation by the Peierls mechanism — doubling the lattice period, halves the Brillouin zone. It is often idealised as an alternation of double and single bonds, viz. $-C=C-C=C-$.

The figure Fig. 9.5 shows that there are two different but symmetry-related ground states that can be formed by the dimerisation. One can readily imagine that in a long chain, these two states might join up next to each other, and that situation is visualised in Fig. 9.6. The

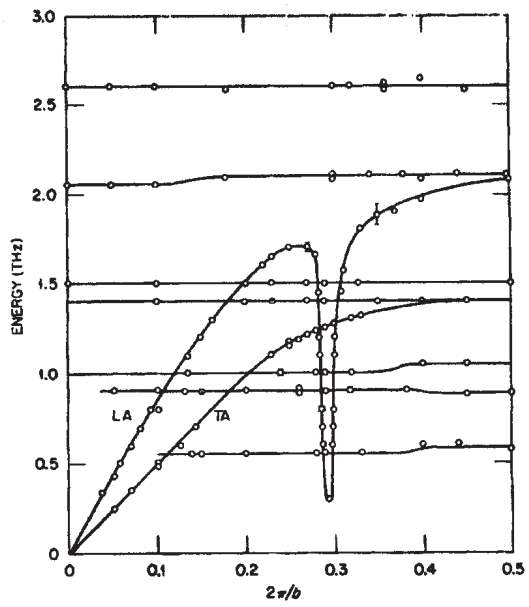


Figure 9.4: Phonon dispersion curves in the quasi- one-dimensional organic compound TTF-TCNQ (tetrathiofulvalene tetracyanoquinone) along the direction of the chains in which there is a prominent soft phonon that turns into a periodic lattice distortion at low temperature. (There are many non-dispersing optical modes in the complicated unit cell.)[Mook and Watson, Phys. Rev. Lett. **36**, 801 (1976)]

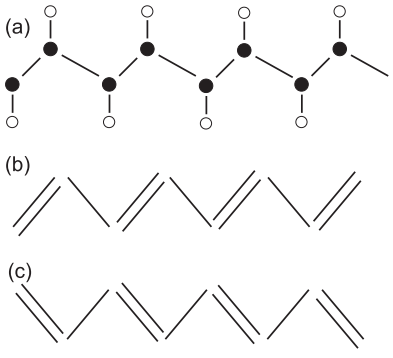


Figure 9.5: (a) shows a sketch of the atomic arrangement of polyacetylene, with the C atoms as solid circles, and the H atoms as open circles. The C atoms are not equally spaced, and the structure is often idealised as an alternation of "double" and "single" bonds, (b) and (c). The two isomers in (b) and (c) are related by a mirror symmetry.

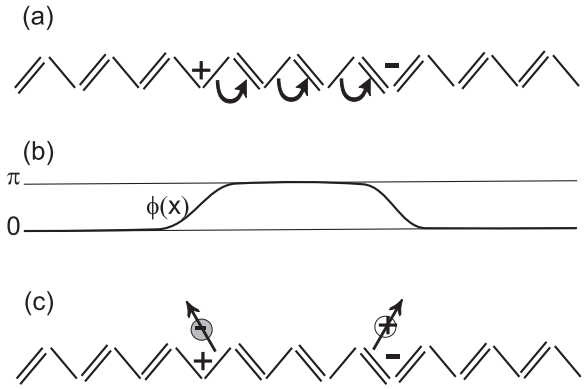


Figure 9.6: In (a) we create a region of dimerisation of the opposite sign in the middle of another domain. The boundaries between the regions are "solitons" (that are in practice several lattice constants wide), and one can see from the schematic arrows that they are charged. (b) These are often described using a phase variable $\phi(\mathbf{r})$, which describes the position of the charge density wave (see Eq. 9.6). The solitons act as potentials that can trap either electrons or hole (c), producing new kinds of quasiparticles that have the spin of the carrier but are electrically neutral.

boundary between the two regions cannot be locally "unwound", because a large number of atoms will have to be displaced to do so. There is a topological distinction between the two states.

9.1.3 Alignment of the charge density wave

The *solitons* that form the boundaries are similar in character to a domain wall in a magnet, (there separating a homogeneous region of spin-up from spin-down). A convenient semiclassical description is to write the modulated CDW as

$$\rho(\mathbf{r}) = \rho_c + \rho_o \cos(\mathbf{Q}\cdot\mathbf{r} + \phi(\mathbf{r})) \quad , \tag{9.6}$$

where the CDW is described by an amplitude ρ_o and a phase $\phi(\mathbf{r})$. ρ_c is the (uniform) background density of the electron gas. If the phase is a constant, it just defines the alignment of the density wave relative to the underlying lattice - and in the case of polyacetylene, the two states (b) and (c) of Fig. 9.5 are described by phases different by exactly π.

9.1.4 Incommensurate density waves, sliding

Polyacetylene is a simple case where the CDW is commensurate with the underlying lattice - a doubling of periodicity. Here there are two inequivalent states, and the charge on the domain wall is $2e/2$. In the two-dimensional material $2H-TaSe_2$ the period is 3 (i.e. $Q = G/3$, where G is a reciprocal lattice vector of the undistorted lattice). In this case, there are three different (but identical in energy) topological states — the domain walls have charge $2e/3$ per unit cell — and the domain walls can combine in three only at a vortex like defect.

Depending on the chemistry of the material, we may have CDW's that are entirely incommensurate with the underlying lattice, meaning that the periods bear no rational relationship.

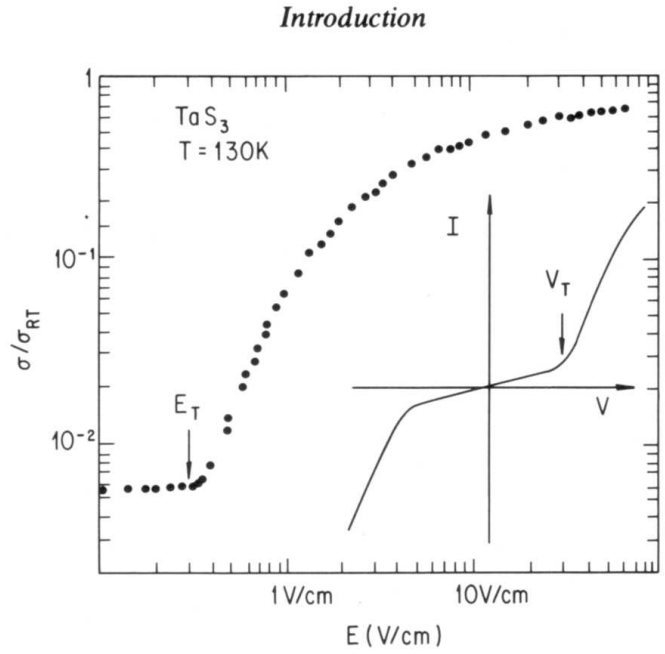


Figure 9.7: The inset shows the current induced in the cdw material TaS_3, and the main panel is the differential conductance $\sigma = dI/dV$. Below a threshold voltage, the CDW is pinned, and the only contribution to the current is the thermal activation of carriers across the CDW gap. Above V_T, there is a large contribution to the current that grows in a nonlinear fashion with increasing voltage, which is the contribution of the sliding of the CDW.[From Gruner and Gorkov,]

Notice that this means that the broken symmetry is now continuous - i.e. any value of the phase ϕ is equally good.

It can be shown that an applied electric field $\mathbf{E}$ couples to the *phase* ϕ of the CDW, with an energy $\mathbf{P} \cdot \mathbf{E} = (e/\pi)\phi E$, picking out the component of the electric field in the direction of $\mathbf{Q}$. An electric field therefore exerts a force on the CDW.

What is there to hold the CDW back? If the CDW is uniform, its energy does not depend on where it sits, but in a real solid there are always defects and impurities pin the CDW by locally deforming it. This means that a finite field has to be applied before the CDW will move, but beyond this threshold field the whole CDW slides through the lattice, reaching an equilibrium velocity determined by the "frictional force" induced by the relative motion with the underlying lattice. Because the CDW is rather stiff, it is not easy to pin, and the electrical fields required to get it to move can be small - no more than Vm^{-1} in some cases.

Above a threshold field E_T, the current associated with the sliding motion of a CDW grows in a non-linear fashion, $I \propto (E - E_T)^\nu$. It turns out that this is a true dynamic critical phenomenon - just as at a phase transition with temperature an order parameter turns on in a non-analytic fashion, so here the CDW current plays the role of the order parameter (Fig. 9.7).

9.2 Magnetism

By magnetism, in the widest sense, we understand the capacity of materials to change the magnetic field in their environment. It is not possible for classical systems in thermal equilibrium to be magnetic. This remarkable result, the Bohr-van Leeuwen theorem [2] (see, e.g., Feynman Lectures on Physics, Vol. 2) implies that all magnetic phenomena are rooted in the protection of orbital or spin angular momentum afforded by quantum mechanics.

Phenomenologically, we distinguish between materials which are *diamagnetic*, *paramagnetic* or *magnetically ordered*. A magnetically ordered material can exhibit magnetism even without an applied magnetic field. Diamagnets and paramagnets, on the other hand, only have a non-zero magnetisation when a field is applied. They differ in the direction of the response to the field. In a diamagnet, the magnetisation induced by an applied magnetic field will point in the direction opposite to that of the applied magnetic field, whereas the magnetisation in a paramagnet points along the direction of the applied field.

A diamagnetic response is a fundamental property of charged, quantum mechanical particles in a magnetic field. Because it is a very weak effect, however, it is only usually observed in materials with completely filled shells. When there are partially filled shells, or unpaired electrons, then the orbital and spin angular momentum of these electrons gives rise to a paramagnetic response, which usually far exceeds the diamagnetic moment produced by the remaining, paired electrons.

9.2.1 Local moments, Curie law susceptibility

The paramagnetic response of an isolated magnetic moment, or 'local moment', provides a very useful model system. Here, we discuss a classical calculation of the magnetic susceptibility of local moments; the corresponding quantum mechanical calculation forms one of the problems on the accompanying problem sheet. The fact that a classical calculation in this case gives rise to a finite magnetisation in an applied magnetic field does not contradict the Bohr-van Leeuwen theorem mentioned above, because the starting point of the calculation, an isolated magnetic moment of fixed magnitude, can only arise from quantum mechanics.

We consider a local moment $\mathbf{m}$ of fixed magnitude $\mu = |\mathbf{m}|$, in a small applied magnetic field $\mathbf{H} \to \mathbf{0}$. The dipole energy of this moment is given by $E = -\mu_0 \mathbf{m} \cdot \mathbf{H}$, and the probability of finding the moment pointing in a particular direction, at finite temperature T, is

$$p(\mathbf{m}) = e^{-E(\mathbf{m})\beta}/Z \tag{9.7}$$

where $\beta = \frac{1}{k_B T}$ and

$$Z = \int_{|\mathbf{m}|=\mu} p(\mathbf{m})\, d^2\mathbf{m} \tag{9.8}$$

[2] "At any finite temperature, and in all finite applied electrical or magnetical fields, the net magnetization of a collection of electrons in thermal equilibrium vanishes identically." (van Vleck, 1932).

We obtain the average moment:

$$\langle \mathbf{m} \rangle = \int\limits_{|\mathbf{m}|=\mu} \mathbf{m}\, p(\mathbf{m})\, d^2\mathbf{m} \tag{9.9}$$

The magnetic susceptibility is then given by $\chi_{ij}(T) = \frac{d\mathbf{m}_i}{d\mathbf{H}_j}$. Here, as in many more complex cases, the susceptibility is isotropic, i.e., $\chi_{ij}(T) = \delta_{ij}\chi(T)$

We can express $\chi_{ij}(T)$ as

$$\begin{aligned} \frac{d\mathbf{m}_i}{d\mathbf{H}_j} &= \frac{1}{Z} \int\limits_{|\mathbf{m}|=\mu} \mathbf{m}_i \frac{d}{d\mathbf{H}_j} e^{\mu_0 \mathbf{m}\mathbf{H}\beta} - \frac{1}{Z^2} \frac{dZ}{d\mathbf{H}_j} \int\limits_{|\mathbf{m}|=\mu} \mathbf{m}_i e^{\mu_0 \mathbf{m}\mathbf{H}\beta} && (9.10) \\ &= \mu_0\beta \left(\langle \mathbf{m}_i \mathbf{m}_j \rangle - \langle \mathbf{m}_i \rangle \langle \mathbf{m}_j \rangle \right) && (9.11) \\ &= \mu_0\beta \left(\langle \mathbf{m}_i \mathbf{m}_j \rangle \right) && (9.12) \end{aligned}$$

(because $\langle \mathbf{m}_i \rangle = \frac{1}{\mu_0\beta} \frac{1}{Z} \frac{dZ}{d\mathbf{H}_i} = 0$ for $\mathbf{H} = 0$)

In the limit $\mathbf{H} \to 0$, $e^{\mu_0 \mathbf{m}.\mathbf{H}\beta} \to 1$ and $\langle \mathbf{m}_x^2 \rangle = \langle \mathbf{m}_y^2 \rangle = \langle \mathbf{m}_z^2 \rangle = \frac{1}{3}\langle |\mathbf{m}|^2 \rangle = \frac{1}{3}\mu^2$

If we do not have just a single local moment, but rather N moments distributed over the volume V, then the associated susceptibility is

$$\chi = \frac{1}{3} \frac{N}{V} \mu_0 \mu^2 \frac{1}{k_B T} \tag{9.13}$$

9.2.2 Types of magnetic interactions

In many materials, a finite magnetisation is observed even in the absence of an applied magnetic field. This phenomenon must be produced by interactions coupling to the magnetic moment of the electrons. This is surprising, because the large Coulomb interaction between the electrons couples only to the charge, not to the spin of the electrons. The first idea might just be that the moments could couple through the dipole magnetic fields they generate. However, this is very small: the energy of interaction of two magnetic dipoles of strength m at a distance r is of order $\mu_o m^2/4\pi r^3$. Putting in a magnetic moment of order a Bohr magneton, we obtain

$$U_{dipolar} \approx \frac{\mu_o}{4\pi} \left(\frac{e\hbar}{2m} \right)^2 \frac{1}{r^3} \approx \pi\alpha^2 \left(\frac{a_{Bohr}}{r} \right)^3 \text{ Ryd.} \tag{9.14}$$

where $\alpha \approx 1/137$ is the fine structure constant. At typical atomic separations of 2 nm, this is about 4×10^{-5} eV, or less than a degree Kelvin — far too small to explain ordering temperatures of many hundreds of Kelvin as observed, for instance, in iron.

If the dipolar interactions are too weak to explain the robust magnetic order observed in real materials, then how can a spin-dependent interaction arise from the starting Hamiltonian governing the electrons and nuclei in the solid, in which only Coulomb interactions appear to be relevant? A number of distinct mechanisms have been identified, all of which consider particular situations in which the electrons are confined to a particular sub-set of low energy states, and in which the effects of Pauli exclusion are such that the total energy depends on the

spin configuration. There are complementary views of magnetism as originating either from the alignment of *local moments* or from a spontaneous spin polarisation of itinerant electrons. We begin with the former.

Direct exchange

As a first model system, let us consider two electrons in two orbitals, $|a>$, $|b>$, which are mutually orthogonal, and which are eigenstates of the single-particle Hamiltonian $\hat{H}_0$. Because the electrons are indistinguishable, the two-body wavefunction is antisymmetric under particle exchange:

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = -\Psi(\mathbf{r}_2, \mathbf{r}_1)$$

A simple approximation to the full two-body wavefunction can be formed from antisymmetrised product wavefunctions:

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}} \left(|ab\rangle - |ba\rangle \right) \quad ,$$

where the first slot in the Dirac-ket vector denotes the state occupied by electron 1, and the second slot the state of electron 2.

If we now consider spin, as well, then we find four possible antisymmetrised two-particle states, which can be grouped into one state with a singlet spin wavefunction, for which the spatial state is symmetric under particle exchange

$$\frac{1}{2}(|ab\rangle + |ba\rangle)(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)$$

and three states with triplet spin wavefunctions, for which the spatial state is antisymmetric under particle exchange

$$\frac{1}{2}(|ab\rangle - |ba\rangle) \begin{pmatrix} |\uparrow\uparrow\rangle \\ |\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle \\ |\downarrow\downarrow\rangle \end{pmatrix}$$

We find that subject to the full Hamiltonian $\hat{H} = \hat{H}_0 + \hat{H}_{1,2}$, where the interaction part $H_{1,2} = V(\mathbf{r}_1 - \mathbf{r}_2)$, the singlet state has a higher energy than the triplet state.

We introduce some shorthand notation:

$$\begin{aligned} E_0 &\equiv \langle ab|\hat{H}|ab\rangle = E_a + E_b + E_{Coul} && (9.15) \\ E_{Coul} &\equiv \langle ab|\hat{H}_{1,2}|ab\rangle && (9.16) \\ &= \int d^3\mathbf{r}_1 d^3\mathbf{r}_2 |\psi_a(\mathbf{r}_1)|^2 |\psi_b(\mathbf{r}_2)|^2 V(\mathbf{r}_1 - \mathbf{r}_2) && (9.17) \\ E_{ex} &\equiv \langle ba|\hat{H}_{1,2}|ab\rangle && (9.18) \\ &= \int d^3\mathbf{r}_1 d^3\mathbf{r}_2 \psi_b^*(\mathbf{r}_1)\psi_a(\mathbf{r}_1)\psi_a^*(\mathbf{r}_2)\psi_b(\mathbf{r}_2) V(\mathbf{r}_1 - \mathbf{r}_2) && (9.19) \end{aligned}$$

Here, E_{Coul} looks like Coulomb repulsion between charge densities, and E_{ex} resembles E_{Coul}, but the electrons have traded places ($\Rightarrow$ exchange term). For short range interactions, such as $V = \delta(\mathbf{r}_1 - \mathbf{r}_2)$, $E_{Coul} \to E_{ex}$.

We find that the energy of the singlet state is

$$E_{singlet} = \frac{1}{2}\Big(\langle ab + ba|\hat{H}|ab + ba\rangle\Big) \tag{9.20}$$
$$= E_0 + E_{ex} \tag{9.21}$$

The energy of the triplet state, however, is lower:

$$E_{triplet} = \frac{1}{2}\Big(\langle ab - ba|\hat{H}|ab - ba\rangle\Big) \tag{9.22}$$
$$= E_0 - E_{ex} \tag{9.23}$$

There is, therefore, a spin dependent effective interaction in this simple model system. Note that this interaction arises, because the electrons have been constrained to single occupancy of the two orbitals, leaving only spin flips as the remaining degrees of freedom.

This simple example reflects a general phenomenon: the spin triplet state is symmetric under particle exchange and must therefore be multiplied by an antisymmetric spatial wavefunction. An antisymmetric spatial wavefunction must have nodes whenever two spatial coordinates are equal: $\psi(...., r_i = r, ...r_j = r, ...) = 0$. So it is then clear that the particles stay farther apart in an antisymmetrised spatial state than in a symmetric state. This reduces the effect of the repulsive Coulomb interaction. Therefore it is because of the combination of Pauli principle and Coulomb repulsion that states with antisymmetric spatial wavefunction (which will generally have high spin) have lower energy.

When the orbitals concerned are orthogonal, E_{ex} is *positive* in sign, i.e., the lowest energy state is a triplet. However, if the overlapping orbitals are not orthogonal – as will happen between two orbitals between neighbouring atoms – the interaction may be *negative*, so the lowest energy is a singlet.

Heisenberg Hamiltonian

We can express the spin-dependent interaction between the electrons, which has arisen from the direct exchange term E_{ex}, in terms of the spin states of the two electrons, which are probed by the spin operators $\hat{S}_1$ for electron 1 and $\hat{S}_2$ for electron 2. Because triplet and singlet states differ in the expectation value of the magnitude of the total spin $\hat{S} = \hat{S}_1 + \hat{S}_2$, we can use this to distinguish between the singlet and triplet states:

$$\hat{S}^2 = (\hat{S}_1 + \hat{S}_2)^2 = \frac{3}{2} + 2\hat{S}_1 \cdot \hat{S}_2$$

This leads to the definition of a new operator $\hat{H}_{spin}$

$$\hat{H}_{spin} = \frac{1}{4}(E_{singlet} + 3E_{triplet}) - (E_{singlet} - E_{triplet})\hat{S}_1 \cdot \hat{S}_2$$

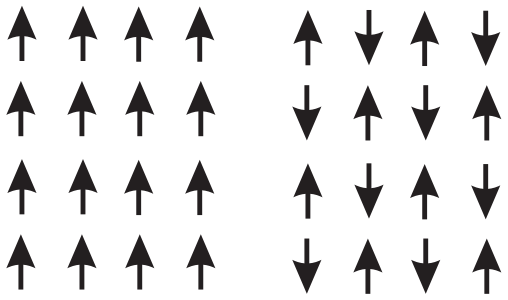


Figure 9.8: . Schematic picture of the ground state of a ferromagnet and an antiferromagnet. The order parameter for the ferromagnet is the uniform magnetisation, and for an antiferromagnet it is the staggered magnetisation $< \mathbf{S}(\mathbf{Q}) >$, where $\mathbf{Q}$ is the wavevector corresponding to the period of the order

The eigenvalues of this operator are $E_{singlet}$ for the singlet state, and $E_{triplet}$ for the triplet state. They therefore reproduce the spectrum of the full Hamiltonian, provided that only spin state changes are allowed, i.e., we restrict ourselves to low energy excitations.

By defining $J = (E_{singlet} - E_{triplet})/2$ and shifting the zero in energy, we then obtain the *Heisenberg Hamiltonian* for two electrons

$$\hat{H}_2 = -2J\hat{S}_1 \cdot \hat{S}_2 \quad ,$$

which can be extended naturally to a collection of spins

$$\hat{H}_{Heisenberg} = -\sum_{ij} J_{ij}\hat{S}_i\hat{S}_j.$$

Depending on the sign of J, the ground state of the Heisenberg model will be ferromagnetic (aligned spins) or anti-ferromagnetic (anti-aligned spins on neighbouring sites, Fig. 9.8); more complicated magnetic states can arise if we have different magnetic ions in the unit cell, and also on taking account of magnetic anisotropy.

Superexchange and insulating antiferromagnets

When there is strong overlap between orbitals, as in a typical covalent bond, then it is advantageous for the system to form hybridised molecular orbitals and to occupy them fully. In this case, the singlet state has far lower energy than the triplet state, and the system has no magnetic character. However, a special class of much weaker interactions can be important when two magnetic moments are separated by a non-magnetic ion (often O^{2-}) in an insulator (Fig. 9.2.2). Direct exchange between the two local moments is unimportant, because they are too far apart. We consider a ground state in which the relevant valence state of each magnetic ion is singly occupied and that of the non-magnetic ion is doubly occupied. The spectrum of excitations from this ground state is now dependent on the spin orientation of the electrons on the magnetic moments: if the two spins are antiparallel, then it is possible for an electron from the non-magnetic ion to hop onto one of the magnetic ions, and be replaced by an electron from the other magnetic ion. Although the state created in this way has a higher energy than the ground state, it can be admixed to the initial ground state and will – in second order

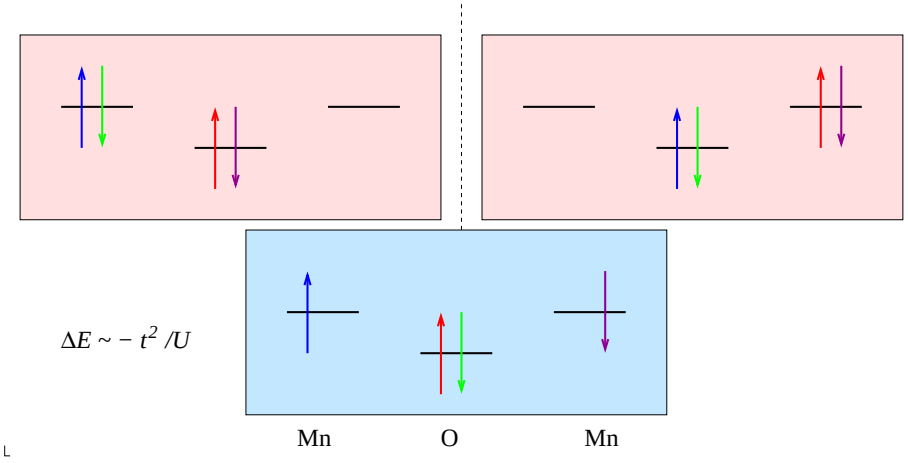


Figure 9.9: An illustration of the superexchange mechanism in antiferromagnetic insulators. Two magnetic moments are separated by a non-magnetic ion. The large separation rules out a direct exchange interaction between the magnetic moments. The excited states (top left and top right), which involve double occupancy of an atomic orbital on the magnetic ions, are only possible, if the original spin state of the two magnetic ions was anti-aligned. Only for anti-aligned magnetic moments can the system therefore benefit from the lower energy, which admixing an excited state brings in second order perturbation theory. This results in a ground state energy, which depends on the mutual spin orientation of the two magnetic moments.

perturbation theory – always cause the new, perturbed, ground state energy to be lowered. This admixture would not be possible if the two magnetic moments were aligned. We arrive, therefore at a total energy for the system which depends on the mutual orientation of the two magnetic moments.

Second order perturbation theory suggests that this effective *superexchange interaction* is of order $J \sim -t^2/U < 0$, where t is the matrix element governing hopping between the magnetic moment and the non-magnetic ion, and U is the Coulomb repulsion energy on the magnetic moment. When extended to a lattice, it favours an antiferromagnetic ground state, in which alternate sites have antiparallel spins. On complicated lattices, very complex arrangements of spins can result.

The magnitude of this interaction is often quite small, in the range of a few to a few hundred degrees Kelvin. Consequently, these systems will often exhibit phase transitions from a magnetically ordered to a disordered paramagnetic state at room temperature or below.

Band magnetism in metals

Let us start with Pauli paramagnetism – the response of a metal to an applied magnetic field. We consider a Fermi gas with energy dispersion $\epsilon_{\mathbf{k}}$ in a magnetic field H. In a magnetic field, the spin-up and spin-down bands will be Zeeman-split (see Fig. 9.10):

$$\begin{aligned} \epsilon_{\mathbf{k}\uparrow} &= \epsilon_{\mathbf{k}} - \mu_B B_a \quad , \\ \epsilon_{\mathbf{k}\downarrow} &= \epsilon_{\mathbf{k}} + \mu_B B_a \quad . \end{aligned} \tag{9.24}$$

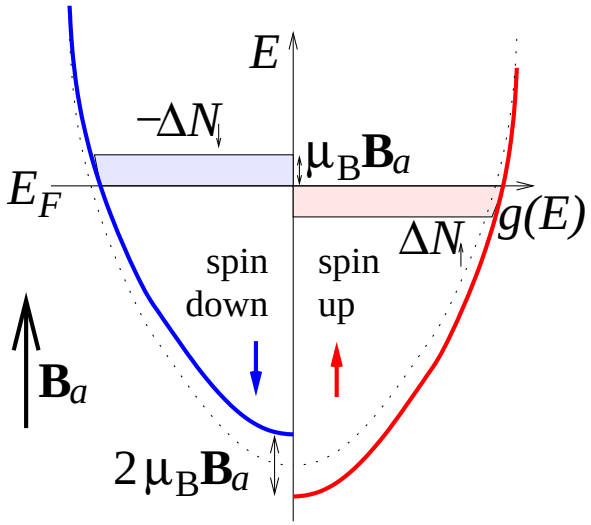


Figure 9.10: Effect of an applied field $B_a = \mu_0 H$ on the spin-up and spin-down electrons in a metal. The spin-dependent shift of the energy levels – while keeping the chemical potential (and the Fermi energy) constant – produces a population imbalance between the two spin species. This converts to a magnetic moment, which is proportional to the density of states at the Fermi level.

Since the chemical potential must be the same for both spins, there must be a transfer of carriers from the minority spin band to the majority spin band:

$$n_\uparrow - n_\downarrow = \mu_B B_a g_v(E_F) \tag{9.25}$$

where $g_v(E_F)$ is the density of states at the Fermi level per unit volume.[3] We could relate g_v to the density of states per atom as $g_v = \frac{N}{V} g$. The magnetisation is $M = \mu_B(n_\uparrow - n_\downarrow)$, and $B_a = \mu_0 H$, which gives the static spin susceptibility

$$\frac{M}{H} = \chi_\sigma = \mu_0 \mu_B^2 g(E_F) \quad . \tag{9.26}$$

We should compare this result to the susceptibility obtained for local moments (Eq. 9.13): $\chi(T) = \frac{1}{3}\mu_0 \mu^2 \frac{N}{V} \frac{1}{k_B T}$. Whereas the Curie law susceptibility is strongly temperature dependent, the Pauli susceptibility of metals is temperature independent (at least at low temperatures, where thermal broadening of the Fermi function is unimportant). On the other hand, there are also similarities: both expressions contain the square of the fluctuating moment, and both have an energy scale in the denominator. Note that the density of states per unit volume can be written roughly as $g(E_F) = \frac{N}{V} \frac{1}{E_F}$, up to a constant of order one. For the Curie law, the energy scale in the denominator is the thermal energy $k_B T$, whereas in the Pauli expression, it is the Fermi energy. To approach this problem from a different angle, we could also use the argument made when explaining the linear heat capacity in metals, namely that only a fraction $k_B T/E_F$ of the electrons are sufficiently close to the Fermi level to act like classical particles. Here, we argue that this fraction of electrons can act like local moments. Multiplying the Curie law for local moments by the fraction $k_B T/E_F$ transforms it, up to a constant of order one, into the Pauli expression for metals.

[3]For this, we must assume that the splitting is small enough that the density of states can be taken to be a constant. We define $g(\mu)$ to be the density of states for both spins.

Now let us include in a very simple fashion the effect of interactions. The Stoner-Hubbard model, which provides arguably the simplest way forward, includes an energy penalty U for lattice sites which are doubly occupied, i.e., they hold both an up- and a down-spin electron.

$$\hat{H}_{int} = \sum_{sitesi} U n_{i\uparrow} n_{i\downarrow} \quad , \tag{9.27}$$

If we treat this interaction in a mean-field approximation, it leads to a shift of the energies of the two spin bands

$$\begin{aligned} \epsilon_{\mathbf{k}\uparrow} &= \epsilon_{\mathbf{k}} + U\bar{n}_{\downarrow} - \mu_0\mu_B H \\ \epsilon_{\mathbf{k}\downarrow} &= \epsilon_{\mathbf{k}} + U\bar{n}_{\uparrow} + \mu_0\mu_B H \end{aligned} \tag{9.28}$$

We see that the presence of spin-down electrons increases the energy of the spin-up electrons in the same way as a magnetic field pointing down would. Conversely, spin-up electrons cause the energy of spin-down electrons to increase in the same way as a magnetic field pointing up. The interactions between the electrons appear formally in the same way as an additional magnetic field. This so-called exchange field is not physical in the sense that it could deflect a compass needle, it is a book-keeping device to handle the effects of the Coulomb interaction between the electrons.

With the same approximation as before - that the density of states can be taken to be a constant, we can then self-consistently determine the average spin density

$$\frac{N}{V}(\bar{n}_{\uparrow} - \bar{n}_{\downarrow}) = [U(\bar{n}_{\uparrow} - \bar{n}_{\downarrow}) + 2\mu_0\mu_B H]\frac{1}{2}g_v(E_F) \quad . \tag{9.29}$$

The magnetisation is $M = \mu_B(n_{\uparrow} - n_{\downarrow})$ which then gives us the static spin susceptibility

$$\chi_{\sigma} = \mu_0 \frac{\mu_B^2 g(E_F)}{1 - \frac{Ug(E_F)}{2}} \quad . \tag{9.30}$$

Here, g denotes the density of states per atom, in contrast to $g_v = \frac{N}{V}g$, which is the density of states per unit volume. In comparison to the non-interacting case, the magnetic susceptibility is enhanced, and will diverge if U is large enough that the Stoner criterion is satisfied

$$\frac{Ug(E_F)}{2} > 1 \quad , \tag{9.31}$$

which marks the onset of ferromagnetism in this model.

The Stoner criterion for ferromagnetic order has a very fundamental interpretation: because the density of states per atom is of order $\frac{1}{E_F}$, the Stoner criterion expresses the balance between interaction energy U and kinetic energy E_F. If the kinetic energy of the electrons is high, then they will not form a magnetically ordered state. If, on the other hand, the interaction strength is higher than the kinetic energy, then the electron system can lower its energy by aligning its spins. Variations on this criterion surface in many other areas of correlated electron physics.

Local moment magnetism in metals – indirect exchange

In a d-band metal, such as iron, or in f-band metals, such as gadolinium or erbium there are both localised electrons with a moment derived from the tightly bound orbitals, and *itinerant* electrons derived from the *sp* bands. The itinerant bands are weakly, if at all, spin-polarised by themselves because the exchange interactions are small and the kinetic energy large. However, the itinerant carrier acquires an *induced* spin polarisation due to its interaction with the core spin on one atom. This spin polarisation can then be transmitted to a neigbouring ion, where it attempts to align the neighbouring spin. There is then an interaction between the localised electron spins, which is mediated by the itinerant electrons, often called RKKY (for Ruderman-Kittel-Kasuya-Yoshida).

A more detailed view of this process can be given. If we have an ion of spin $\mathbf{S}$ embedded in the conduction electrons, one would expect that the local direct exchange will give rise to a contact interaction of the form

$$H_{int} = -J\mathbf{S}\cdot\mathbf{s}\,\delta(\mathbf{r}) \quad , \tag{9.32}$$

with $\mathbf{s}$ the conduction electron spin density, and J a direct exchange interaction. The spin density is not otherwise polarised, but the perturbation will induce a weak spin density modulation in the conduction cloud, which will of course decay away to zero at large distance from the ion. The induced spin density is just

$$s(\mathbf{r}) = J\chi_\sigma(\mathbf{r})S \tag{9.33}$$

where we have introduced the spin susceptibility χ_σ. (Above we considered the average spin susceptibility to a uniform field, this is a generalisation to non-uniform fields).

At a nearby lattice site (say $\mathbf{r}$), the induced spin density caused by the polarisation of one atom interacts with the spin of another, and the energy is then

$$-J\mathbf{S}(\mathbf{r})\cdot\mathbf{s}(\mathbf{r}) = J^2\chi_\sigma(\mathbf{r})\mathbf{S}(\mathbf{r})\cdot\mathbf{S}(0) \quad , \tag{9.34}$$

Summing over all pairs of sites in the crystal we obtain

$$H_{RKKY} = -\sum_{ij} J^2\chi_\sigma(\mathbf{r}_{ij})\mathbf{S}(\mathbf{r}_i)\cdot\mathbf{S}(\mathbf{r}_j) \quad . \tag{9.35}$$

If we could replace $\chi_\sigma(\mathbf{r}_{ij})$ by its average (say Eq. (9.26)) then one would predict a long range ferromagnetic interaction, which is not far from the truth for many materials. Of course, in a more accurate theory, χ decays as a function of distance. A careful analysis shows in fact that χ *oscillates*, changing sign as it decays, with a wavelength π/k_F. These *Friedel oscillations* are connected with the sharp change in occupation numbers at the Fermi surface.

9.2.3 Magnetic order and Weiss exchange field

An alternative approach to the problem of magnetism could start with a phenomenological equation of state, linking magnetisation and magnetic field

$$H = aM + bM^3 \quad , \tag{9.36}$$

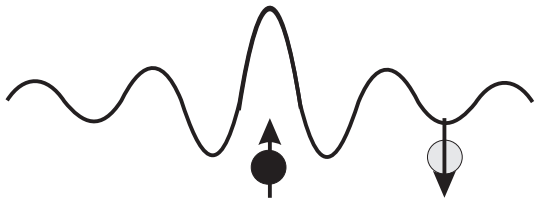


Figure 9.11: In metals, a local moment will polarise the conduction electron spins, producing a spin density that decays with distance and oscillates in sign with period $1/2k_F$. The interaction of the induced spin density with a neighbouring local moment produces the RKKY interaction.

where, to keep things simple, we neglect the vector nature of H and M. In this equation of state, a takes the role of the inverse susceptibility $\chi^{-1} = dH/dM$ and b ensures that the magnetisation bends over towards saturation for high fields.

For a material to be ferromagnetic, we require a finite M, a *remanent* magnetisation, even for zero H. This would appear to be possible only if the parameter a in the equation of state is negative. Systems of non-interacting electrons, however, do not exhibit a negative susceptibility: the Curie law for isolated local moments gives $a \propto T$, whereas in metals, the Pauli susceptibility is positive and only very weakly temperature dependent. We therefore need to introduce a further term which captures the effect of interactions between the electrons.

The simplest way to incorporate these interactions is to introduce an *exchange molecular field*, h, into the equation of state: $H + h = aM + bM^3$. The exchange molecular field is not a real magnetic field, which could deflect a compass needle or induce voltages in a pick-up coil. It is a way to represent, in a mean field sense, the effect of the exchange interaction produced by other electrons on a test electron. If we assume that the exchange field is simply proportional to the overall magnetisation (with constant of proportionality λ, this is the Weiss molecular field concept), then we arrive at a feed-back equation:

$$H + \lambda M = aM + bM^3 \quad ,$$

which can be recast in the form of the original equation of state, with a modified linear coefficient $a^* = a - \lambda$:

$$H = (a - \lambda)M + bM^3 = a^*M + bM^3$$

We see that although the noninteracting susceptibility is finite, interactions between the electrons give rise to a feed-back effect, which boosts the magnetic susceptibility $\chi = 1/a^* = \chi_0/(1 - \lambda)\chi_0$, where $\chi_0 = 1/a$ is the noninteracting susceptibility. This leads to a magnetic instability, if

$$\lambda\chi_0 > 1 \tag{9.37}$$

The above equation represents a more general form of the *Stoner criterion* (Eq. 9.31).

Chapter 10

Fermi liquid theory

10.1 The problem with the Fermi gas

Our modelling of electrons in solids so far has been based on a fairly simple-minded approach: instead of looking for the eigenstates of the many-particle system we are interested in, we instead calculate the electronic eigenstates of a single-particle Hamiltonian (subject to the periodic lattice potential). We then fill these eigenstates with electrons according to the Fermi occupation factor, treating our system as a degenerate Fermi gas. This separation of a many-particle problem into single-particle states relies on being able to separate the many-body wave function into an antisymmetrised product of single-particle wavefunctions. For a two-electron state, for instance, we could write $\Psi(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}} (\psi_a(\mathbf{r}_1)\psi_b(\mathbf{r}_2) - \psi_a(\mathbf{r}_2)\psi_b(\mathbf{r}_1))$. Such product states, however, implicitly ignore correlations between the electrons. For example, we would find that expectation values of products such as $\langle \mathbf{r}_1 \mathbf{r}_2 \rangle$ decompose into the products of expectation values $\langle \mathbf{r}_1 \rangle \langle \mathbf{r}_2 \rangle$.

In the presence of strong electron-electron interactions the the electronic motion must be correlated. We saw last term that in the Thomas-Fermi approximation, the electrons in a Fermi gas react to the introduction of a charged impurity, in such a way as to screen the impurity potential at long distances. Taking this idea to the next level, we could say that for any one electron under consideration, which of course carries a negative charge, the other electrons in the metal execute a correlated screening motion, which would reduces their density in the vicinity of the first electron. This reduces the effective range of the Coulomb potential due to the electron under consideration, which we might take as justification for ignoring the Coulomb interaction. However, this also implies that the electrons undergo correlated. Such correlations are not contained in a single-particle description.

On the other hand, the band structure approach which we have used so far has been remarkably successful in modelling a wide range of materials and phenomena. It explains electronic transport and thermodynamic properties such as the heat capacity in metals, we have used it to understand semiconductors and semiconductor devices, and it is consistent with Fermi surface probes such as quantum oscillations in high magnetic fields, as well as other probes of electronic structure. These successes suggests that it is not altogether wrong. How can we reconcile the success of the single particle picture with its conceptual difficulties?

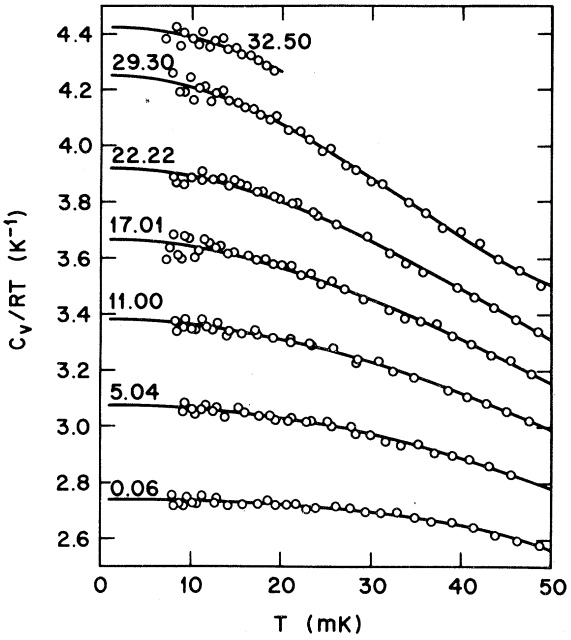


Figure 10.1: Heat capacity of ^3He at low temperature (from Greywall, Phys. Rev. B **27**, 2747 (1983)) at different pressures p in bar. The data illustrates that even in a densely packed assembly of atoms, the low temperature properties are consistent with what we expect of a Fermi *gas*, namely the heat capacity becomes linearly dependent on temperature T. The Sommerfeld coefficient $\gamma = C/T$ increases with increasing pressure, as the ^3He system approaches solidification.

10.1.1 The extreme case of liquid helium 3

The example of ^3He at low temperature illustrates the problem we face. The helium isotopes ^4He (bosons) and ^3He (fermions) are special in that they do not solidify, at least at ambient pressure, down to zero temperature. This is caused by their very weak mutual interaction and their low mass, which boosts their quantum-mechanical zero point motion. Because of this, they are examples of *quantum fluids*, and fermionic ^3He can be regarded as an uncharged analogue of electrons in solids.

Because of their hard-core repulsion at small separation and weak van der Waals attraction, we can picture helium atoms at low temperature as forming a closely-packed assembly of hard spheres, which does not form a solid but must clearly be very strongly correlated. Nevertheless, measurements of all key properties at low temperature show good agreement with results from Fermi gas theory. For instance, the molar heat capacity of a degenerate Fermi gas $C_m/T \rightarrow$ const. at low T, and this is exactly what is found in ^3He (Fig. 10.1).

If we look more closely, however, then we realise that – while the general form of the heat capacity and other properties is the same as for a Fermi gas – the detailed prefactors may be different. For instance, in a Fermi gas, $C_m/T \simeq \frac{\pi^2}{2}\frac{R}{T_F}$, where T_F is the Fermi temperature, which is determined by the density and the mass of the particles. In the case of ^3He, C_m/T rises rapidly with increasing pressure, whereas a simple-minded Fermi gas calculation would actually predict an increase in T_F with increasing pressure (as the density rises), and thereby a decrease in C_m/T. This suggests that the basic relationships worked out for a degenerate

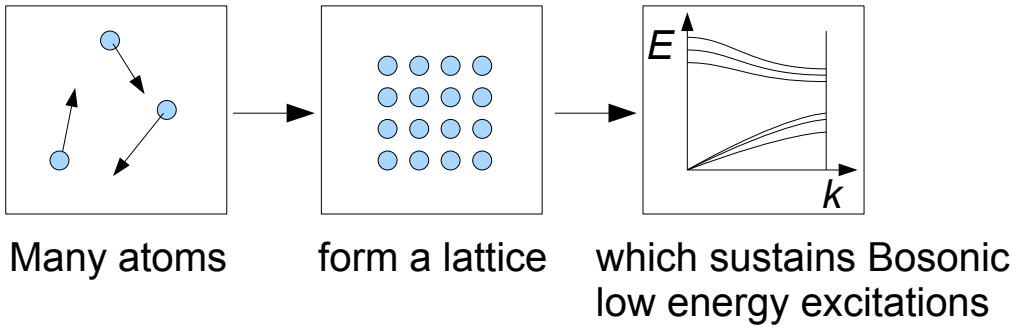


Figure 10.2: Modelling of the crystal lattice at low temperature in terms of its elementary excitations, the lattice vibrations or phonons.

Fermi gas carry over for strongly correlated ^3He, but that parameters such as the particle mass are renormalised.

10.2 Collective excitations

In dealing with the problem of the above-mentioned interacting Fermi system formed by the ^3He atoms, or by the electrons in a metal, we seek inspiration from the very successful modelling of the collective motion of the lattice atoms in a solid. There, too, we have an interacting many body system, which is formed in this case by the ionic cores in the crystal. This lattice system is successfully modelled by cataloguing its low energy excitations, the lattice vibrations.

Because we know that the atoms in a crystal form a regular lattice, we do not need to model the motion of every single atom. Instead, we can concentrate on the deviations of the system from its ground state structure. These deviations, or excitations, are deformations of the lattice, and we can decompose them into a set of normal modes, or elementary excitations, which we can label with a wavevector k, and for which we can compute a frequency of vibration ω. Because the vibrations of a harmonic lattice at a particular wavevector k are described by the same Hamiltonian as a quantum harmonic oscillator, creation and annihilation operators can be used to generate excited states. Exciting the vibrational state of the lattice is thought of as 'creating a phonon'. The commutation relations between the creation and annihilation operators, which allow multiple excitation of the same $k-$state, cause the phonons to follow Bose statistics. What we find, is that a collection of atoms (which themselves may be fermions or bosons!) can form a lattice, and the low energy excitations of the lattice can behave like a Bose gas. By making this step we have achieved a tremendous simplification of the original problem: where there were originally many atoms, coupled via a strongly anharmonic interaction, we now have a small number of elementary excitations, which interact only weakly, and whose effect at low temperatures can be modelled conveniently in terms of a relatively simple Bose gas calculation.

Can we do the same for electrons? This is the idea behind Landau's Fermi liquid theory. When a collection of fermions forms a strongly interacting, correlated assembly, the low energy excitations of the 'liquid' formed from the interacting fermions behaves like a gas of weakly interacting fermions, but with parameters (e.g. particle mass) different from those of the interacting particles from which it arises (Fig. 10.3). This would explain why a single-particle

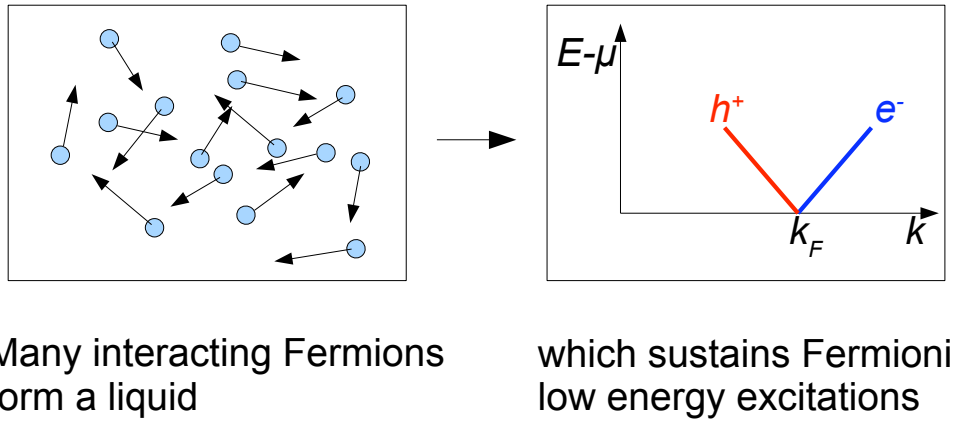


Figure 10.3: Modelling of a Fermi system in terms of its elementary excitations, electrons or holes close to the Fermi surface. This is the essence of Fermi liquid theory.

description works so well in many materials. In many cases, the properties of the electrons making up the Fermi liquid carry over with only slight modification to the properties of the fermionic excitations of the Fermi liquid.

10.2.1 Adiabatic continuity

We could approach the interacting electron state in this way: let us begin by imagining an assembly of electrons that do not interact. We know that in this case, the electron system will form a Fermi gas, which means that the ground state is represented by completely filled states inside the Fermi surface and empty states outside the Fermi surface, and that the low energy excitations from the ground state are electrons just outside the Fermi surface and holes just inside the Fermi surface. We then very gradually turn on the interaction between the electrons and follow the evolution of the energy levels of the system. The principle of *adiabatic continuity* (Fig. 10.4) suggests that we can continue to label the energy eigenstates in the same way as for the non-interacting system: energy eigenstates shift, when the system is tuned, but their labels remain useful. We can therefore assume that the excitations of the interacting Fermi system, the Fermi liquid, follow the same basic rules as those of the Fermi gas.

One important consequence of this one-to-one correspondence between the quasiparticle states and the states of the non-interacting system is that the volume of the Fermi surface is unchanged, as the interaction is turned on. This is called Luttinger's theorem.

As usual, there are many hidden pitfalls in this argument. For example, it only holds if the energy levels do not cross as the interaction is turned on. This is not guaranteed, and in fact it is difficult to find a non-trivial example of an interacting system in which the energy levels do not cross on tuning. It is safer to say that for interacting Fermi systems, a Fermi liquid state in the sense discussed above is possible, but that not every Fermi system will necessarily be described in this way.

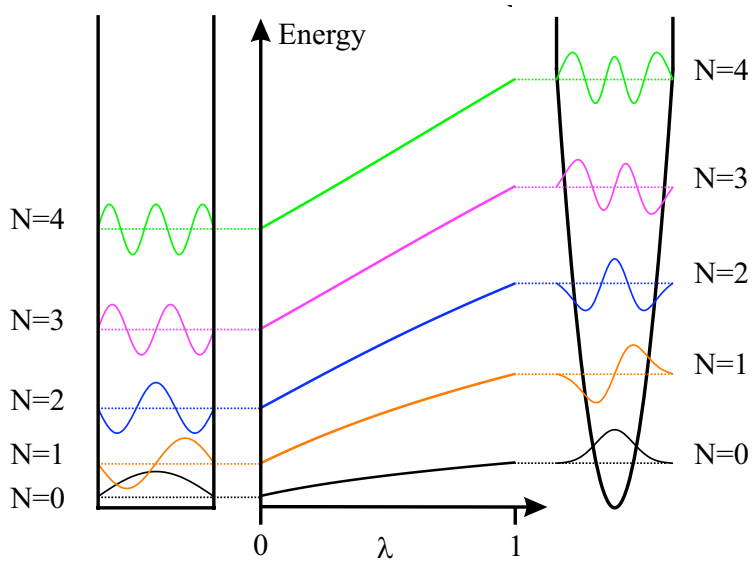


Figure 10.4: An example of adiabatic continuity (from Schofield, Contemporary Physics **40**, 95 (1999)): as the Hamiltonian of a system is changed – here, we gradually change a square well potential into a simple harmonic oscillator potential – the energy levels and details of the eigenstates drift, but we can continue to label the states with the same labels as before.

10.3 Total energy expansion for Landau Fermi liquid

Adopting Landau's Fermi liquid approach, we label excited states of the interacting system by quantum numbers of the non-interacting system, such as wavevector $\mathbf{k}$, spin, band index, etc. In analogy with the case of lattice vibrations, where excited states are expressed in terms of a new particle called the phonon, we talk of a 'quasiparticle' at wavevector $\mathbf{k}$, if the system is in an excited state labelled with that wavevector, which would have to lie outside the Fermi surface. Of course, there can also be 'quasiholes', corresponding to excitations at wavevectors inside the Fermi surface.

We can then express the total energy of the interacting system as a functional of the occupation numbers of the various states labelled in this way. At low temperatures, when the number of excitations is small, this functional could be approximated by a Taylor expansion.

$$E[n_{\mathbf{k}}] = \sum_{\mathbf{k}} \epsilon(\mathbf{k}) n(\mathbf{k}) + \frac{1}{2} \sum_{\mathbf{k}\mathbf{k}'} f(\mathbf{k}, \mathbf{k}') n(\mathbf{k}) n(\mathbf{k}') \tag{10.1}$$

The first term on the right-hand side is familiar. It expresses the energy of having quasiparticles in band states of energy $\epsilon(\mathbf{k})$. The next term on the right-hand side, which is second-order in occupation number $n(\mathbf{k})$, accounts for interactions between excited state, or quasiparticles. By postulating such a relatively simple expression for the total energy, Landau was able to arrive at a variety of key expressions that link material properties such as heat capacity, magnetic susceptibility and compressibility to properties of the quasiparticles and their interaction function f.

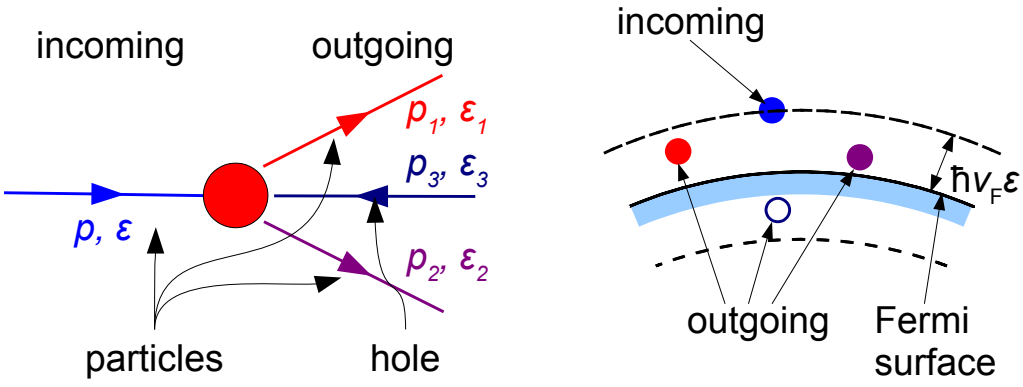


Figure 10.5: Quasiparticle scattering event. The rate at which quasiparticles scatter from the rest of the electron system is largely determined by phase-space constraints, which result in the scattering rate rising with the square of the quasiparticle energy, referenced against the Fermi energy.

10.3.1 Energy dependence of quasiparticle scattering rate

Why is it that the quasi-particle scattering rate Γ can be small in a metal where the typical separation between electrons is only an Angstrom or so? Consider the following scattering event: a quasiparticle just outside the Fermi surface scatters by creating an electron-hole pair. In other words, one particle with energy ϵ and momentum p comes in, and two particles and one hole come out (Fig. 10.5).

To satisfy the Pauli exclusion principle, all the outgoing particles must have energy (referenced to E_F) > 0. Remember that the hole energy $= -\epsilon_3$. Moreover, by conservation of energy, $\epsilon = \epsilon_1 + \epsilon_2 - \epsilon_3$ is fixed by the energy of the incoming particle. This implies that ϵ_1 and ϵ_2 can be chosen freely from the range $[0, \epsilon]$, and this fixes ϵ_3. Hence, the number of available final states for the scattering event rises with increasing incident quasiparticle energy as $\propto \epsilon^2$, which means that the scattering rate itself decreases with the square of the excitation energy at low energies:

$$\text{Scattering rate } \Gamma \sim \epsilon^2 \quad . \tag{10.2}$$

We find, therefore, that quasiparticles with energy $\epsilon \to 0$, i.e., close to the Fermi surface, scatter extremely rarely. Despite the fact that they are moving through a strongly interacting and dense liquid, these quasiparticles can travel long distances before they scatter. Their free motion is protected by the Pauli exclusion principle, because there are very few empty states into which these particles could scatter. This is a key result in Landau's Fermi liquid theory: close to E_F, particles interact but do not scatter, and are therefore long-lived.

This is an extraordinarily important result for metals. It explains why it is that the mean free path in, e.g., copper, is very long at low temperatures if the material is pure enough, despite the fact that the characteristic separation between electrons is of order a lattice constant and their interaction energy is of order a few eV. The electrical current is carried by a quasiparticle excitation that is a collective mode of the Fermi system. In the language of perturbation theory, the quasi-particle is a "dressed" excitation, that involves a correlated motion of the

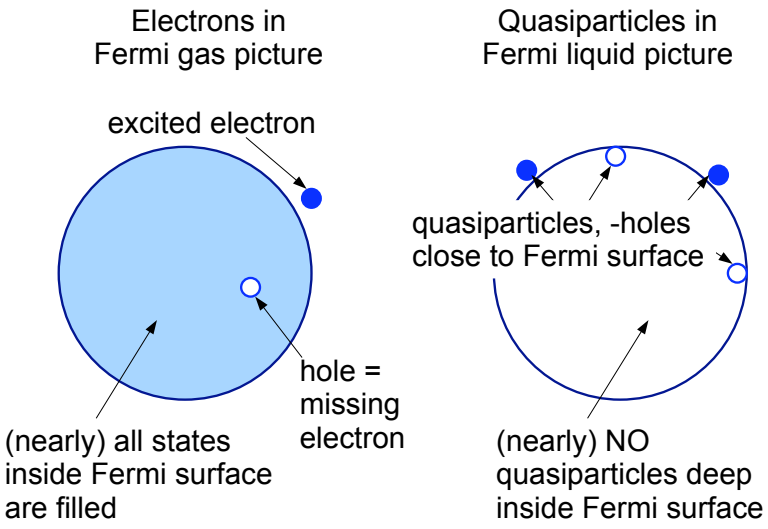


Figure 10.6: Correspondence and differences between the Fermi gas and Fermi liquid pictures.

added electron together with the many-body background.

Conversely, with increasing $\epsilon = E - E_F$, the scattering rate grows more quickly than ϵ. When the scattering rate exceeds $\epsilon/\hbar$, the quasiparticles are no longer well-defined, because they scatter before their wavefunction can undergo a full oscillation. The quasiparticles become overdamped. Landau's Fermi liquid approach can therefore only apply to low energy excitations. It will only work well at low temperature, so that $k_B T << E_F$. This is still a very wide range, because in most metals E_F corresponds to thousands of Kelvin.

10.3.2 Quasiparticles and -holes live near the Fermi surface

Because Fermi liquid quasiparticles can only be defined close to the Fermi surface, we have to be careful in thinking about the ground state of the Fermi liquid (Fig. 10.6). For a Fermi gas, it is unproblematic to think of the ground state in terms of a volume in **k**-space which is enclosed by the Fermi surface and in which all **k**-states are filled by electrons. This does not carry over easily to the Fermi liquid. For a Fermi liquid, the wavevectors **k** label *quasiparticle* states. These are only well-defined for **k** close to the Fermi surface, because only there does the lifetime of the quasiparticles become large enough. We cannot picture the ground state, then, as a filled Fermi sea, but rather we have to think in terms of an unusual 'vacuum' ground state, from which the low energy excitations are fermionic particles and holes which sit on either side of a surface in momentum space, namely the Fermi surface.

10.3.3 Quasiparticle spectral function

We can also approach the Fermi liquid state from another angle by considering the response of the many-particle system to addition or removal of a quasiparticle.

Let us begin by considering the case of a non-interacting system. If we place a particle into a single-particle eigenstate of the Hamiltonian labelled by its momentum **k**, then the wavefunction

will evolve in time according to the Schrödinger prescription

$$\psi_k(\mathbf{r},t) = \psi_k(\mathbf{r})e^{-i\epsilon_{\mathbf{k}}t/\hbar} \quad . \tag{10.3}$$

Here ψ_k is the Bloch wavefunction satisfying the time-independent Schrödinger equation, and the time-dependent solution oscillates in time with a single frequency, given by $\hbar\omega = \epsilon_{\mathbf{k}}$, the band energy.

If we look at this in Fourier space, we would say that

$$\psi_k(\mathbf{r},\omega) = 2\pi\psi_k(\mathbf{r})\delta(\omega - \epsilon_{\mathbf{k}}) \tag{10.4}$$

so that the wavefunction has spectral weight only at $\omega = \epsilon_{\mathbf{k}}/\hbar$.[1]

We can say that the probability amplitude of finding an electronic state with energy ω and momentum $\mathbf{k}$ is

$$A(\mathbf{k},\omega) = \delta(\omega - \epsilon_{\mathbf{k}}) \quad , \tag{10.5}$$

where the quantity $A(\mathbf{k},\omega)$ is usually called the *electron spectral function*.

What about an interacting system? In this case, (i) interactions modify the precise form of the dispersion relation, so we should replace $\epsilon_{\mathbf{k}}$ by a renormalised $\tilde{\epsilon_{\mathbf{k}}}$. This latter is often referred to as a mass renormalisation: $m^*/m = \epsilon_{\mathbf{k}}/\tilde{\epsilon_{\mathbf{k}}}$. Moreover, (ii), if we add a quasiparticle, it will scatter from other quasiparticles or by creating particle-hole pairs. Because of this, the probability amplitude of finding a quasiparticle of momentum $\mathbf{k}$ at time t will decay exponentially at a rate given by the quasiparticle scattering rate $\Gamma_{\mathbf{k}}$. These two effects change the time-dependence of the state to $e^{(i\epsilon_{\mathbf{k}}/\hbar - \Gamma_k)t}$ and after a Fourier transform lead to an *ansatz* for the spectral function in an interacting system of the form

$$A(\mathbf{k},\omega) = -\frac{1}{\pi}\Im\left[\frac{1}{\omega - \tilde{\epsilon_{\mathbf{k}}}/\hbar + i\Gamma(\mathbf{k})}\right] \quad . \tag{10.6}$$

Notice that if the inverse lifetime $\Gamma_{\mathbf{k}} \to 0$, the spectral function reduces to the noninteracting (10.5). (10.6) describes *quasiparticles* with a dispersion curve $\tilde{\epsilon_{\mathbf{k}}}$ and a decay rate (inverse lifetime) $\Gamma_{\mathbf{k}}$.

We must be careful about the chemical potential. If we are in equilibrium (and at T=0), we cannot add fermionic excitation at an energy $\omega < \mu$. So we shall infer that for $\omega > \mu$, (10.6) is the spectral function for particle-like excitations, whereas for $\omega < \mu$ it is the spectral function for *holes*.

If $\Gamma_{\mathbf{k}}$ is small, the quasiparticles are long-lived and have some real meaning. More precisely: a quasiparticle resonance at energy $\hbar\omega = \tilde{\epsilon_{\mathbf{k}}}$ can be resolved, if the peak width in the spectral function is less than the peak's centre position, or, borrowing the terminology developed for damped harmonic oscillators, if the quality factor (ratio of peak position over peak width) is larger than 1. For an interacting system (right panel in Fig. 10.7), A has width $= 2 \times \Gamma_{\mathbf{k}}$. The phase-space argument made in section 10.3.1 showed that $\Gamma_{\mathbf{k}} \propto \epsilon^2$. The quality factor $\tilde{\epsilon}_{\mathbf{k}}/\Gamma \propto \tilde{\epsilon}_{\mathbf{k}}/\tilde{\epsilon}^2_{\mathbf{k}}$ therefore diverges for $\epsilon \to 0$. We find that quasiparticles are overdamped at high energies, but well-defined for $E \to E_F$.

[1] Getting the sign here requires one to adopt a sign convention for Fourier transforms that is opposite to the one often used in maths books. We have also set $\hbar = 1$.

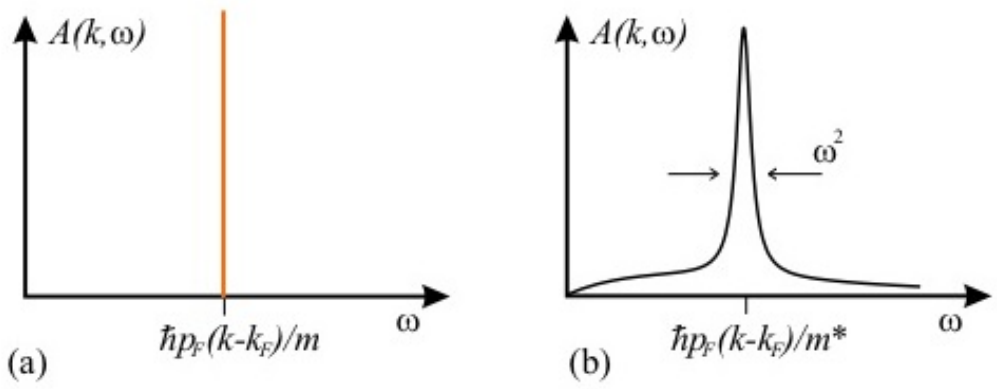


Figure 10.7: Electron spectral functions for Fermi gas (left) and Fermi liquid (right)

This shows that the quasiparticle concept is self-consistent: it is possible to have quasiparticles which scatter so rarely that their lifetime is sufficiently long to observe them. However, this does not guarantee that the Fermi liquid state always exists in every metal. The conditions under which Fermi liquids exist or not is an active field of both experimental and theoretical research.

10.3.4 Tunability of the quasiparticle interaction

The interaction term $\frac{1}{2}\sum_{\mathbf{k}\mathbf{k}'} f(\mathbf{k},\mathbf{k}')n(\mathbf{k})n(\mathbf{k}')$ in the Landau expansion for the total energy causes various quasiparticle properties to be changed with respect to the free electron value. Most importantly, the effective mass m^* can be orders of magnitude larger than the bare electron mass. The quasiparticle interaction function $f(\mathbf{k},\mathbf{k}')$ is completely different from the Coulomb repulsion, which acts on the underlying electrons. In particular, $f(\mathbf{k},\mathbf{k}')$ can be spin-dependent. This is an example of the tunability of correlated electron systems: although the underlying Coulomb interaction is fixed, the effective interaction in the low energy model depends strongly on details of the system, and can therefore be tuned over a wide range by changing, for example, magnetic field, pressure or doping.

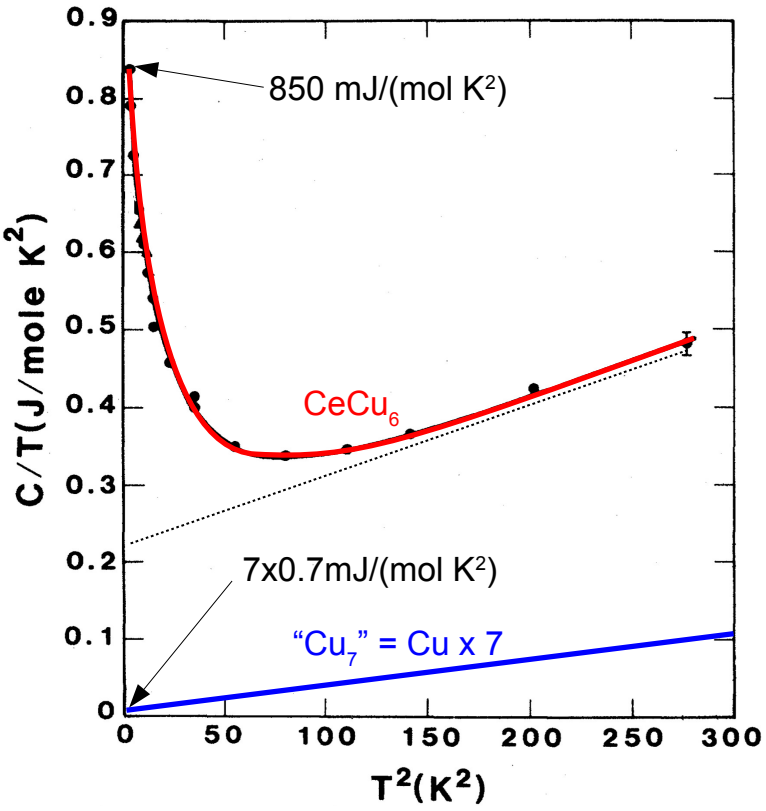


Figure 10.8: Low temperature heat capacity of $CeCu_6$ and of Cu. The molar heat capacity of pure Cu has been multiplied by 7 to allow direct comparison of the heat capacity per atom.

10.4 Fermi liquids at the limit: heavy fermions

10.4.1 What are heavy fermions?

A key result from Fermi liquid theory is the realisation that the effective mass of the quasiparticles in an interacting system can be very different from the band mass, which is determined solely from the band structure of the non-interacting system. Strong interactions can in principle cause high effective masses. Heavy fermion materials have very high Sommerfeld coefficients of the heat capacity $C/T \sim 1$ J/(molK), and high, weakly temperature-dependent magnetic susceptibility. This suggests they follow Fermi liquid theory, but the effective quasiparticle masses are strongly enhanced: in some cases up to 1000 times m_e. Usually, heavy fermion materials contain Cerium, Ytterbium or Uranium, which contribute partially filled f-orbitals to the band structure. These highly localised states are important, because in a lattice, they lead to very narrow bands. The strong Coulomb repulsion prevents double occupancy of these states. There are hundreds of heavy fermion materials. Examples include $CeCu_2Si_2$, $CeCu_6$, $CeCoIn_5$, $YbCu_2Si_2$, UPt_3.

Fig. 10.8 shows the Sommerfeld coefficient of the heat capacity, C/T for the typical heavy fermion material $CeCu_6$. Note that in the low temperature limit, C/T approaches 850 mJ/(molK^2). If we compare this value to that expected for pure Cu, we find that the heat capacity per atom is boosted by a factor of 150. In other words, if we replace 14% of the copper atoms in a sample of pure Cu by Ce, we increase the low temperature heat capacity 150-fold!

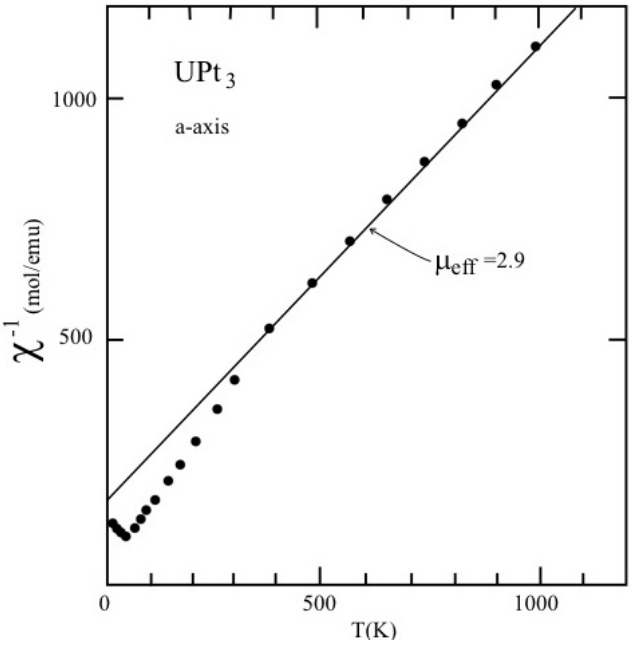


Figure 10.9: Inverse of the magnetic susceptibility χ of UPt_3 plotted versus temperature T. At high temperature, $\chi^{-1} \propto T$, suggesting local-moment behaviour.

The Sommerfeld coefficient of the heat capacity is linked to the electronic density of states at the Fermi level, $g(E_F)$, which in turn is inversely proportional to the Fermi velocity v_F. This is because $v_F = \frac{1}{\hbar} |\nabla_{\mathbf{k}} E|$, and the density of states is obtained by integrating dk/dE over the Fermi surface: $g(E_F) \propto \int dA \frac{1}{|\nabla_{\mathbf{k}} E|}$. On the other hand, the momentum of quasiparticles at the Fermi surface is $\hbar k_F$, but can also be written as $m^* v_F$. As the conduction electron density in $CeCu_6$ is not expected to be very different from that in Cu, we can expect k_F to be similar in both metals. In fact, k_F is of order 0.5-1.0 $Å^{-1}$ in many metals. This value is not affected by the strength of electronic interactions, because the size of the Fermi surface is essentially fixed only by the number of electrons, not by their interactions. We find, then, that the effective quasiparticle mass $m^* \propto g(E_F)$. All else being equal, the quasiparticles in $CeCu_6$ would appear to have a 150 times higher effective mass than those in Cu.

Note also that the high value for C/T in $CeCu_6$ is only reached at the lowest temperatures. Such a strong upturn of C/T at low T is observed in many cases. It suggests that the heavy fermion state develops fully only at low temperature.

10.4.2 High T: "local moments", low T: Fermi liquid

While the behaviour of $CeCu_6$ and other heavy fermion materials at low temperatures is consistent with the key results of Fermi liquid theory, the picture changes dramatically at high temperature (Fig. 10.9).

Recall that an isolated magnetic moment will display a strongly temperature dependent susceptibility $\chi \propto 1/T$ according to the Curie-law. A similar form is observed in many heavy fermion materials at high temperature, often with a slope that is consistent with the Curie constant expected from the electronic configuration of the corresponding magnetic ions (Uranium, Cerium or Ytterbium, typically). This suggests that the partially filled f-orbitals act as local

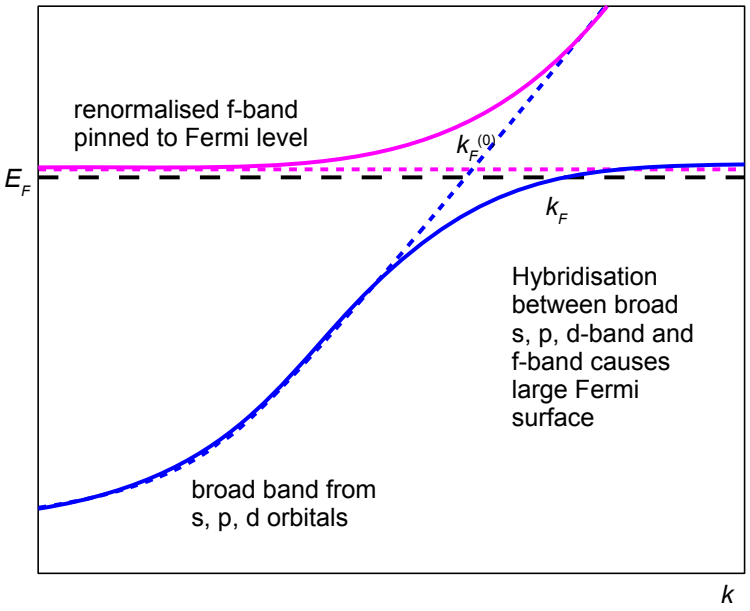


Figure 10.10: Schematic band structure for heavy fermion materials

moments at high temperatures, as if the formation of band states could be ignored.

On the other hand, the Curie form of the susceptibility does not extend all the way to zero temperature, but rather it crosses over the to a constant value in the low temperature limit, just as the Sommerfeld coefficient of the heat capacity did. This suggests that at low temperatures, a local moment picture of these materials is not appropriate, and we must instead consider them as heavy Fermi liquids. It is difficult to reconcile these two ways of thinking about the same material.

10.4.3 Renormalised band picture for heavy fermion systems

A qualitative understanding of the origin of the heavy fermion state can be obtained by considering the hybridisation between the bands associated with the more extended s, p, and d-orbitals on the atoms and the bands which arise from the very tightly localised atomic f-orbitals.

Because a partially filled f-orbital will always lie *inside* filled s, p and even d orbitals with a higher major quantum number, there is negligible hybridisation between f-orbitals on neighbouring atoms – they are just too far apart. This results in a very flat band from the atomic f-states.

If we consider single-electron states naively, then we find that the f-band formed from the atomic f-orbitals is well below the chemical potential, and should therefore be completely full. In such a scheme there would be no local moments at high temperature and no heavy fermion behaviour at low temperature. The scheme fails, because it ignores the strong Coulomb repulsion between electrons sharing the same f-state.

Instead, once a single electron has occupied an f-orbital, the energy cost for a second electron hopping onto the same orbital is very high, almost prohibitive. We can modify our single particle picture to ensure that the f-orbitals have an average occupancy of one, by

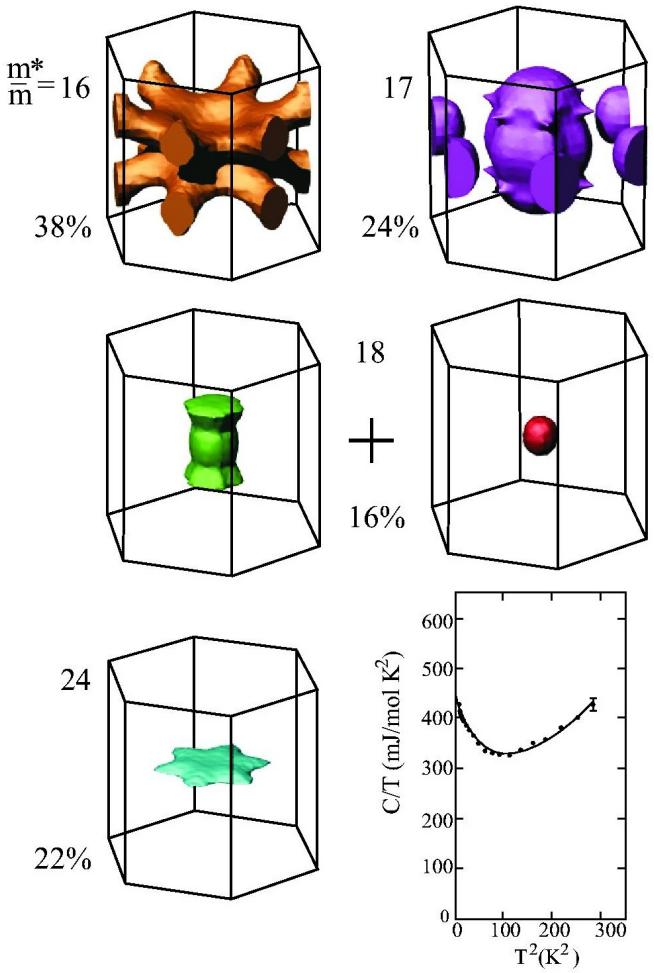


Figure 10.11: Fermi surface sheets of heavy quasiparticles, detected by quantum oscillation measurements in UPt_3.

renormalising (or shifting) the energy of the f-states close to the chemical potential, near E_F. This is the *renormalised band picture*, shown in Fig. 10.10.

Similar to what happened in the band structure of copper in the Lent term problem, the renormalised, narrow f-band and the broad band arising from atomic s, p, and d orbitals hybridise, producing an anticrossing very close to the Fermi level. This causes the new dispersion $E(k)$ to cross E_F at a much reduced slope compared to the broad s, p, d-band. As the slope dE/dk gives the Fermi velocity v_F and is inversely proportional to the effective mass, this scheme can explain the enhanced effective masses observed in materials in which partially occupied f-orbitals are present. Moreover, note that k_F is larger for the hybridised system than it would be without the hybridisation. In fact, the volume of the Fermi surface in heavy fermion systems is large enough to contain not only the electrons on s, p, and d-bands but also the f-electrons.

10.4.4 Quasiparticles detected in de Haas-van Alphen experiments

Direct evidence for the existence of a heavy Fermi liquid state has come from the observation of quantum oscillations (see Lent term section) in a number of heavy fermion materials. Fig. 10.11 shows one of the clearest examples, UPt_3. The volume enclosed by the Fermi surface is a very stringent criterion, which can be used to decide between the heavy Fermi liquid scenario – in which the f-electrons contribute to the Fermi surface – and local moment models. Moreover, the temperature dependence of the signal observed in quantum oscillation measurements can be used to determine the effective mass of the quasiparticles. This can be compared against the measured heat capacity. Where these comparisons were possible, such as in UPt_3, the measured heat capacity was consistent with what would be expected from the measured effective masses.

10.4.5 Heavy fermions, summary:

- Many intermetallic compounds containing elements with partially filled $4f$ (Ce, Yb) or $5f$ (U) orbitals show heavy fermion behaviour.
- Their low temperature properties are consistent with Fermi liquid theory, if we assume very high effective carrier masses.
- What is the role of electrons in partially filled f-orbitals? Do they behave like local moments or like conduction electrons? Where quantum oscillation studies have been successful, they indicate that at low temperature the electrons contribute to the Fermi surface, like in a normal metal. At high temperature, on the other hand, they behave like local moments.
- Because $g(E_F)$ is so high in these materials, they tend to order magnetically or even become superconducting. There are many different ordered low temperature states in these metals, some simple, some very exotic.
- There is an increasing number of materials (e.g., $YbRh_2Si_2$), which do not follow Fermi liquid theory at low T. Are they 'non-Fermi' liquids? Do the f-electrons remain as local moments down to absolute 0 in this case, not contributing to the Fermi surface? This is being investigated at the moment.