Solid State Physics
PHYS 40352

by Mike Godfrey

Spring 2012

Last changed on May 22, 2017

Contents

Preface

This document will eventually be a summary of the material taught in the course. In a few places you may find that derivations and examples of applying the results are not given, or are very much abbreviated. Conversely, it is often convenient to present some material in a different way from in the lectures. A little common sense should therefore be used when reading the notes and a good textbook consulted from time to time.

I would suggest you come back to these notes from time to time, as they are (and are likely to remain) a work in progress. Please let me know if you find any typos or other slips, so I can correct them.

Chapter 1

Crystal structure

In preparation: Much of the material in this chapter has been adapted, with permission, from notes and diagrams made by Monique Henson in 2013.

You are strongly recommended to make sure that you understand and (where appropriate) can solve problems that involve:

- the meaning of the terms *lattice* and *motif* [or *basis*]
- simple cubic, face-centered cubic, and body-centered cubic lattices
- *lattice vectors* and *primitive lattice vectors*; *unit cells* and *primitive unit cells*
- diffraction of X rays by a crystal in terms of the Bragg equation *and* the reciprocal lattice vectors
- the relation between lattice planes and reciprocal lattice vectors
- be sure you know (and can derive) the reciprocal lattices for the simple cubic, FCC, and BCC lattices [these are useful for the kinds of problems that can be set on nearly-free electron theory and X-ray diffraction]
- the "indexing" of X-ray diffraction patterns (i.e., given the Bragg angles θ, find plausible $\mathbf{G}$ vectors, or lattice planes—you can get additional practice from past paper questions)

1.1 Lattice and basis

A fundamental property of a crystalline solid is its *periodicity*: a crystal consists of a regular array of identical "structural units". The structural unit, which is called the *basis* [or *motif*] can be simple, consisting of just one atom (as in sodium or ion), or complex, consisting of two or more atoms (as in diamond or in haemoglobin); see Fig. 1.1.

The positions in space of these structural units define the points of a *lattice*.[1] Although any real crystal has only a finite number of atoms, this number can be very large indeed (10^{23}, say), so that it is often useful to imagine the crystal and its corresponding lattice to be infinite, extending through all space. The environment of every lattice point is identical in all respects, including orientation, so that we can get from one lattice point to any other by a simple translation. A vector connecting any two points of the lattice [and hence a possible translation vector] is called a *lattice vector*, and can be expressed in the form

$$\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3\,, \tag{1.1}$$

where n_1, n_2, n_3 can take any of the integer values $0, \pm1, \pm2, \ldots$.

[1] Usually named after *Bravais*, who made a systematic study [ca. 1845] of the lattices possible in two and three dimensions.

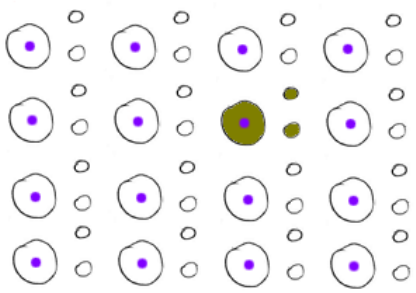

Figure 1.1: An example of a crystal structure. Atoms are represented by grey circles. Three atoms (shaded green) make up the *motif* (or *basis*) of the structure. Lattice points are indicated by blue dots.

Any direction in the lattice can be specified by the coefficients $[n_1, n_2, n_3]$. A set of non-coplanar lattice vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ that can be used to generate all of the lattice vectors in accordance with (1.1) is said to be *primitive*. The choice of these vectors is not unique; in particular, they need not be the shortest possible lattice vectors. This is illustrated in Fig. 1.2.

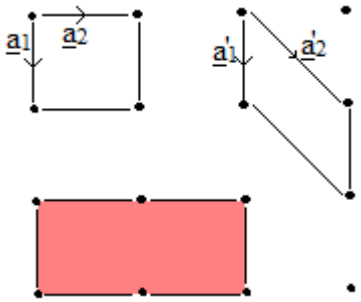

Figure 1.2: Two choices of primitive vectors for a 2D lattice. Primitive unit cells are also shown. The shaded region is a non-primitive cell with twice the area of a primitive cell.

1.1.1 Unit cells

Any region of space that contains only one lattice point and can be translated by lattice vectors $\mathbf{R}$ to fill the whole of space without leaving gaps or forming overlaps is called a *primitive unit cell*. For example, the parallelepiped whose edges are the primitive vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ is always a primitive unit cell.[2] There are many possible choices of unit cell.

In Fig. 1.2 $\mathbf{a}_1$ and $\mathbf{a}_2$ are primitive vectors and the rectangle they span is a primitive unit cell. The choice of vectors is not unique: we could equally well choose $\mathbf{a}_1'$ and $\mathbf{a}_2'$. They are the edges of a primitive cell (a parallelogram) of the same area as the rectangle with edges $\mathbf{a}_1$ and $\mathbf{a}_2$. These are just two simple examples – they don't necessarily have to be that simple.

Consider the shaded rectangle in Fig. 1.2. It is still a unit cell, as it could be repeated to fill the whole of space, but it is not a *primitive* cell as it does not have the smallest possible area. It is not always convenient to work with a primitive unit cell. For example, when discussing the lattices of the cubic system we generally use a unit cell that has the shape of a cube, even though for BCC and FCC this conventional cubic unit cell is non-primitive; see Figs. 1.13 and 1.14 later.

[2] A parallelepiped is illustrated in Fig. 1.11.

1.1.2 Crystal symmetry

Lattice symmetries include translation by a lattice vector, discrete rotations (discussed below), and reflections.

Mirror symmetry (reflections)

Mirror symmetry should be familiar enough not to need discussion.

Rotational symmetry

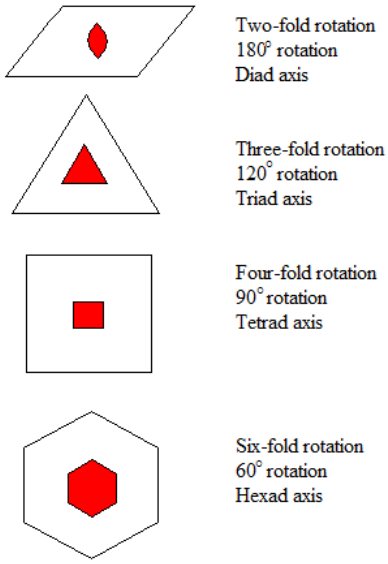

Figure 1.3: Conventional symbols for axes of pure rotational symmetry. These are shown shaded in this figure, as is usual in textbooks of crystallography, but elsewhere in this chapter (and in the lectures) the symbols have been left unfilled.

The notation for axes of rotational symmetry is shown in Fig. 1.3. These are the only four possible rotational symmetries that are consistent with the periodicity of a crystal. Why? Consider an n-fold rotation axis A, in two dimensions, such that a rotation through an angle $\phi = 2\pi/n$ about A maps the crystal onto itself. This is illustrated in Fig. 1.4. Now consider a second n-fold axis, B, that is related to axis A by a lattice translation, which we suppose to be the shortest possible. Let a denote the length AB. After rotation by ϕ anticlockwise about A, B maps onto B$'$. Likewise, when the lattice is rotated clockwise about B, A maps to the point A$'$. The distance between A$'$ and B$'$ must be an integer multiple of a, as a is the length of the shortest lattice vector. Therefore,

$$\mathrm{A'B'} = pa = a - 2a\cos\phi\,, \tag{1.2}$$

where p is an integer. This can be rearranged as

$$\cos\phi = \frac{1-p}{2}\,. \tag{1.3}$$

This gives the possible values of ϕ to be those shown in Table 1.1. Hence, only rotation axes with orders $n = 2$, 3, 4 and 6 are possible in a crystal.

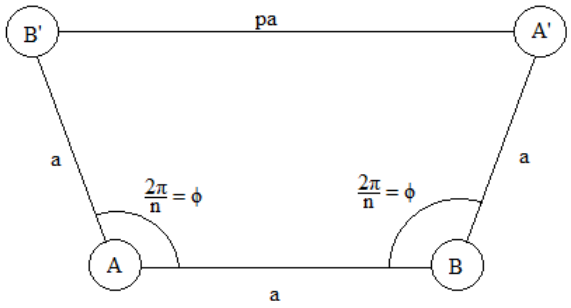

Figure 1.4: Diagram to illustrate the restrictions on the possible orders of rotational symmetry elements in a 2D crystal. A and B are two n-fold axes of rotation, separated by a shortest lattice vector of length a. The rotational symmetry about B and A requires the presence of further axes A′ and B′, and the translational symmetry of the lattice (in the direction parallel to AB) requires their separation A′B′ to be an integer multiple of a.

p	$\cos\phi$	ϕ	n
0	$\frac{1}{2}$	60°	6
1	0	90°	4
2	$-\frac{1}{2}$	120°	3
3	-1	180°	2

Table 1.1: The orders, n, of rotation axes that are consistent with the translational symmetry of a lattice. Refer to Fig. 1.4 for p and $\phi = 360°/n$, and to Eq. (1.3) for the relation between them. Note that if $p = 3$, B′, A, B and A′ all lie on a straight line, giving the maximum possible separation of A′ and B′.

Aside: other symmetries

There are other, more complex, symmetries that may also be considered when classifying crystal structures. For example, rotations may be combined with reflection in a plane perpendicular to the axis of rotation: these combinations are called *rotation–reflection* axes. Translation by a suitable fraction of a lattice vector can be combined with a reflection or a rotation to give the symmetry operations known as *glide* and *screw*, respectively. And it is even possible to combine spatial symmetry operations with *time reversal*[3]: the resulting so-called *anti*-symmetry operations can be helpful in classifying the structure of magnetic materials. None of these more exotic symmetry operations will be discussed further in this course.

1.1.3 Two-dimensional lattices

In two dimensions there are only five lattice types. We list them here, because it can be helpful to regard the three-dimensional lattices as being built up from a stack of 2D lattices. Some of the non-translational symmetries will be mentioned, but it should be understood that these are symmetries *only* of the lattice, regarded as an array of points: the symmetry of a crystal *structure* would also take into account the presence of the *motif*. Our discussion of symmetry is therefore very incomplete.

i. oblique lattice

$a_1 \neq a_2$ and the angle ϕ between the two lattice vectors doesn't take any of the special values listed in Table 1.1; i.e., $\phi \neq 90°, 60°, 120°$. The oblique lattice is shown in Fig. 1.5.

[3] "Time reversal" in practice means reversal of the direction of the microscopic magnetic moments or electric currents in the solid.

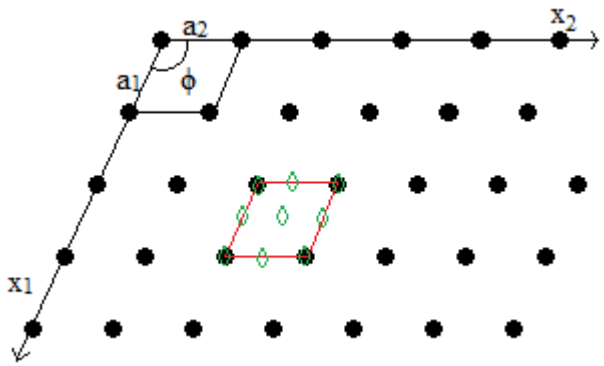

Figure 1.5: An oblique lattice. Axes of 2-fold rotational symmetry [diads] are indicated in green.

An example of a primitive unit cell is the parallelogram shown in red, with the symmetries of it highlighted in green using notation from Fig. 1.3. In an oblique lattice, all of the lattice points are diads. The point at the middle of the cell is also a rotation diad, as are the midpoints of the sides; in fact, there are four inequivalent diad axes. There are no lines of mirror symmetry.

ii. rhombic lattice (or rectangular c lattice)

$a_1 = a_2$ $(= a)$ and $\phi \neq 90°, 60°, 120°$. The primitive cell for a rhombic lattice is the rhombus shown in black in Fig. 1.6, but a conventional non-primitive cell is also shown in red. Note the presence of lines of mirror symmetry, and that rotation diads occur where two mirror lines intersect at right angles – though these are not the *only* diads.

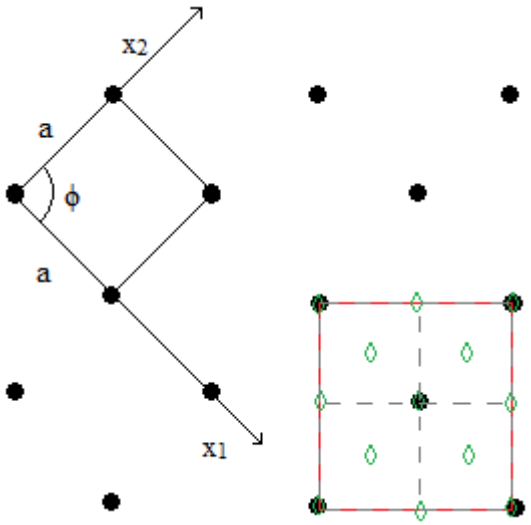

Figure 1.6: A rhombic lattice; $\phi \neq 90°$. Mirror lines are shown as grey, dotted lines. A primitive cell is outlined in black and a non-primitive rectangular c-cell is picked out in red: it contains two lattice points.

The rhombic lattice is also known as the rectangular c lattice. The c is a reminder that there is a second lattice point at the centre of the non-primitive rectangular cell.

iii. rectangular p lattice

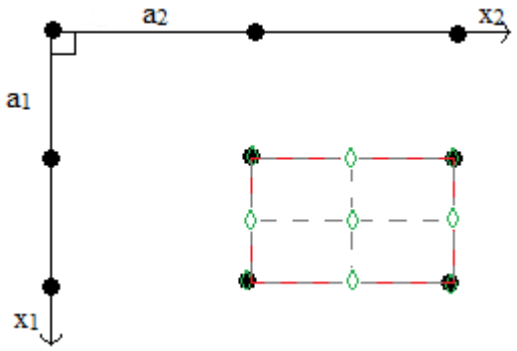

Figure 1.7: A rectangular p lattice. Mirror lines are shown as grey, dotted lines. A primitive cell is picked out in red.

$a_1 \neq a_2$ and $\phi = 90°$. The rectangular p lattice is shown in Fig. 1.7.

iv. square lattice

The square lattice is a special case of the rectangular p lattice with $a_1 = a_2\ (= a)$ and $\phi = 90°$. This is shown in Fig. 1.8. Wherever four mirror lines intersect at 45°, there is a rotation tetrad.

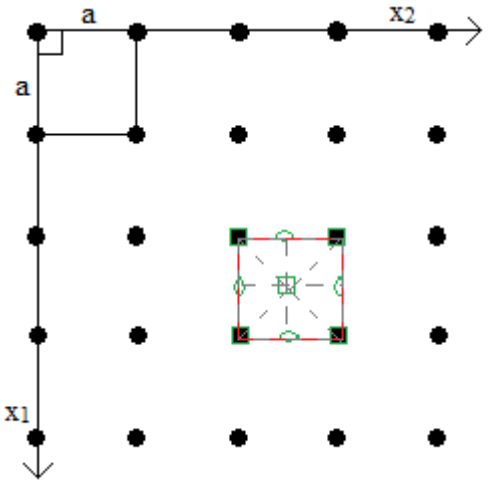

Figure 1.8: A square lattice. Lines of symmetry are shown as grey, dotted lines. In the primitive cell picked out in red, there are tetrads at the corners (where the lattice points are) and at the centre.

The square lattice can also be regarded as a special case of the rhombic lattice with $\phi = 90°$.

v. hexagonal (or triangular) lattice

$a_1 = a_2\ (= a)$ and $\phi = 120°$ (or 60°). The 2D hexagonal lattice is shown in Fig. 1.9. It is a special case of the rhombic lattice.

The symmetry of the hexagonal lattice is better illustrated by choosing a hexagon-shaped unit cell (shown in red in Fig. 1.9), which has three times the area of the primitive unit cell. Hexads are present where mirror lines intersect at 30° and triads where mirror lines intersect at 60°.

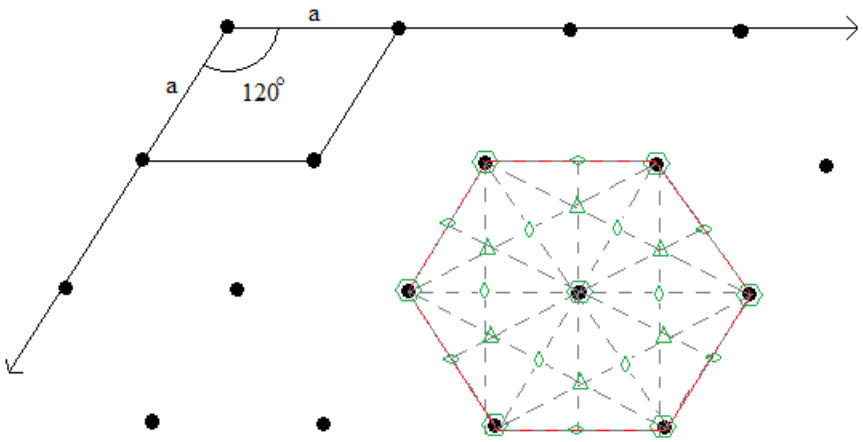

Figure 1.9: A hexagonal lattice. Mirror lines are shown as grey, dotted lines. A hexagon-shaped non-primitive cell is picked out in red: it contains three lattice points.

Exercise 1.1:

When two lines of mirror symmetry intersect, there is always an axis of rotational symmetry passing through the point of intersection, perpendicular to the mirror lines. Explain. On the other hand, rotational symmetry does not require the presence of intersecting mirror lines. Illustrate by giving an example.

Exercise 1.2:

The square lattice is a special case of the rectangular p lattice. Why is there no need to introduce a "square c lattice", analogous to the rectangular c lattice?

A two-dimensional crystal *structure*

The 2D lattices are mainly of interest in describing the planar surfaces of 3D crystals. A notable exception is graphene, which consists of carbon atoms bonded to form a 2D sheet with the structure illustrated in Fig. 1.10. It is a single atomic layer of the graphite structure.

The atoms labelled A all have the same environment (including orientation), and it is easy to see that they lie at the points of a 2D hexagonal lattice. The same can be said of the atoms labelled B, but it should be noted that A and B are *not* related by translation by a lattice vector: their environments are different, being related by a 180° rotation. Thus, if the lattice points of the hexagonal lattice are chosen to be at the A atoms (as in the diagram), the B atoms cannot be at lattice points. The lattice is hexagonal, with a motif consisting of one carbon atom at A and another at B.

This structure is best referred to as the *honeycomb net*, the *honeycomb structure*, or simply the *graphene structure*.

1.1.4 Three-dimensional lattices

There are fourteen three-dimensional Bravais lattices. They are conventionally grouped into seven *lattice systems*.[4]

[4]These are not quite the same as the seven *crystal systems*, in which the classification is based on the point-group symmetry of the crystal structure. Don't worry about this distinction: my primary aim is to illustrate some of the different kinds of *lattice* that a

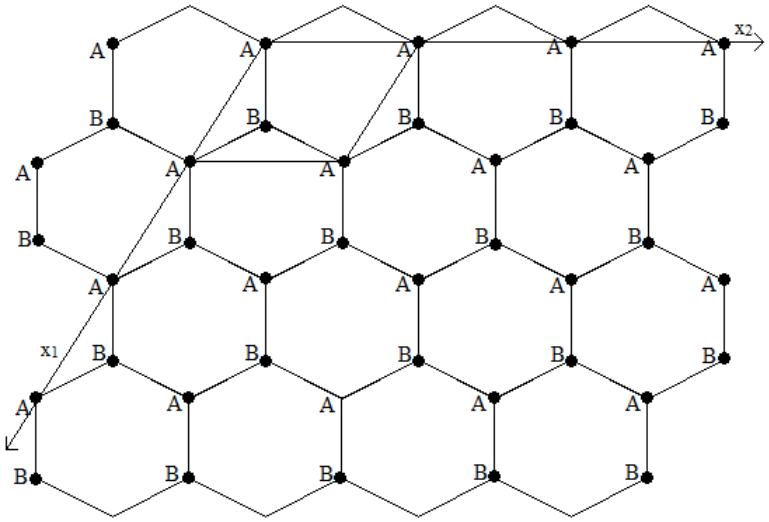

Figure 1.10: The structure of graphene. In this honeycomb net, the bonds around the carbon atoms labelled A and B have different orientations. The honeycomb net is a 2D hexagonal lattice with two atoms per lattice point. In this diagram, the lattice points have been chosen to coincide with the A atoms.

i. triclinic system

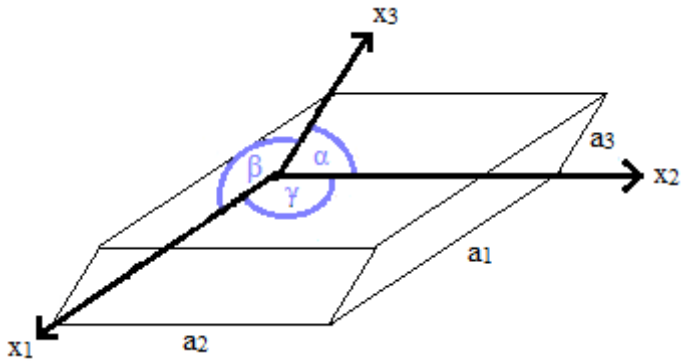

Figure 1.11: A unit cell of the triclinic lattice, in which $\alpha \neq \beta \neq \gamma$ and $a_1 \neq a_2 \neq a_3$. This shape is known as a parallelepiped.

The triclinic system is the least symmetric of all of the lattice types. The three axes, as shown in Fig. 1.11 can be any three non-coplanar lattice directions. There is only one possible lattice in this system. The unit cell is a general parallelepiped. The lattice has only one kind of non-translational symmetry: a centre of inversion symmetry. The triclinic lattice is the only 3D lattice type that has no mirror symmetry.

Once a set of axes x_1, x_2, x_3 has been decided on, the repeat distances along these axes are denoted a_1, a_2, a_3. The angle between axes 2 and 3 is conventionally called α; between axes 1 and 3 the angle is β; and between axes 1 and 2 it is γ. This convention is decidedly awkward,[5] but it will nevertheless be used below to describe the special features of the other lattice systems.

crystal may have, rather than the different kinds of *symmetry*.

[5] It would be more consistent to use $\alpha_1 = \alpha$, $\alpha_2 = \beta$, and $\alpha_3 = \gamma$.

ii. monoclinic system

The monoclinic system has $\alpha = \gamma = 90^\circ$ and $a_1 \neq a_2 \neq a_3$: the x_2 axis is perpendicular to the x_1–x_3 plane. This case retains all of the symmetries of the 2D oblique lattice. In addition, there are mirror planes perpendicular to the axis x_2. There are two distinct monoclinic lattices (P,C).

iii. orthorhombic system

The orthorhombic system has $\alpha = \beta = \gamma = 90^\circ$ and $a_1 \neq a_2 \neq a_3$. The unit cell can be taken to be a general cuboid. There are four lattice types in this system (P,C,I,F).

iv. tetragonal system

The tetragonal system has $\alpha = \beta = \gamma = 90^\circ$ and $a_1 = a_2 \ (= a) \neq a_3$. $a_1 = a_2 = a$ allows for four-fold rotation around the x_3 axis. There are two lattice types in this system (P,I). Plan views of the unit cells of the P and I lattices are shown in Figs. 1.12 and 1.13; the x_3-axis is directed out of the page and heights above the page are given in units of a_3. [Note that the same figures are used to illustrate the unit cells of the cubic P and I lattices. There is *no* tetragonal F lattice: Why not?]

v. cubic system

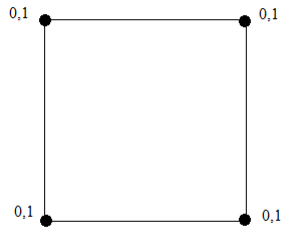

Figure 1.12: P lattice unit cell

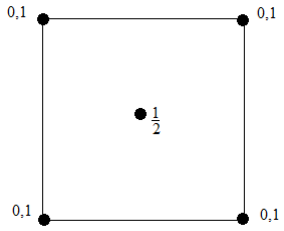

Figure 1.13: I lattice unit cell

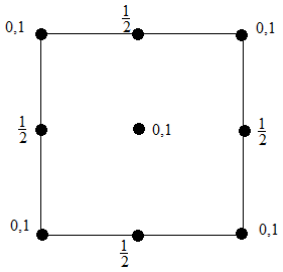

Figure 1.14: F lattice unit cell

In the cubic system, $\alpha = \beta = \gamma = 90^\circ$ and $a_1 = a_2 = a_3 \ (= a)$. One can take the unit cell to be a cube. There are three lattice types in this system, which we illustrate by sketches of the conventional cubic unit cells:

1. The simple cubic lattice, also known as the cubic P lattice, shown in Fig. 1.12.
2. The body-centered cubic lattice, also known as the cubic I lattice, shown in Fig. 1.13. Note that the cell shown in the figure is not a primitive cell: it contains two lattice points.
3. The face-centered cubic lattice, also known as the cubic F lattice, shown in Fig. 1.14. Again, the cell shown in Fig. 1.14 is non-primitive: it contains four lattice points.

The lattices have all the symmetries of the cubic unit cell, which include mirror planes, diads, tetrads, and triads (along the four body-diagonals of the cube).

vi. rhombohedral system

There is one lattice type in this system. A primitive cell of the rhombohedral lattice is shown in Fig. 1.15. The key feature to notice is that $\alpha = \beta = \gamma$ and $a_1 = a_2 = a_3 \ (= a)$. The rhombohedral lattice has three-fold rotational symmetry about an axis inclined at equal angles to x_1, x_2, x_3.

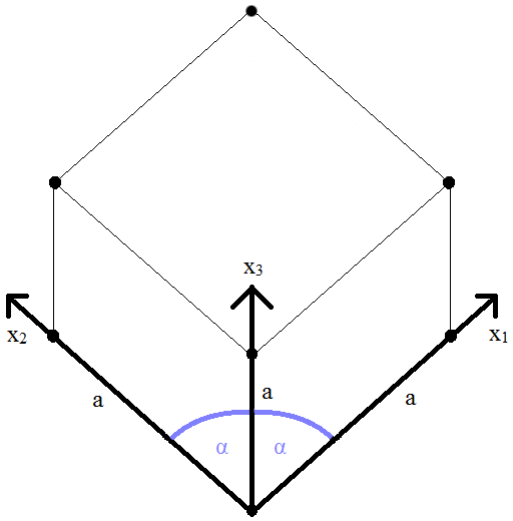

Figure 1.15: Primitive cell of the rhombohedral lattice. Each of the axes x_1, x_2, x_3 is inclined at angle α to the other two. The lattice has 3-fold rotational symmetry about the axis $[1,1,1]$, which is vertical in this diagram.

vii. hexagonal system

There is only one 3D hexagonal lattice type. It consists of 2D hexagonal lattices [as shown in Fig. 1.9] stacked on top of each other, along the x_3-direction. $a_1 = a_2\ (= a) \neq a_3$, $\gamma = 120°$ (or 60°) and $\alpha = \beta = 90°$.

1.1.5 Some cubic crystal *structures*

Section illustrating the caesium chloride [Fig. 1.16], zincblende and diamond structures [Fig. 1.17].

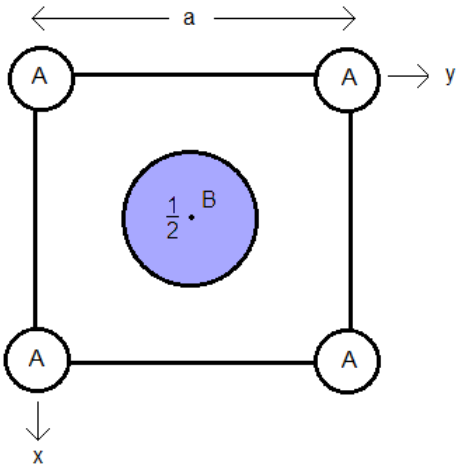

Figure 1.16: Primitive cell of the caesium chloride structure, which has the cubic P lattice. Atoms of type A are at heights 0 and a above the base of the cell and an atom of type B is in the centre, at height $\frac{1}{2}a$. Every atom is surrounded by 8 atoms of the opposite type. If the atoms were all of the *same* type, this would be BCC, and so would have a different Bravais lattice.

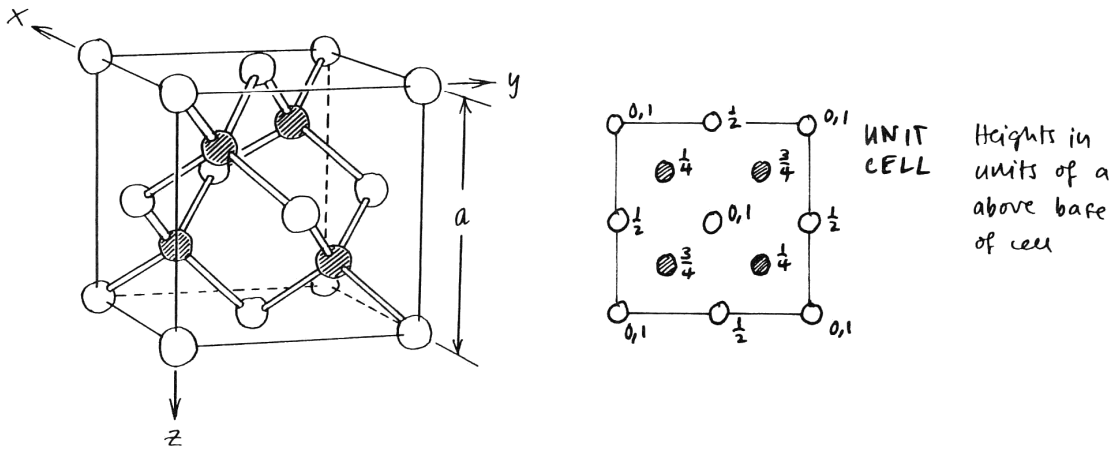

Figure 1.17: Cubic unit cell of the zincblende and diamond structures, which both have the cubic F lattice. In each case there are four lattice points in the cubic unit cell, but the motifs are different. The *zincblende structure* has a motif consisting of two atoms of *different* kinds: unshaded disks represent atoms of one element (e.g., zinc) and shaded disks are another element (e.g., sulphur). Every atom is surrounded by 4 others in a tetrahedral arrangement. In the *diamond structure*, the two atoms of the motif are the *same* element (e.g., carbon).

1.2 X-ray crystallography

1.2.1 Diffraction by a crystal

When X rays are incident on a crystal, most of the energy passes through undeflected, but some is scattered in other directions by the atoms of the crystal. It is normally assumed that this scattering occurs without change of phase. The total amplitude of the diffracted wave is obtained using the principle of superposition: it is the sum of all the waves scattered by the individual atoms. In forming the sum we must take into account the phase that the incident wave has at the atom and also the phase change that occurs in propagating from the atom to the point of observation.

Suppose that the wave is scattered by a single atom at a lattice point $\mathbf{R}$. The scattered wave is detected [e.g., by a photographic plate] at a point $\mathbf{r}$ that is far from the atom. At large distances [large compared with what?] the spherical wave fronts of the scattered wave are almost planar, with a wave vector $\mathbf{k}'$ directed from $\mathbf{R}$ to $\mathbf{r}$. Assuming that the frequency of the wave is unchanged on scattering, the wavelength will also be the same, so that $|\mathbf{k}'| = |\mathbf{k}|$.

The scattered wave can be built up from two factors. The first factor, $\exp[i\mathbf{k}\cdot\mathbf{R}]$, is the phase of the incident wave at the atom.[6] But we must also take into accout the phase change $\mathbf{k}'\cdot(\mathbf{r}-\mathbf{R})$ in propagating from $\mathbf{R}$ to $\mathbf{r}$, so that we have

$$\begin{aligned}\psi_{\rm sc}(\mathbf{r}) &= Ae^{i\mathbf{k}\cdot\mathbf{R}} \times e^{i\mathbf{k}'\cdot(\mathbf{r}-\mathbf{R})} \\ &= Ae^{i\mathbf{k}'\cdot\mathbf{r}} \times e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}}.\end{aligned} \tag{1.4}$$

Here $\psi_{\rm sc}$ stands for, say, the x-component of the electric or magnetic field in the wave. The product of exponentials has been rearranged in the second line. The coefficient A (with $|A| \ll 1$) is just a reminder that only a small part of the incident radiation is scattered by the atom.

[6] Note that we omit a time-dependent factor $\exp[-i\omega t]$ throughout.

Finally we sum the individual waves (1.4) over all the atoms of the crystal to obtain the total scattered wave:

$$\psi_{\rm sc}^{\rm tot}(\mathbf{r}) = Ae^{i\mathbf{k}'\cdot\mathbf{r}} \times \sum_{\mathbf{R}} e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}}. \tag{1.5}$$

The first factor here simply expresses the fact that the scattered wave has wave vector $\mathbf{k}'$, but the second factor

$$\sum_{\mathbf{R}} e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}} \tag{1.6}$$

depends crucially on the particular lattice and it determines whether waves will be scattered strongly in the direction $\mathbf{k}'$. The scattered wave is large only when all of the terms in (1.5) have the same phase. For a given incident wave with wave vector $\mathbf{k}$ this can happen only for a discrete set of outgoing waves with wave vectors $\mathbf{k}'$. For every pair of pair of terms in the sum (1.6) we have

$$e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}_1} = e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}_2}, \quad \text{or} \quad e^{i(\mathbf{k}'-\mathbf{k})\cdot(\mathbf{R}_1-\mathbf{R}_2)} = 1\,. \tag{1.7}$$

But $\mathbf{R}_1 - \mathbf{R}_2$ is just a lattice vector, so that the condition for constructive interference becomes

$$e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}} = 1 \tag{1.8}$$

for every vector $\mathbf{R}$ of the lattice. The solution of this equation will lead us to the concept of a *reciprocal lattice*.

1.2.2 The reciprocal lattice

We need to solve Eq. (1.8), the condition for constructive intereference, to obtain the wave vectors $\mathbf{k}'$. Writing the difference $\mathbf{k}' - \mathbf{k}$ as $\mathbf{G}$, we have

$$e^{i\mathbf{G}\cdot\mathbf{R}} = 1 \quad \text{or} \quad \mathbf{G}\cdot\mathbf{R} = 2\pi P\,, \tag{1.9}$$

where P is an integer [positive, negative or zero], and $\mathbf{R}$ is any lattice vector. In particular, the condition (1.9) must be satisfied for a set of primitive lattice vectors,

$$\mathbf{G}\cdot\mathbf{a}_1 = 2\pi p_1\,, \quad \mathbf{G}\cdot\mathbf{a}_2 = 2\pi p_2\,, \quad \mathbf{G}\cdot\mathbf{a}_3 = 2\pi p_3\,, \tag{1.10}$$

where p_1, p_2, p_3 are all integers. The last three equations, the Laue equations, can be solved for any lattice type.

Before tackling the general case it is helpful first to look at the simpler case of the orthorhombic P lattice, for which we can take the primitive vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ to be orthogonal. In this case, $\mathbf{a}_1 = a_1\hat{\mathbf{x}}$, $\mathbf{a}_2 = a_2\hat{\mathbf{y}}$, $\mathbf{a}_3 = a_3\hat{\mathbf{z}}$, so that the Laue equations (1.10) reduce to

$$G_x = 2\pi p_1/a_1\,, \quad G_y = 2\pi p_2/a_2\,, \quad G_x = 2\pi p_3/a_3\,. \tag{1.11}$$

The solutions

$$\mathbf{G} = 2\pi\,(p_1\hat{\mathbf{x}}/a_1 + p_2\hat{\mathbf{y}}/a_2 + p_3\hat{\mathbf{z}}/a_3) \tag{1.12}$$

define the points of an orthorhombic *reciprocal lattice* with repeat distances $b_1 = 2\pi/a_1$, $b_2 = 2\pi/a_2$, $b_3 = 2\pi/a_3$ along the coordinate axes. Note that these repeat distances are not the same as in the original lattice; in fact, they have the dimensions of reciprocal length. To emphasize this distinction, the lattice of the original crystal structure is often called the *direct* lattice.

General solution of the Laue equations

The general solution of (1.10) for non-orthogonal primitive vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ of the *direct* lattice can be written as

$$\mathbf{G} = p_1\mathbf{b}_1 + p_2\mathbf{b}_2 + p_3\mathbf{b}_3\,, \tag{1.13}$$

where the primitive vectors $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ of the *reciprocal* lattice are given by

$$\mathbf{b}_1 = \frac{2\pi\,\mathbf{a}_2\times\mathbf{a}_3}{\mathbf{a}_1\cdot[\mathbf{a}_2\times\mathbf{a}_3]}\,, \quad \mathbf{b}_2 = \frac{2\pi\,\mathbf{a}_3\times\mathbf{a}_1}{\mathbf{a}_1\cdot[\mathbf{a}_2\times\mathbf{a}_3]}\,, \quad \mathbf{b}_3 = \frac{2\pi\,\mathbf{a}_1\times\mathbf{a}_2}{\mathbf{a}_1\cdot[\mathbf{a}_2\times\mathbf{a}_3]}\,. \tag{1.14}$$

Exercise 1.3:

Verify that a reciprocal vector given by (1.13) and (1.14) satisfies Eq. (1.9) for any $\mathbf{R}$ of the form $n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$, where $P = p_1n_1 + p_2n_2 + p_3n_3$ in Eq. (1.9).

Geometrically, the scalar triple product $\mathbf{a}_1 \cdot [\mathbf{a}_2 \times \mathbf{a}_3]$ appearing in the denominators in Eq. (1.14) represents the volume of a parallelepiped with edges $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$; i.e., it is the volume of the primitive unit cell. The products in the numerators, e.g. $\mathbf{a}_1 \times \mathbf{a}_2$, are the vector areas of the faces of this cell. In magnitude, therefore, the vectors $\mathbf{b}_i$ are 2π times the reciprocal altitudes of the parallelepiped, and they are perpendicular to its faces.

Notation for vectors of the direct and reciprocal lattices

Once a set of primitive vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ has been chosen, vectors of the direct lattice are often specified simply by giving the coefficients of the $\mathbf{a}_i$ in *square* brackets:

$$\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \equiv [n_1, n_2, n_3]\,. \tag{1.15}$$

Similarly, reciprocal lattice vectors are specified by giving the coefficients of the vectors $\mathbf{b}_i$ in *round* brackets:

$$\mathbf{G} = p_1\mathbf{b}_1 + p_2\mathbf{b}_2 + p_3\mathbf{b}_3 \equiv (p_1, p_2, p_3)\,. \tag{1.16}$$

This distinction between square and round brackets is not always observed by physicists, but it is a convention that we shall use throughout this course.

In the above notation, crystallographers like to omit the commas in cases where there would be no ambiguity and to use overbars to indicate negative coefficients. For example, $(1\bar{3}2)$ means exactly the same thing as $(1, -3, 2)$. This convention can save a little space in diagrams and tables.[7]

1.2.3 Reciprocal lattice vectors and lattice planes

There is a very close connection between reciprocal lattice vectors and lattice planes, which is illustrated in Fig. 1.20. We imagine a plane wave $\exp[i\mathbf{G}\cdot\mathbf{r}]$ in a crystal. By the definition (1.9) of reciprocal lattice vectors, this wave would have the same phase $\mathbf{G}\cdot\mathbf{r} = 0$ (modulo 2π) at every lattice point $\mathbf{r} = \mathbf{R}$. Every lattice *plane* perpendicular to $\mathbf{G}$ is therefore a wavefront of the wave $\exp[i\mathbf{G}\cdot\mathbf{r}]$. There may, however, be additional wavefronts that lie between the lattice planes, as shown in the diagram. It follows that the spacing of the lattice planes must be an integer multiple of the wavelength $\lambda_{\mathbf{G}}$ $(= 2\pi/|\mathbf{G}|)$ of this fictitious wave:

$$d = n\lambda_{\mathbf{G}} = n \times \frac{2\pi}{|\mathbf{G}|}\,, \quad \text{where} \quad n = 1, 2, 3, \ldots. \tag{1.17}$$

Let us write $\mathbf{G} = n\mathbf{G}_s$, where $\mathbf{G}_s = (l_1, l_2, l_3)$ is the *shortest* reciprocal lattice vector in the direction of $\mathbf{G} = (p_1, p_2, p_3)$ and n is the greatest common divisor of p_1, p_2, p_3. Then

$$d = \frac{2\pi}{|\mathbf{G}_s|} = \frac{2\pi}{|l_1\mathbf{b}_1 + l_2\mathbf{b}_2 + l_3\mathbf{b}_3|}\,. \tag{1.18}$$

The integers l_1, l_2, l_3 are called the *Miller indices* of the lattice planes perpendicular to $\mathbf{G}$.

Equation (1.18) is a very easy way to calculate the interplanar spacing from the length of the reciprocal lattice vector $\mathbf{G}_s = (l_1, l_2, l_3)$, and it is well worth understanding it fully.

[7] Don't feel obliged to follow the convention of omitting commas and using overbars.

Aside: Alternative definition of the Miller indices

Last year, the Miller indices of a lattice plane were introduced to you in a different way, by using the reciprocals of the plane's intercepts on the coordinate axes. We should check that our approach is equivalent.

Any position vector $\mathbf{r}$ can be written in the form

$$\mathbf{r} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3\,;$$

the expansion is analogous to Eq. (1.1), but the dimensionless coefficients x_1, x_2, x_3 need not be integers. Consider a family of lattice planes specified by the reciprocal lattice vector $\mathbf{G}_s = (l_1, l_2, l_3)$, where the integers l_i have no common divisor greater than 1. The equation of the lattice plane closest to the origin (but not passing through it) is $\mathbf{G}_s \cdot \mathbf{r} = 2\pi$. Then, in terms of the x_i, this becomes

$$\mathbf{G}_s \cdot \mathbf{r} = 2\pi(l_1x_1 + l_2x_2 + l_3x_3) = 2\pi\,.$$

By setting $x_2 = x_3 = 0$ we find that the plane crosses the x_1-axis at $x_1 = 1/l_1$, where the distance is measured in units of a_1. Similarly, the intercepts on the x_2- and x_3-axes are $1/l_2$ and $1/l_3$.

Hence, the Miller indices l_1, l_2, l_3 of a lattice place could be defined as the reciprocals of its intercepts on the axes x_1, x_2, x_3.

1.2.4 The Bragg construction

We have found above that the condition for diffraction is $\mathbf{k}' - \mathbf{k} = \mathbf{G}$, where $\mathbf{G}$ is a reciprocal lattice vector. W. L. Bragg [in 1913] used this result to find an alternative formulation of the diffraction condition.

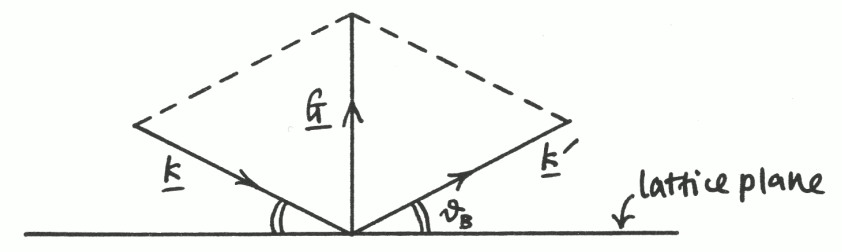

Figure 1.18: Definition of the angle θ_B that appears in Bragg's equation (1.21).

In the vector diagram of $\mathbf{k}$, $\mathbf{k}'$ and $\mathbf{G}$, shown Fig. 1.18, the lengths of $\mathbf{k}$ and $\mathbf{k}'$ are both equal to $2\pi/\lambda$, so the incoming and outgoing wave vectors are inclined at equal angles to the plane normal $\mathbf{G}$. It is as if the diffracted wave had been *reflected* by a set of lattice planes perpendicular to $\mathbf{G}$; for that reason, the process is often called *Bragg reflection*. The angle between the plane and the wave vectors of the incoming and outgoing waves is the so-called *Bragg angle*, and is denoted by θ_B. The angle between the forward directions of the two wave vectors [i.e., the scattering angle] is $2\theta_B$.

Only a few values of θ_B can give rise to diffraction. By inspection of Fig. 1.18,

$$|\mathbf{G}| = 2|\mathbf{k}|\sin\theta_B = \frac{4\pi}{\lambda}\sin\theta_B\,. \tag{1.19}$$

From Eq. (1.17) we also have

$$|\mathbf{G}| = \frac{2\pi n}{d} \quad \text{with} \quad n = 1, 2, 3, \ldots, \tag{1.20}$$

where d is the spacing of lattice planes perpendicular to $\mathbf{G}$. By equating $|\mathbf{G}|$ from the last two equations we find

$$n\lambda = 2d\sin\theta_B\,, \quad n = 1, 2, 3, \ldots, \tag{1.21}$$

which is *Bragg's equation*; it is fully equivalent to the vector condition $\mathbf{k}' - \mathbf{k} = \mathbf{G}$. Since the left-hand side of (1.21) cannot be less than λ and the sine function cannot be greater than unity, Bragg's equation shows that for diffraction to occur, the wavelength λ must be less than twice the spacing between lattice planes.

Bragg's equation is often interpreted [and derived, sort of] as the condition for constructive interference of waves "reflected" by neighbouring lattice planes: the path difference for the two waves is a multiple n of the wavelength. In this context, n is often called the *order of diffraction*, by analogy with diffraction by a ruled grating or a pair of Young's slits.

As an example we consider a simple cubic solid with lattice parameter a. A diffracted wave corresponding to $\mathbf{G} = (222) = 2 \cdot (111)$ is a second-order ($n = 2$) reflection from the (111) planes of the direct lattice. These planes are perpendicular to the $[111]$ direction, with spacing $d_{111} = a/\sqrt{3}$. [Check this.] It would normally be called the (222) reflection, with or without the round brackets, as this tells us both the order of reflection and the planes involved. Similarly, (630) would denote the third-order reflection from the (210) planes (spacing $d_{210} = a/\sqrt{5}$).

In practice, it is not very easy to apply Bragg's equation without having recourse to the reciprocal lattice: it is hard to visualize all the sets of planes and the geometrical relationships between them, and the interplanar spacings are, in any case, most easily calculated using Eq. (1.18). Once the initial conceptual problems have been overcome, it is often simpler to use $\mathbf{k}' - \mathbf{k} = \mathbf{G}$ directly, since this only requires us to visualize the points of the reciprocal lattice.

1.2.5 Structure factor

In Sec. 1.2.1 we obtained the amplitude for a a wave diffracted by a crystal in the case where there is only one atom per lattice point. How are the results changed if the structural motif consists of more than one atom?

If the motif contains M atoms labelled $j = 1, 2, \ldots, M$, the atoms associated with a particular lattice point $\mathbf{R}$ can be taken to have positions $\mathbf{R} + \mathbf{r}_j$. For example, if $\mathbf{R}$ is taken to be at one corner of the unit cell, the vectors $\mathbf{r}_j$ are the positions of the atoms within the unit cell, measured relative to that corner.

The sum in (1.5) can be generalized to

$$\begin{aligned}\psi_{\mathrm{sc}}^{\mathrm{tot}}(\mathbf{r}) &= A e^{i\mathbf{k}'\cdot\mathbf{r}} \times \sum_{\mathbf{R}} \Big\{ \sum_{j=1}^{M} f_j \, e^{-i(\mathbf{k}'-\mathbf{k})\cdot(\mathbf{R}+\mathbf{r}_j)} \Big\} \\ &= A e^{i\mathbf{k}'\cdot\mathbf{r}} \times \sum_{\mathbf{R}} e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}} \times \sum_{j=1}^{M} f_j \, e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{r}_j}, \end{aligned} \tag{1.22}$$

where the factors f_j (discussed shortly) take account of the fact that the atoms need not be all of the same kind, and so may scatter differently. The second factor in (1.22), involving a sum over lattice points $\mathbf{R}$, has been discussed following (1.6); it is large only if $\mathbf{k}' - \mathbf{k}$ is a reciprocal lattice vector. Accordingly, the last factor, involving the sum over j, needs to be considered only for those special values $\mathbf{k}' - \mathbf{k} = \mathbf{G}$:

$$S(\mathbf{G}) = \sum_{j=1}^{M} f_j \, e^{-i\mathbf{G}\cdot\mathbf{r}_j}; \tag{1.23}$$

$S(\mathbf{G})$ is known as the *structure factor*. Equation (1.22) is an expression for the (complex) amplitude of the wave scattered by the crystal; the X-ray intensity is proportional to the *square* of the amplitude, so that the various diffracted waves will have intensities proportional to $|S(\mathbf{G})|^2$.

The atomic form factors f_j, which determine the strength of scattering by individual atoms, are given by

$$f_j = \int_{\text{atom } j} n_j(\mathbf{r}) \, e^{-i\mathbf{G}\cdot\mathbf{r}} \, d^3\mathbf{r}, \tag{1.24}$$

where $n_j(\mathbf{r})$ is the number density of electrons in atom j and the origin of coordinates for the integration is taken to be the centre of the atom. Atomic form factors decrease with increasing $|\mathbf{G}|$, so that diffraction tends to be strongest for small values of the indices p_1, p_2, p_3.

Structure factor for the caesium chloride structure

A unit cell of the structure is illustrated in Fig. 1.16. There are atoms A and B at $\mathbf{r}_A = [0,0,0]a$ and $\mathbf{r}_B = [\frac{1}{2},\frac{1}{2},\frac{1}{2}]a$ in the simple-cubic unit cell. The reciprocal lattice vectors are $\mathbf{G} = (p_1,p_2,p_3)2\pi/a$, where the $\{p_i\}$ are all integers. Hence, the structure factor is given by

$$\begin{aligned} S(\mathbf{G}) &= f_A e^{-i\mathbf{G}\cdot\mathbf{r}_A} + f_B e^{-i\mathbf{G}\cdot\mathbf{r}_B} \\ &= f_A + f_B \exp\left[-i(2\pi/a)(p_1,p_2,p_3)\cdot[\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2}]a\right] \\ &= f_A + f_B \exp[-i\pi(p_1+p_2+p_3)]\,. \end{aligned}$$

Bearing in mind that $\exp[\pm i\pi] = -1$, we find

$$S(\mathbf{G}) = f_A + (-1)^{p_1+p_2+p_3} f_B = \begin{cases} f_A + f_B & \text{if } p_1+p_2+p_3 \text{ is even} \\ f_A - f_B & \text{if } p_1+p_2+p_3 \text{ is odd.} \end{cases} \tag{1.25}$$

The intensities of the diffracted waves are proportional to $|S(\mathbf{G})|^2$, so one expects to see diffraction spots with two different intensities in the ratio $|f_A + f_B|^2 : |f_A - f_B|^2$, depending on the values of p_1, p_2, p_3.

Structure factor of the monatomic BCC structure

The monatomic BCC structure is simply the cubic I lattice [see Fig. 1.13] with one atom per lattice point; the metals sodium and iron, for example, have this structure at room temperature.

It is often convenient to describe cubic structures such as the BCC structure *as if* they were simple cubic with a motif consisting of more than one atom. As noted in the caption to Fig. 1.16, BCC can be thought of as a special case of the caesium chloride structure in which the atoms A and B are identical. We have already done the work of calculating the structure factor for caesium chloride, and to obtain the structure factor for BCC we simply set $f_A = f_B \equiv f$ in Eq. (1.25):

$$S(\mathbf{G}) = \begin{cases} f + f = 2f & \text{if } p_1+p_2+p_3 \text{ is even} \\ f - f = 0 & \text{if } p_1+p_2+p_3 \text{ is odd.} \end{cases} \tag{1.26}$$

Note that there is no diffraction when $p_1+p_2+p_3$ is odd; the physical reason is a destructive interference between the waves scattered by the atoms at $[0,0,0]a$ and $[\frac{1}{2},\frac{1}{2},\frac{1}{2}]a$.

The "allowed" values of $\mathbf{G}$ are, of course, the vectors of the lattice reciprocal to the BCC lattice. You should be able to show that the allowed reciprocal lattice vectors lie on the points of an FCC lattice in reciprocal space, and that this FCC lattice has conventional cubic lattice constant $4\pi/a$. [This is shown in a different way in Examples 1, Q.5, where you use Eqs. (1.14) to calculate a set of primitive vectors for the reciprocal lattice.]

Exercise 1.4:

Use a similar method to calculate the structure factor of the monatomic FCC structure, regarding it as simple cubic with a motif consisting of identical atoms at $[0,0,0]a$, $[0,\frac{1}{2},\frac{1}{2}]a$, $[\frac{1}{2},0,\frac{1}{2}]a$ and $[\frac{1}{2},\frac{1}{2},0]a$. You should find that the structure factor is

$$S(\mathbf{G}) = f\left\{1 + (-1)^{p_2+p_3} + (-1)^{p_3+p_1} + (-1)^{p_1+p_2}\right\},$$

which equals $4f$ if p_1, p_2, p_3 are either all even or all odd integers; in other cases $S(\mathbf{G}) = 0$. Here the allowed values of $\mathbf{G}$ lie on the points of a BCC lattice, which is the reciprocal of the FCC lattice.

[Alternative derivations of the reciprocal lattice for FCC are in your solution to Examples 1, Q.5 and in Sec. 2.4.1 of these notes.]

1.2.6 Further geometry of diffraction

An incident wave $\exp[i\mathbf{k}\cdot\mathbf{r}]$ is diffracted by a crystal only if there is a solution of $\mathbf{k}' - \mathbf{k} = \mathbf{G}$, where $\mathbf{G}$ is a reciprocal lattice vector and $|\mathbf{k}'|$ is restricted to have the same value $2\pi/\lambda$ as the incident wave.

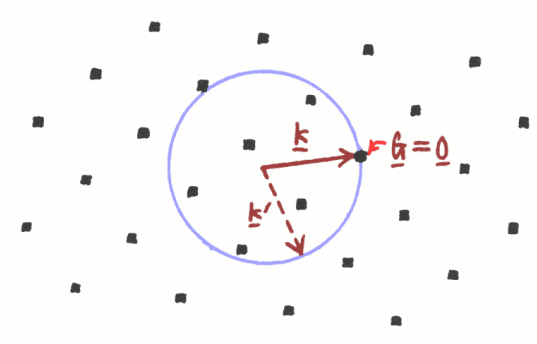

Figure 1.19: For an arbitrary incident $\mathbf{k}$, there are unlikely to be any solutions of $\mathbf{k}' - \mathbf{k} = \mathbf{G}$, apart from the trivial one, $\mathbf{k}' - \mathbf{k} = \mathbf{0}$. Typically, only the point $\mathbf{G} = \mathbf{0}$ will lie on the blue sphere of radius $2\pi/\lambda$.

In Fig. 1.19, the black dots are points of the reciprocal lattice and the vector $\mathbf{k}$ has been taken to end on one of these, which we take to be $\mathbf{G} = \mathbf{0}$. A vector $\mathbf{k}'$ with $|\mathbf{k}'| = 2\pi/\lambda$ is shown as a dashed line; it has the same root as $\mathbf{k}$ and its head must lie on the surface of a sphere of radius $2\pi/\lambda$. Diffraction can occur only if a reciprocal lattice point lies on this surface, which is called the *Ewald sphere*. Since a surface is two-dimensional, it is actually very unlikely that it will pass through any of the reciprocal lattice points, apart from $\mathbf{G} = \mathbf{0}$. To improve the chances of observing a reflection we must therefore relax some of the conditions. This can be done by varying $|\mathbf{k}|$, i.e., by using a non-monochromatic source of X rays, or by varying the orientation of the crystal [and hence the reciprocal lattice] with respect to the beam. We consider each of these possibilities.

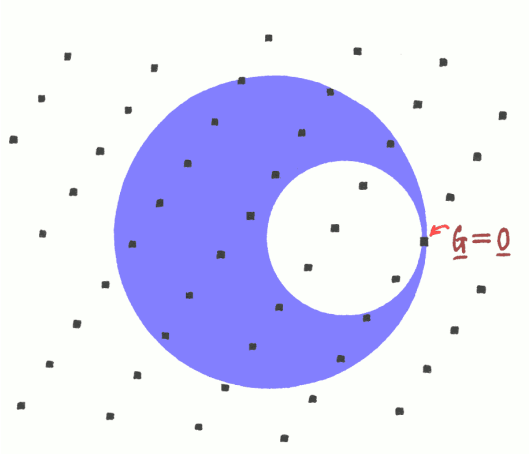

Figure 1.20: If the X rays have a range of wavelengths $\lambda_1 < \lambda < \lambda_2$, each of the reciprocal lattice points $\mathbf{G}$ that fall in the shaded region will give a possible wave vector for a diffracted wave, $\mathbf{k}' = \mathbf{k} + \mathbf{G}$.

Laue method:

The crystal and the beam of X rays have fixed orientations, but a range of wavelengths $\lambda_1 < \lambda < \lambda_2$ is used. To modify Fig. 1.19 to account for this, we suppose that spherical surfaces are drawn with all

possible radii in the range $2\pi/\lambda_2$ to $2\pi/\lambda_1$; these surfaces fill the shaded region shown in Fig. 1.20. Every reciprocal lattice point in the shaded region corresponds to an X-ray diffraction that is allowed for some value of the wavelength in the range λ_1 to λ_2..

The diffracted X rays are intercepted by a flat photographic plate behind the crystal. Increasing the range of λ increases the number of observed reflections, but it also increases the complexity of the diffraction pattern. In practice the Laue method is most useful when the orientation of the crystal and some details of its crystal structure are already know, which is a fairly common situation.

Oscillation photography:

We have very little to say about this technique: the crystal is simply made to oscillate about a fixed axis. In Fig. 1.19 this corresponds to rotating the reciprocal lattice to and fro about an axis passing through $\mathbf{G} = \mathbf{0}$. Diffraction is observed whenever a reciprocal lattice point passes through the spherical surface shown in that diagram. There are several variations on this idea, including *rotation photography*, in which the crystal rotates through 360°. [Indicate why this might be useful?]

Debye–Scherrer (powder) method:

A monochromatic source of X rays is used, and the sample is in the form of a powder containing crystals with many different orientations. All reflections with $|\mathbf{G}| = |\mathbf{k}' - \mathbf{k}| < 4\pi/\lambda$ will be observed. Because crystals are present in all orientations, each diffracted ray gives rise now to a *cone* of diffracted rays with semi-vertex-angle $2\theta_B$, the scattering angle: see Fig. . Measuring these angles gives the lengths of the reciprocal lattice vectors shorter than $4\pi/\lambda$; or, equivalently, the spacing of all lattice planes with $d > \frac{1}{2}\lambda$.

The Debye–Scherrer method is particularly useful for finding out the basic structure of the solid, including the dimensions of the unit cell.

Chapter 2

Electrons in crystals

2.1 Summary of free-electron theory, etc.

Not yet formatted for LaTeX. Your notes from PHYS20252, PHYS30151 and J. Pearson's notes, available online, contain all the material needed on free electrons.

To cut down the number of broken cross-references within this document, we note here that the radius of the Fermi sphere (or Fermi wave vector) k_F and the electron number density n are related by

$$k_F{}^3 = 3\pi^2 n \quad \text{or} \quad n = \frac{k_F{}^3}{3\pi^2}\,. \tag{2.1}$$

2.2 Electrons in a periodic potential

We suppose that an electron in a perfect crystal moves in a spatially periodic field of force due to the ions and the averaged effect of all the electrons. This is an idealization because the Coulomb repulsion between electrons tends to keep them apart (their motion is *correlated*); nevertheless, it is still the starting point for understanding the behaviour of electrons in solids and it is remarkably successful in practice.

2.2.1 Bloch's theorem

Electrons moving in a periodic potential $V(\mathbf{r})$ are often called *Bloch electrons*. Their wave functions obey the Schrödinger equation

$$-\frac{\hbar^2}{2m}\nabla^2\psi_i(\mathbf{r}) + V(\mathbf{r})\psi_i(\mathbf{r}) = E_i\psi_i(\mathbf{r}). \tag{2.2}$$

Some of the most important and most general properties of the solutions $\psi_i(\mathbf{r})$ of (2.2) depend only on the periodicity of the potential: $V(\mathbf{r}) = V(\mathbf{r}+\mathbf{R})$, where $\mathbf{R}$ is any vector (a *lattice vector*) connecting similar points of the crystal lattice.

First we notice that if $\psi_i(\mathbf{r})$ is a solution of (2.2) having energy E_i, then so too is $\psi_i(\mathbf{r}+\mathbf{R})$: the two functions are related by a constant factor, $\psi_i(\mathbf{r}+\mathbf{R}) = c(\mathbf{R})\psi_i(\mathbf{r})$. The modulus $|c(\mathbf{R})|$ will be unity since the probability $|\psi_i(\mathbf{r})|^2$ of finding the electron in the neighbourhood of $\mathbf{r}$ should be the same as the probability $|\psi_i(\mathbf{r}+\mathbf{R})|^2$ for finding it near $\mathbf{r}+\mathbf{R}$, where the local environment is identical. The factor $c(\mathbf{R})$ will also obey $c(\mathbf{R}+\mathbf{R}') = c(\mathbf{R})\,c(\mathbf{R}')$, because translation by $\mathbf{R}+\mathbf{R}'$ is identical to successive translations by lattice vectors $\mathbf{R}$ and $\mathbf{R}'$. Only a complex exponential function of $\mathbf{R}$ has both properties: the dependence of the wave function on $\mathbf{R}$ reduces to a factor $\exp[i\mathbf{k}\cdot\mathbf{R}]$,

$$\psi_{\mathbf{k}}(\mathbf{r}+\mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{R}}\psi_{\mathbf{k}}(\mathbf{r}). \tag{2.3}$$

This is *Bloch's theorem*.[1] It enables us to write the wave function in the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}), \tag{2.4}$$

where $u_{\mathbf{k}}(\mathbf{r})$ is a periodic function of $\mathbf{r}$, $u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r}+\mathbf{R})$.

Exercise 2.1:

Show that the results (2.3) and (2.4) are equivalent.

The property (2.3) is also possessed by the free-electron wave function

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

where $\hbar\mathbf{k}$ is the momentum of the electron. Although for an electron in a lattice the wave function is not a plane wave and the momentum is not conserved (it cannot be, because the electron is subject to forces due to the ions and the other electrons), it is worth noting that the electron is still characterized by a constant vector $\mathbf{k}$: $\hbar\mathbf{k}$ is called the *crystal momentum* of the electron.

Fourier series in three dimensions

The potential $V(\mathbf{r})$ and the function $u_{\mathbf{k}}(\mathbf{r})$ appearing in (2.4) are both periodic in space: $V(\mathbf{r}+\mathbf{R}) = V(\mathbf{r})$, where $\mathbf{R}$ is a lattice vector. In PHYS20171, you learned that periodic functions of *one* variable can be expanded in a Fourier series. The complex exponential functions $\exp[2\pi i n x/a]$ have period a and can be used as a basis for expanding a periodic function $f(x) = f(x+a)$:

$$f(x) = \sum_{n=-\infty}^{\infty} c_n\, e^{2\pi i n x/a}, \qquad \text{where} \qquad c_n = \frac{1}{a}\int_0^a e^{-2\pi i n x/a}\, f(x)\, dx. \tag{2.5}$$

We can do the same for a periodic function in three dimensions. We have already discussed the plane waves that have the same periodicity as the crystal lattice. They are the functions $\exp[i\mathbf{G}\cdot\mathbf{r}]$, where $\mathbf{G}$ is a reciprocal lattice vector. By analogy with (2.5) we can write

$$V(\mathbf{r}) = \sum_{\mathbf{G}} V_{\mathbf{G}}\, e^{i\mathbf{G}\cdot\mathbf{r}}, \qquad \text{where} \qquad V_{\mathbf{G}} = \frac{1}{v_{\text{cell}}}\int_{\text{cell}} e^{-i\mathbf{G}\cdot\mathbf{r}}\, V(\mathbf{r})\, d^3\mathbf{r}, \tag{2.6}$$

where the sum over $\mathbf{G}$ in the Fourier representation of $V(\mathbf{r})$ includes all reciprocal lattice vectors, and the formula for the Fourier coefficient (which we are unlikely to use in this course) involves an integral over a primitive unit cell of the crystal lattice.

If $f(x)$ is a real function of x, the coefficients c_n may still be complex, but satisfy the symmetry property ${c_n}^* = c_{-n}$, which follows directly from the formula used to calculate them. Similarly, it should be easy for you to see that ${V_{\mathbf{G}}}^* = V_{-\mathbf{G}}$.

Representation of the wave function in terms of plane waves

By using the results described in the preceding section, we can write the periodic function $u_{\mathbf{k}}(\mathbf{r})$ as a Fourier series,

$$u_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} C_{\mathbf{G}}\, e^{i\mathbf{G}\cdot\mathbf{r}},$$

so that $\psi_{\mathbf{k}}(\mathbf{r})$ can be written

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}\, u_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} C_{\mathbf{G}}\, e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}. \tag{2.7}$$

[1] We have argued here only for the plausibility of (2.3). The books by Kittel and Ashcroft and Mermin give proofs based on the Fourier representations of $\psi_{\mathbf{k}}(\mathbf{r})$ and $V(\mathbf{r})$ discussed later in this section.

This last expression shows explicitly that $\psi_{\mathbf{k}}$ is sum of *many* plane waves, and so is not a momentum eigenstate; this confirms that the crystal momentum $\hbar\mathbf{k}$ is not the momentum of the electron. In fact, a measurement of the momentum of the electron could give any of the results $\hbar(\mathbf{k}+\mathbf{G})$ and the squared coefficients $|C_{\mathbf{G}}|^2$ give the relative probabilities of obtaining these different results.

Nevertheless, a measurement of the momentum cannot give just *any* value. A completely general function, satisfying periodic boundary conditions at the surface of a crystal of volume L^3, can be expanded as a Fourier series

$$\psi(\mathbf{r}) = \sum_{\mathbf{q}} F_{\mathbf{q}}\, e^{i\mathbf{q}\cdot\mathbf{r}}, \qquad \text{where} \qquad F_{\mathbf{q}} = \frac{1}{L^3}\int_{\text{crystal}} e^{-i\mathbf{q}\cdot\mathbf{r}}\,\psi(\mathbf{r})\,d^3\mathbf{r}\,. \tag{2.8}$$

Comparison of (2.7) and (2.8) shows that $\psi_{\mathbf{k}}(\mathbf{r})$ contains only a tiny fraction of the possible waves with wave vector $\mathbf{q}$; namely, the waves with $\mathbf{q} = \mathbf{k}+\mathbf{G}$. For consistency with the notation used in (2.8) we shall in future write the expansion of $\psi_{\mathbf{k}}$ in the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} F_{\mathbf{k}+\mathbf{G}}\, e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}. \tag{2.9}$$

The form of the last expression makes it clear that $\psi_{\mathbf{k}}$ is identical to $\psi_{\mathbf{k}+\mathbf{K}}$, where $\mathbf{K}$ is any reciprocal lattice vector. Because $\psi_{\mathbf{k}}$ has the periodicity of the reciprocal lattice, the same is true for any observable property of the electron, such as its energy, $E(\mathbf{k}) = E(\mathbf{k}+\mathbf{K})$.

Exercise 2.2:

> The periodicity of $\psi_{\mathbf{k}}$, regarded as a function of $\mathbf{k}$, follows from the fact that the summand in (2.9) depends only on $\mathbf{k}+\mathbf{G}$. If the periodicity is *not* obvious to you, prove it by replacing $\mathbf{k}$ by $\mathbf{k}+\mathbf{K}$ on both sides of (2.9). Change the variable of summation to $\mathbf{G}' = \mathbf{G}+\mathbf{K}$, and note that a sum over all $\mathbf{G}$ of the reciprocal lattice is equivalent to a sum over all $\mathbf{G}'$.

2.2.2 Brillouin zones

Given that any physical property of an electron has the periodicity of the reciprocal lattice, it is only ever necessary to plot the energy in a primitive unit cell of reciprocal space. Any primitive unit cell would do for this, but a conventional choice of unit cell is the so-called *first Brillouin zone*, which is the set of vectors $\mathbf{k}$ that are closer to $\mathbf{k} = \mathbf{0}$ than to any other reciprocal lattice point $\mathbf{G} \neq \mathbf{0}$.

For example, consider a crystal with a simple-cubic lattice, with lattice constant a. In this case, the reciprocal lattice is also a simple-cubic lattice with lattice constant $2\pi/a$. The first Brillouin zone is then a cube of side $2\pi/a$ whose faces are planes bisecting (at right-angles) the shortest reciprocal lattice vectors $(100)2\pi/a$, $(010)2\pi/a$ and $(001)2\pi/a$.

Second, third, and nth Brillouin zones

The perpendicular bisectors of the reciprocal lattice vectors $\mathbf{G} \neq \mathbf{0}$ are sometimes called *zone boundaries* or *Bragg planes*.[2] Given the concept of a zone boundary, the first Brillouin zone could be defined as the set of points $\mathbf{k}$ that can be reached from $\mathbf{k} = \mathbf{0}$ without crossing a zone boundary.

Higher Brillouin zones can be defined in a similar way. The second Brillouin zone is the set of $\mathbf{k}$s that can be reached from $\mathbf{k} = \mathbf{0}$ by crossing exactly one zone boundary, and the third Brillouin zone is the set that can be reached by crossing *two*—and no fewer than two—zone boundaries. Each of these zones is a primitive unit cell of reciprocal space. Generalizing this idea, the nth Brillouin zone is the set of $\mathbf{k}$s that can be reached from $\mathbf{k} = \mathbf{0}$ by crossing $n-1$ (and no fewer than $n-1$) zone boundaries. As n increases, the shapes of the zones become increasingly complex; this is illustrated in Figure 2.1 for the case of a two-dimensional square lattice.

[2] The term "Bragg plane" is avoided in these notes because there is a chance of confusing it with the *lattice planes* responsible for Bragg reflection. Zone boundaries (and Bragg planes) are planes in *reciprocal* space; lattice planes, of course, are planes of lattice points in *direct* space, i.e., in $\mathbf{r}$-space.

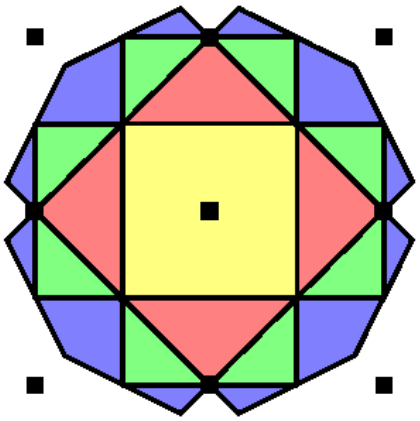

Figure 2.1: First four Brillouin zones for the two-dimensional square lattice. Solid squares represent reciprocal lattice points and solid lines are zone boundaries. Areas of a single colour are different parts of the same Brillouin zone; for example, the central square (yellow) is the first Brillouin zone, and the various blue polygons are all parts of the fourth Brillouin zone. The Brillouin zones all have the same area $(2\pi/a)^2$ in 2D.

As far as this author knows, the only practical use for these higher zones is in relating the features of energy-band structures of real materials to the dispersion relation $E^0(\mathbf{k}) = \hbar^2 k^2/2m$ of the free-electron model. This, of course, is an important application, and we shall see some examples below.

Number of states in a Brillouin zone

At this point we can work out the number of distinct values of $\mathbf{k}$ in a Brillouin zone, from which we can obtain the number of states in an energy band. The volume per state is defined by the boundary conditions on the electron wave functions at the surfaces of the crystal. For periodic boundary conditions, the result (as for free electrons) is $(2\pi)^3/V$, where V is the volume of the crystal. Hence the number of values of $\mathbf{k}$ falling in the simple cubic Brillouin zone of volume $(2\pi/a)^3$ is

$$\frac{(2\pi/a)^3}{(2\pi)^3/V} = \frac{V}{a^3} = N\,,$$

the number of unit cells in the solid. For each of these N values of $\mathbf{k}$ the z-component of the electron spin can take two values, $s_z = \pm\frac{1}{2}\hbar$. Accordingly, each energy band can accommodate $2N$ electrons: i.e., two electrons per primitive unit cell of the solid. The last statement is general, and is *not* restricted to solids with the simple cubic structure.

2.2.3 Schrödinger's equation in k-space

The Fourier representations of the potential $V(\mathbf{r})$ and the Bloch functions $\psi_{\mathbf{k}}(\mathbf{r})$ discussed in Sec. 2.2.1 can be used to obtain a new (and practically important) way of formulating Schrödinger's equation for electrons in a crystal.

We substitute the Fourier expansions (2.6) and (2.9) into the Schrödinger equation (2.2), giving

$$\begin{aligned} E\psi_{\mathbf{k}}(\mathbf{r}) &= -\frac{\hbar^2}{2m}\nabla^2\psi_{\mathbf{k}}(\mathbf{r}) + V(\mathbf{r})\,\psi_{\mathbf{k}}(\mathbf{r}) \\ &= \sum_{\mathbf{G}} \frac{\hbar^2}{2m}(\mathbf{k}+\mathbf{G})^2\, F_{\mathbf{k}+\mathbf{G}}\, e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} + V(\mathbf{r})\psi_{\mathbf{k}}(\mathbf{r}) \\ &\equiv \sum_{\mathbf{G}} E^0(\mathbf{k}+\mathbf{G})\, F_{\mathbf{k}+\mathbf{G}}\, e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} + \sum_{\mathbf{G}'} V_{\mathbf{G}'}\, e^{i\mathbf{G}'\cdot\mathbf{r}} \times \sum_{\mathbf{K}} F_{\mathbf{k}+\mathbf{K}}\, e^{i(\mathbf{k}+\mathbf{K})\cdot\mathbf{r}}, \end{aligned} \tag{2.10}$$

where we have written $E^0(\mathbf{k}+\mathbf{G})$ for the free-electron kinetic energy expression $\hbar^2(\mathbf{k}+\mathbf{G})^2/2m$. We would like to find the equations satisfied by the coefficients $F_{\mathbf{k}+\mathbf{G}}$. The tricky term is the one involving the product of V with $\psi_{\mathbf{k}}$. To help us to pick out the term in $\exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}]$ from the product we have renamed the summation variables $\mathbf{G}'$ and $\mathbf{K}$. This makes it a little easier to show that

$$\begin{aligned} \sum_{\mathbf{G}'} V_{\mathbf{G}'}\, e^{i\mathbf{G}'\cdot\mathbf{r}} \times \sum_{\mathbf{K}} F_{\mathbf{k}+\mathbf{K}}\, e^{i(\mathbf{k}+\mathbf{K})\cdot\mathbf{r}} &= \sum_{\mathbf{K}} F_{\mathbf{k}+\mathbf{K}} \sum_{\mathbf{G}'} V_{\mathbf{G}'}\, e^{i(\mathbf{k}+\mathbf{K}+\mathbf{G}')\cdot\mathbf{r}} \\ &= \sum_{\mathbf{K}} F_{\mathbf{k}+\mathbf{K}} \sum_{\mathbf{G}} V_{\mathbf{G}-\mathbf{K}}\, e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \quad \text{where} \quad \mathbf{G} = \mathbf{G}' + \mathbf{K} \\ &= \sum_{\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \sum_{\mathbf{K}} V_{\mathbf{G}-\mathbf{K}}\, F_{\mathbf{k}+\mathbf{K}}\,; \end{aligned} \tag{2.11}$$

a change of summation variable has been made in the second line and the order of summation is reversed in going from the second to the third line.[3] After inserting (2.11) into (2.10), it is easy to read off the coefficients of $\exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}]$ on each side of (2.10). We find

$$E\,F_{\mathbf{k}+\mathbf{G}} = E^0(\mathbf{k}+\mathbf{G})\,F_{\mathbf{k}+\mathbf{G}} + \sum_{\mathbf{K}} V_{\mathbf{G}-\mathbf{K}}\,F_{\mathbf{k}+\mathbf{K}} \tag{2.12}$$

as the equation satisfied by the coefficients $F_{\mathbf{k}+\mathbf{G}}$: it is fully equivalent to the original Schrödinger equation (2.2). For want of a better name, we call (2.12) *Schrödinger's equation in k-space*.

2.2.4 Weak periodic potential: Nearly-free electrons

In the case of a weak periodic potential $V(\mathbf{r})$, we anticipate that the wave functions $\psi_{\mathbf{k}}(\mathbf{r})$ will be approximately free-electron-like,

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} + [\text{small corrections proportional to } V],$$

and that $E(\mathbf{k}) \simeq E^0(\mathbf{k}) \equiv \hbar^2k^2/2m$. We can verify that this guess is consistent with (2.12) for many (but not all) values of $\mathbf{k}$. For if $F_{\mathbf{k}+\mathbf{G}} \ll F_{\mathbf{k}}$ for $\mathbf{G} \neq \mathbf{0}$ and $E \simeq E^0(\mathbf{k})$, only one term in the sum on the right-hand side of (2.12) needs to be kept (the term with $\mathbf{K} = \mathbf{0}$), and we get

$$F_{\mathbf{k}+\mathbf{G}} \simeq \frac{1}{E^0(\mathbf{k}) - E^0(\mathbf{k}+\mathbf{G})}\, V_{\mathbf{G}}\, F_{\mathbf{k}} \quad \text{for} \quad \mathbf{G} \neq \mathbf{0}\,;$$

this is indeed small, provided $|V_{\mathbf{G}}| \ll |E^0(\mathbf{k}) - E^0(\mathbf{k}+\mathbf{G})|$. But it is not hard to see that this approximation for $F_{\mathbf{k}+\mathbf{G}}$ will not always be valid, even for small V; it certainly fails when

$$E^0(\mathbf{k}) = E^0(\mathbf{k}+\mathbf{G}), \quad \text{i.e., when} \quad |\mathbf{k}| = |\mathbf{k}+\mathbf{G}| \quad \text{for some } \mathbf{G}.$$

The last condition states that $\mathbf{k}$ is equidistant from the points $\mathbf{0}$ and $-\mathbf{G}$ in reciprocal space; that is, $\mathbf{k}$ lies anywhere on the zone boundary bisecting the reciprocal lattice vector $-\mathbf{G}$. In this case, we can expect $F_{\mathbf{k}}$ and $F_{\mathbf{k}+\mathbf{G}}$ to be of similar magnitude.

[3] The method here is slightly different from that used in the lecture. Whichever method is used, all we are really doing is establishing a *convolution theorem* for the Fourier transforms of periodic functions: the sum over $\mathbf{K}$ in (2.11) [and in (2.12)] has the form of a convolution.

There is another, more physical way of understanding this last result. An electron initially in the free-electron state $\exp[i\mathbf{k}\cdot\mathbf{r}]$ will be strongly scattered by the crystal potential $V(\mathbf{r})$ into a state $\exp[i\mathbf{k}'\cdot\mathbf{r}]$, provided $\mathbf{k}' = \mathbf{k} + \mathbf{G}$, where $\mathbf{G}$ is a reciprocal lattice vector: this is just the diffraction condition for waves in a crystal, which we derived for elastic scattering of X-rays in Chapter 1. For a real (as opposed to *virtual*) scattering process, energy is conserved, so that $E^0(\mathbf{k}) = E^0(\mathbf{k}+\mathbf{G})$, which implies (as above) that $\mathbf{k}$ lies on a zone boundary. Hence, if $\mathbf{k}$ lies on a zone boundary, the electron wave function must consist (at the very least) of a term $\exp[i\mathbf{k}\cdot\mathbf{r}]$ for the initial state and a term $\exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}]$ for the scattered wave.

Dispersion relation $E(\mathbf{k})$ near a zone boundary

As discussed above, near a zone boundary we must consider a wave function consisting of at least *two* terms,

$$\psi_{\mathbf{k}}(\mathbf{r}) \simeq F_{\mathbf{k}}\, e^{i\mathbf{k}\cdot\mathbf{r}} + F_{\mathbf{k}+\mathbf{G}}\, e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}\,.$$

If no other terms are significant, the equations (2.12) reduce to two equations for the two unknown coefficients,

$$E\,F_{\mathbf{k}} = E^0(\mathbf{k})\,F_{\mathbf{k}} + V_{-\mathbf{G}}\,F_{\mathbf{k}+\mathbf{G}} \quad \text{and} \quad E\,F_{\mathbf{k}+\mathbf{G}} = E^0(\mathbf{k}+\mathbf{G})\,F_{\mathbf{k}+\mathbf{G}} + V_{\mathbf{G}}\,F_{\mathbf{k}}\,.$$

To simplify the equations slightly, we have set $V_{\mathbf{0}} = 0$. This amounts to choosing the arbitrary additive constant in the potential energy in such a way that the average of $V(\mathbf{r})$, taken over a unit cell, is zero.[4] The two equations for $F_{\mathbf{k}}$ and $F_{\mathbf{k}+\mathbf{G}}$ have the form of a 2×2 matrix eigenvalue problem

$$\begin{pmatrix} E^0(\mathbf{k}) & V_{-\mathbf{G}} \\ V_{\mathbf{G}} & E^0(\mathbf{k}+\mathbf{G}) \end{pmatrix} \begin{bmatrix} F_{\mathbf{k}} \\ F_{\mathbf{k}+\mathbf{G}} \end{bmatrix} = E \begin{bmatrix} F_{\mathbf{k}} \\ F_{\mathbf{k}+\mathbf{G}} \end{bmatrix}.$$

The eigenvalue problem can be solved in the usual way, giving

$$E = \tfrac{1}{2}[E^0(\mathbf{k}) + E^0(\mathbf{k}+\mathbf{G})] \pm \{\tfrac{1}{4}[E^0(\mathbf{k}) - E^0(\mathbf{k}+\mathbf{G})]^2 + |V_{\mathbf{G}}|^2\}^{1/2}, \tag{2.13}$$

where we have used $V_{-\mathbf{G}} = V_{\mathbf{G}}{}^*$, which was noted following (2.6).

> Many band-structure calculations for real solids follow an approach similar to the one we have used here. The calculations are extended to include a large (but finite) number of plane waves, so that an $M\times M$ matrix eigenvalue problem must be solved for the coefficients $F_{\mathbf{k}+\mathbf{G}}$ and the energy bands. Large matrix eigenvalue problems are a standard topic in numerical analysis, and C/Fortran subroutine libraries are available to solve them. Similar routines are also available in Matlab and Mathematica.

Exercise 2.3:

> Do the necessary algebra to obtain the result (2.13) for the energy of an electron near a zone boundary.

The dispersion relation (2.13) simplifies in two special cases. If $|V_{\mathbf{G}}| \ll \frac{1}{2}|E^0(\mathbf{k}) - E^0(\mathbf{k}+\mathbf{G})|$, the two solutions are

$$E \simeq E^0(\mathbf{k}) \quad \text{and} \quad E \simeq E^0(\mathbf{k}+\mathbf{G})\,;$$

as we should expect, the effect of the crystal potential is negligible if we are far away from the zone boundary. The other special case is when $\mathbf{k}$ lies exactly on the zone boundary, so that $E^0(\mathbf{k}) = E^0(\mathbf{k}+\mathbf{G})$. In this case,

$$E = \tfrac{1}{2}[E^0(\mathbf{k}) + E^0(\mathbf{k}+\mathbf{G})] \pm |V_{\mathbf{G}}| \quad \text{for } \mathbf{k} \text{ on the zone boundary;} \tag{2.14}$$

the degeneracy $E^0(\mathbf{k}) = E^0(\mathbf{k}+\mathbf{G})$ on the zone boundary has been split by the crystal potential: the amount of the splitting is $2|V_{\mathbf{G}}|$.

[4] See (2.6), which gives $V_{\mathbf{0}} = \int V(\mathbf{r})\, d^3\mathbf{r}/v_{\text{cell}}$. This expression is the average of $V(\mathbf{r})$, taken over a unit cell.

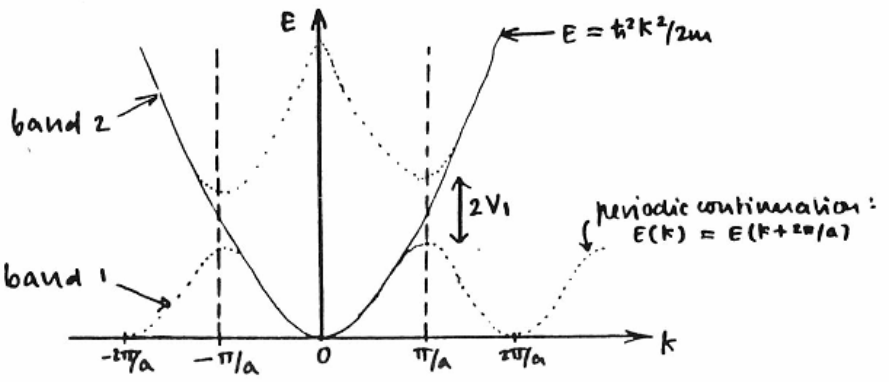

Figure 2.2: Nearly-free electron bands in one dimension, for an electron in the real potential $V(x) = 2V_1 \cos(2\pi x/a)$, with $V_1 > 0$. The free-electron dispersion relation is indicated by a solid curve; the dotted lines are the bands, as modified by the crystal potential. At the zone boundary, $k_x = \pm\pi/a$, the gap between the first and second bands is $2|V_1|$; note that for the given potential, the Fourier coeffcients $V_1 = V_{-1}$ are real and positive, so we could equally well write $2V_1$ for the gap (as in the figure).

In the second-year course PHYS20252 and in J. Pearson's notes available online, an alternative method is given for deriving this energy splitting between the nearly-free electron bands; the argument uses the symmetry of the electron wave function and simple first-order perturbation theory. This is an excellent approach, which isn't superseded by the method given above, but the equation (2.13) contains additional information on the *shape* of the nearly-free electron energy bands. Nearly-free electron bands are sketched in Figure 2.2, for the one-dimensional case.

Equation (2.14) is often misunderstood. It does *not* show that two energy bands are always completely separated in energy—even though this happens to be true for an electron in a one-dimensional periodic potential. In higher dimensions, and if $|V_G|$ is not too large, electrons in different bands *can* have the same energy, *provided they have different wave vectors*. We shall see an important example of this in Sec. 2.2.6, but first we remind ourselves of the band-theory account of the difference between metals and insulators.

2.2.5 Metals and insulators

Having argued that a weak crystal potential can lead to energy gaps between bands, we can now formulate the difference between metals and insulators.

In the ground state of the electron gas in a solid, the electrons occupy the states of lowest energy of the lowest-lying bands, consistent, as usual, with the exclusion principle: no more than two electrons ($s_z = \pm\frac{1}{2}\hbar$) for each value of the wave vector. As we have see already seen, each band can accommodate up to $2N$ electrons, where N is the number of primitive unit cells in the solid. Therefore, if there is an *odd* number of electrons in the unit cell, at least one band will be only partly filled with electrons: the Fermi energy falls within this band. When an electric field is applied to the solid, the electrons redistribute among the states near E_F to form a current-carrying state at very little cost in energy. The material is consequently metallic.

If, however, the lowest-lying energy bands are completely full of electrons—this requires an *even* number of electrons per primitive unit cell—the electrons cannot be redistributed to form a current-carrying state without exciting them from a filled band across an energy gap to one of the higher, empty bands. This costs a lot of energy, and cannot occur for moderate field strengths: the material is an insulator.

We reach the important conclusions that an insulator, such as diamond, must have an even number of electrons per primitive unit cell, and that materials with an odd number, such as sodium, are necessarily metals.

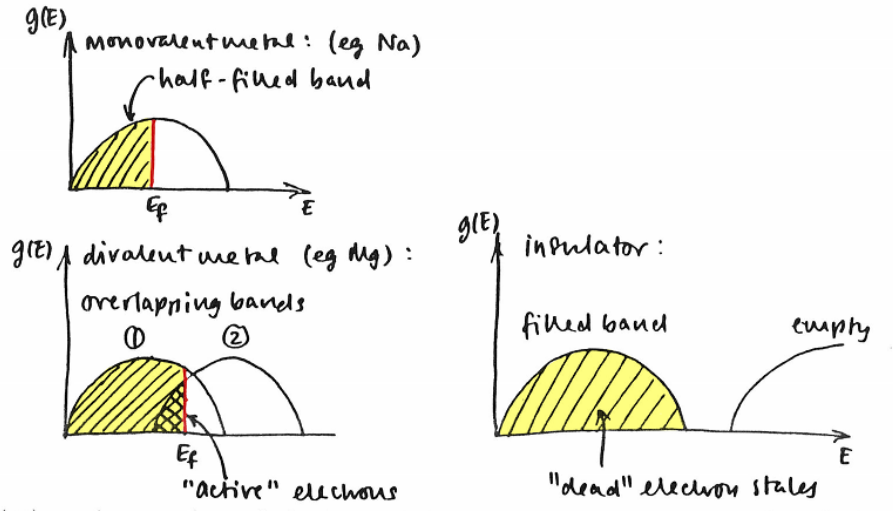

Figure 2.3: Distinction between metals (left) and insulators (right) in the band picture. At $T = 0$, the bands of an insulator are either all full or all empty, whereas a metal has at least one partly-occupied band. The electrical properties of a metal are dominated by the electrons at the Fermi energy, E_F, which have unoccupied states of neighbouring energy that they can transfer to under the action of an applied electric field.

Exercise 2.4:

Check these conclusions against the periodic table of the elements, so far as you can.

The converse statements are not necessarily true: for example, the divalent elements zinc and magnesium are both metals. In all such cases, although there are enough electrons to fill a whole number of bands, two or more bands are found to *overlap* in energy, so that the Fermi energy lies in more than one partly-filled band. We shall investigate the details of this in the next section, but for the time being we note that there are three distinct possibilities for the densities of electron states with respect to energy, taking these to be representative of a monovalent metal, a divalent metal, and an insulator. These possibilities are illustrated in Figure 2.3. In the figure, $g(E)$ is the number of electron states per unit energy interval; this was calculated in detail for free electrons in the second-year course. In this course our only explicit use of $g(E)$ will be, as here, in sketches indicating the presence or absence of energy levels in a given range of energy.

2.2.6 Band overlap in a nearly-free-electron divalent metal

We consider the case of a simple-cubic divalent metal with one atom per primitive unit cell of side a.[5] Diagrams illustrating the occupied states are shown in Figure 2.4 The free-electron Fermi sphere contains $2N$ electrons, so it has the same volume as the first Brillouin zone, which is a cube of side $2\pi/a$. Parts of the Fermi sphere protrude from the faces of the cube. The states "just outside" the Brillouin zone belong to a different band from those just inside. We can see this by plotting the energy as a function of k_x. It resembles the free-electron result for one dimension, with a gap $2V_1$ in the dispersion relation[6] at $k_x = \pi/a$.

[5]The real divalent elements magnesium and zinc both have a close-packed hexagonal structure, and strontium and calcium are face-centered cubic. Similar arguments can be used for these more complex structures.

[6]Actually, this gap is $2|V_{100}|$, since the zone face bisects the reciprocal lattice vector (100), but we simplify the notation here.

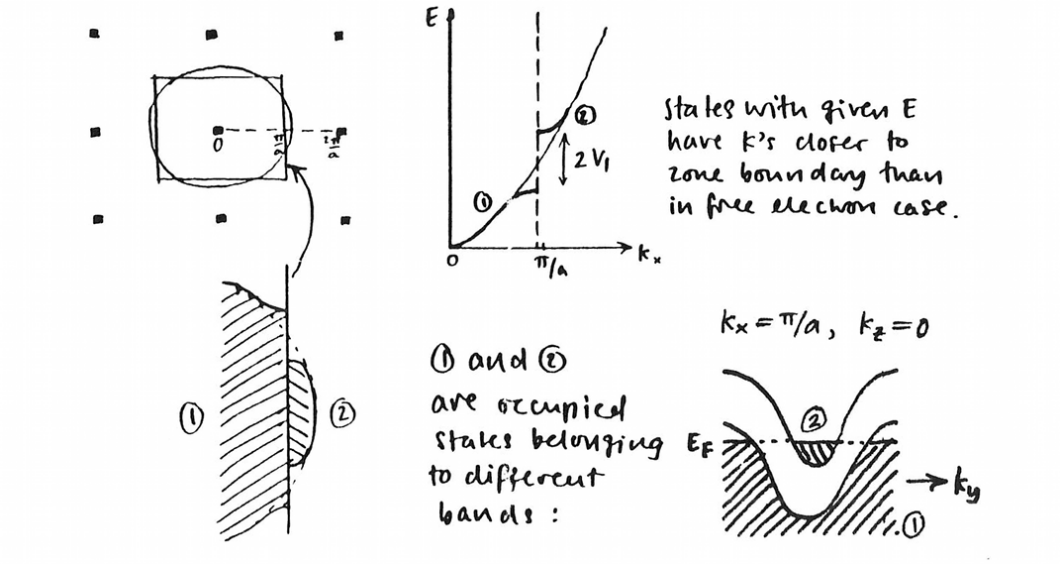

Figure 2.4: Simple-cubic divalent metal. *Top left:* Free-electron Fermi sphere in relation to the first Brillouin zone. *Lower left and top right:* Electron energies are modified near a zone boundary, leading to fragmentation of the Fermi sphere and distortions of the fragments belonging to the first and second bands; this is illustrated in the so-called "extended zone scheme", in which states of the first band are indicated in the first Brillouin zone and states of the second band are shown just outside, in the second zone. *Lower right:* Electron energies plotted as a function of k_y, along a line with fixed $k_z = 0$ on the zone boundary at $k_x = \pi/a$; the aim is to show how the first two bands overlap in energy, provided $2V_1$ is small.

Electrons at the portion of the Fermi surface lying in the first zone have larger energies than electrons *within* the Fermi sphere, just outside a zone face, at $\mathbf{k} = (\pi/a, 0, 0)$. In other words, there is an overlap in energies between the two bands. This is not to say, however, that the states of the two bands have merged together: it is simply that the energies of electrons in different bands may coincide for *different* values of the wave vector $\mathbf{k}$.

In Figure 2.4, note that the free-electron Fermi sphere is slightly distorted near the zone boundary; from the plot of $E(k_x)$ we see that states of a given energy occur with wave vectors closer to the zone boundary than they would for free electrons. The effect is to make a surface of constant energy, such as the Fermi surface, meet the zone boundary at right-angles.

Now, the energy is periodic in reciprocal space, $E(\mathbf{k}) = E(\mathbf{k} + \mathbf{G})$, so that portions of the Fermi sphere "just outside" the zone boundary $k_x = \pi/a$ can be translated back to a point just inside, near $k_x = -\pi/a$. We sketch separately the occupied states of the two bands in Figure 2.5.

Increasing the strength of the periodic potential increases the gap at the zone boundary, lowering the energies of electrons in the first band and raising energies in the second. To minimize their total energy, electrons transfer from the second band to the first, thereby keeping the Fermi energies equal. As V_1 increases there finally comes a point where the two bands cease to overlap: one is full of electrons and the other is empty. When this happens the Fermi surface disappears completely and the solid has become an insulator.

We can obtain a semi-quantitative criterion for this "band-crossing" transition by considering two states from different bands lying at the intersections of the Fermi surface with the zone boundary; see Figure 2.6. Assuming the gap to be constant over the whole surface of the zone we will have

$$E_F = \hbar^2 k_1^2/2m - V_1 = \hbar^2 k_2^2/2m + V_1. \tag{2.15}$$

As V_1 increases, the lower band fills and k_1 increases to its maximum value $\pi\sqrt{3}/a$ for the zone corner. At the same time, k_2 shrinks to its minimum value π/a at the centre of the zone face, and the bands just

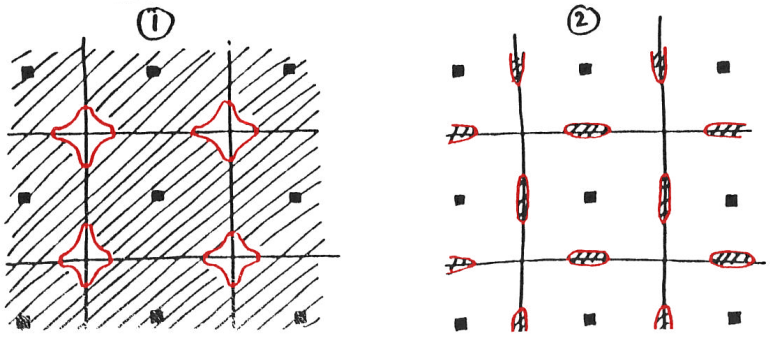

Figure 2.5: Shaded regions indicate occupied states in the first (left) and second (right) bands of a divalent nearly-free electron metal with the simple-cubic structure. Filled squares represent reciprocal lattice points and the vertical and horizontal lines represent the boundaries of the first Brillouin zone. A so-called "repeated zone scheme" is used in this figure: the first Brillouin zone (the central cube) is repeated to fill all of $\mathbf{k}$-space. This makes it easier to see that the occupied states of the second band form lens-shaped "pockets" centered on the zone faces. The Fermi surface (which consists of several disjoint fragments) is picked out in red.

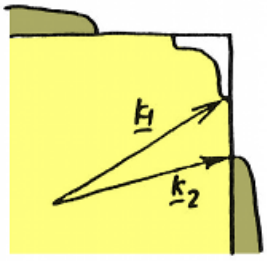

Figure 2.6: Band crossing transition. As the gap $2|V_1|$ increases, electrons transfer from the second band to the first, to keep the Fermi energy equal in each band; $\mathbf{k}_1$ increases till it reaches the corner of the Brillouin zone at $(111)\pi/a$, while $\mathbf{k}_2$ shrinks till it reaches the centre of the zone face at $(100)\pi/a$. In this (fictitious) process, the Fermi *surface* gradually shrinks to nothing, and the material becomes an insulator.

fail to overlap.

Exercise 2.5:

Use (2.15) to show that the critical value of V_1 is $\hbar^2\pi^2/2ma^2$. Evaluate this for a divalent metal with $a = 2.5$ Å.

In this example the perturbing potential would have to be quite large to convert the metal into an insulator. However, in cases where zone boundaries would follow the free-electron Fermi surface more closely [covalent semiconductors are an example] this appears to be the mechanism that opens up a gap between occupied and unoccupied states. We shall consider the case of Group IV elements later.

2.2.7 Tight-binding method

When we considered the motion of electrons in the nearly-free electron approximation, we found that energy gaps $2|V_{\mathbf{G}}|$ appeared, close to the zone boundaries, as a result of the interaction of the electrons with the crystal potential. Our approach relied on perturbation theory, and was perhaps a little abstract. The idea of the *tight-binding method* is to build up the Bloch functions, starting from atomic wave functions. The existence of energy gaps may seem more obvious if we approach the problem in this way, as we are already very familiar with the idea of *atoms* having discrete energy levels.

When reading this section, bear in mind that the important results are contained in equations (2.19)–(2.22). The precise method of derivation of (2.19) is much less important than the intuitive picture of an electron tunnelling between atomic orbitals that overlap only slightly.

Diatomic molecule

To simplify the discussion we suppose that the motion of the electron is purely one-dimensional, so that Schrödinger's equation takes the form

$$\hat{H}\psi(x) = E\psi(x) \quad \text{with} \quad \hat{H} = \frac{\hat{p}^2}{2m} + W(x) + W(x-a). \tag{2.16}$$

Here $W(x)$ is the attractive potential due an ion at $x = 0$, and $W(x-a)$ is due to the other ion at $x = a$. It is assumed that these potentials tend to zero far from an ion. When two atoms are well separated, the electronic wave functions and energy levels of a single atom will be only slightly affected by the presence of the other atom. This suggests that we can build up wave functions for the composite system using linear combinations of atomic orbitals ϕ_n, which are the bound states of an electron in the field of a single ion,

$$\left[\hat{p}^2/2m + W(x)\right]\phi_n(x) = E_n\phi_n(x). \tag{2.17}$$

The lowest-lying states of the composite system (a covalently bonded molecule) should resemble the ground state of the two atoms: substituting a trial function $\psi(x) = c_0\phi_0(x) + c_1\phi_0(x-a)$ into (2.16) we find

$$\begin{aligned}\hat{H}\psi &= c_0\left[E_0\phi_0(x) + W(x-a)\phi_0(x)\right] + c_1\left[E_0\phi_0(x-a) + W(x)\phi_0(x-a)\right] \\ &\simeq E\left[c_0\phi_0(x) + c_1\phi_0(x-a)\right].\end{aligned} \tag{2.18}$$

The correction terms $W(x)\phi_0(x-a)$ and $W(x-a)\phi_0(x)$ should have only a small effect: at large distances from an ion the atomic orbitals decrease exponentially, so that we expect $\phi_0 \simeq e^{-\lambda a}$ in the region of the second ion.

We are interested only in the coefficients c_0 and c_1 appearing in (2.18), so we eliminate the dependence on x by multiplying each side by $\phi_0(x)$ (which is a real function) and integrating from $x = -\infty$ to ∞,

$$\begin{aligned}&c_0E_0 + c_0\int W(x-a)\,\phi_0^2(x)\,dx \\ &\quad + c_1\int W(x)\,\phi_0(x)\,\phi_0(x-a)\,dx + c_1E_0\int \phi_0(x)\,\phi_0(x-a)\,dx \\ &\quad \simeq c_0E + c_1E\int \phi_0(x)\,\phi_0(x-a)\,dx;\end{aligned}$$

where we have used the normalization condition $\int \phi_0^2(x)\,dx = 1$. The remaining integrals that involve products $\phi_0(x)\,\phi_0(x-a)$ all contain the exponential factor $e^{-\lambda a}$; whereas the integral of $W(x-a)\,\phi_0^2(x)$ contains two such factors and will be neglected here. The last terms on each side cancel to the same order, because the difference $E - E_0$ is exponentially small.

Carrying out the same procedure after multiplying (2.18) by $\phi_0(x-a)$ and integrating we obtain two equations for the unknown coefficients and the energy,

$$Ec_0 \simeq E_0 c_0 - Bc_1 \quad \text{and} \quad Ec_1 \simeq E_0 c_1 - Bc_0, \tag{2.19}$$

where

$$-B = \int_{-\infty}^{\infty} W(x)\,\phi_0(x)\,\phi_0(x-a)\,dx.$$

Note that the integral is negative because the two factors ϕ_0 have the same sign and $W(x)$ is attractive. The equations (2.19) can be interpreted intuitively, as a form of Schrödinger's equation. They say that in the absence of tunnelling ($B = 0$) an electron stays on one atom and has energy E_0. But if $B \neq 0$, the electron can tunnel between atoms, an effect contained in the "hopping terms" $-Bc_1$ and $-Bc_0$.

The solutions of (2.19) are

$$E = E_0 \mp B \quad \text{for} \quad c_1 = \pm c_0 \quad \text{respectively,}$$

so that the ground- and first excited-state solutions of (2.16) are the symmetric and antisymmetric wave functions $\psi_{s,a} = [\phi_0(x) \pm \phi_0(x-a)]/\sqrt{2}$.

Exercise 2.6:

Check that you can prove the last results.

The atomic energy level E_0 for the two atoms has been split into two levels separated by an amount $2B$ which decreases rapidly with increasing separation of the atoms; B is sometimes called the *covalent energy*, as it determines the binding energy of the molecule.

Polar molecule

The same picture can be used for a *polar* (or *heteropolar*) molecule, in which the constituent atoms X and Y (lithium and hydrogen, say) are different. The only change to Schrödinger's equation (2.19) is to the atomic energy level E_0 which must be replaced by values E_X, E_Y appropriate to the two atoms,

$$Ec_X = E_X c_X - Bc_Y \quad \text{and} \quad Ec_Y = E_Y c_Y - Bc_X,$$

where $E_Y < E_X$ refer respectively to the anion and the cation. Solving the last pair of equations for E gives

$$E = \tfrac{1}{2}(E_X + E_Y) \pm \left[\tfrac{1}{4}(E_X - E_Y)^2 + B^2\right]^{1/2},$$

so that in the polar molecule the splitting of the bonding and antibonding energy levels is greater than in the purely covalent case considered before.

Exercise 2.7:

Verify the last result, and by calculating the coefficients c_X, c_Y show that in the bonding state an electron is more likely to be found on the anion. [Probabilities are proportional to the squares of the coefficients.]

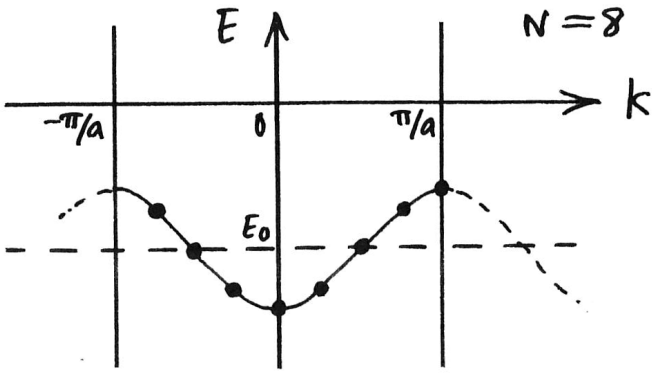

Figure 2.7: Dispersion relation $E(k)$ for an electron in a ring of 8 atoms. Filled circles represent the 8 allowed values of k in the first Brillouin zone.

A ring of atoms

In a similar way we can approximate the wave functions and energies of an electron in a molecule consisting of a ring of $N > 2$ identical atoms. Here the electron can tunnel to the rth atom from either of its neighbours $r-1$ and $r+1$. Clearly the Schrödinger equation (2.19) for the coefficients c_r will be generalized to

$$Ec_r = E_0 c_r - Bc_{r+1} - Bc_{r-1} \tag{2.20}$$

for a trial wave function $\psi = \sum_r c_r \phi_0(x - ra)$. Just as a linear differential equation with constant coefficients can be solved by an exponential substitution, we can attempt the same with (2.20), setting $c_r = A\exp[ikra]$. [You were shown a similar procedure for solving for the classical vibrational modes of a chain of atoms in the second-year course.] After dividing through each side by the common factor c_r we find

$$E = E_0 - Be^{ika} - Be^{-ika} = E_0 - 2B\cos ka. \tag{2.21}$$

Since the chain is closed, the possible values of k are constrained by the fact that $c_N \equiv c_0$, so that $\exp[ikNa] = 1$. Hence

$$k = \frac{2\pi n}{Na}, \quad n = 0, \pm 1, \pm 2, \ldots.$$

Not every value of k corresponds to a different wave function. Only N values can be physically distinct, since the wave function was built up from N orbitals $\phi_0(x - ra)$; we note also that the coefficients $c_r = A\exp[ikra]$ are unchanged by the replacement $k \to k + 2\pi/a$, so that no physical quantity such as the energy (2.21) can be affected. [Verify this.] In representing the dependence of a physical quantity on k we are therefore free to choose the N solutions of longest wavelength; i.e., the range $-\pi/a < k \leq \pi/a$, which is the *first Brillouin zone* described in Sec. 2.2.2. The shortest of the waves, $k \to \pm\pi/a$, have a wavelength $\lambda = 2a$.

Just as the level E_0 was split into two in the diatomic molecule, here it becomes several levels in the range $E_0 - 2B$ to $E_0 + 2B$; see Figure 2.7.

Tight-binding solid

We can apply Bloch's theorem to the electron wave functions in a so-called *tight-binding solid*, in which the wave functions of adjacent atoms are assumed to overlap very little. Although results from the tight binding method are not normally very accurate in applications to real solids, it is easy to understand the basic idea behind the method. Because of its simplicity, the tight-binding approximation is often used as a starting point in theoretical discussions of the d and f electrons in a solid: the f-electron wave functions, in particular, decrease rapidly with distance from an atom, so that the approximation of small overlap between the wave functions on different atoms is quite reasonable.

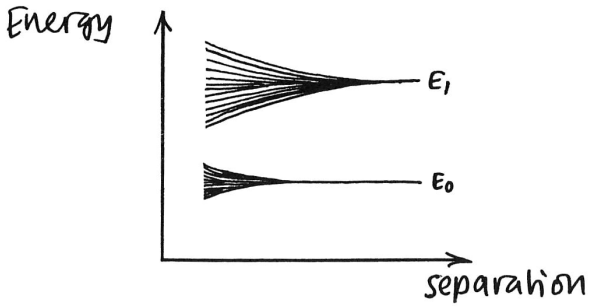

Figure 2.8: Energies of tight-binding electrons as a function of the separation of atoms. The atomic energy levels E_0 and E_1 broaden to form a pair of bands. For the range of separations shown, there is no overlap of bands, but band-overlap (and hybridization) would be expected for typical interatomic separations in solids.

In the neighbourhood of an atom at $\mathbf{R}$ the influence of other atoms is relatively small, so that the wave function should resemble the electron wave functions ϕ_n of a free atom,

$$\psi(\mathbf{r}) \simeq \phi_n(\mathbf{r} - \mathbf{R}).$$

By analogy with the wave functions of electrons in the ring of atoms we approximate the wave function throughout the whole crystal by a linear combination of atomic orbitals,

$$\psi_{n\mathbf{k}}(\mathbf{r}) \simeq \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} \phi_n(\mathbf{r} - \mathbf{R}) e^{i\mathbf{k}\cdot\mathbf{R}}, \tag{2.22}$$

where the factor $1/\sqrt{N}$ ensures that $\psi_{n\mathbf{k}}$ is normalized approximately to unity. The suffix n is a "band index", reminding us that the wave function (2.22) is constructed from a particular atomic orbital ϕ_n.

Exercise 2.8:

Verify that this wave function satisfies Bloch's theorem (2.3).

For well-separated atoms the energies $E_n(\mathbf{k})$ of the wave functions (2.22) will be similar to the corresponding atomic energy levels E_n; but as the atoms are brought closer together the energies will spread out to form practically continuous *bands* of energy levels; Figure 2.8. The total number of levels in each band will be $2N$, where N is the number of primitive unit cells in the solid: as in the discussion of the ring of atoms there are N distinct values of the wave vector, but the electron spin can take two values for each of these.

When the separation of atoms becomes small enough, the bands corresponding to different atomic states begin to overlap in energy. This corresponds to the band overlap discussed in the nearly-free electron approximation. Once the bands have started to overlap, the simple expression (2.22) will no longer be a reasonable approximation throughout the Brillouin zone. Instead, the wave function will contain $\mathbf{k}$-dependent linear combinations of two or more atomic states, a situation known as *hybridization*. Hybridization complicates the simple tight-binding picture significantly and we shall not discuss it further; it is, nevertheless, an essential feature of any tight-binding calculation that attempts to describe the band structure and bonding of real materials.

2.3 Semiclassical dynamics of Bloch electrons

The electrical conductivity of metals was investigated in the second-year course within the free-electron model. Newton's second law was used as the equation of motion for the electrons, assuming that the

crystal momentum $\hbar\mathbf{k}$ played the part of the momentum $\mathbf{p}$. We might instead have guessed $\mathbf{p} = m\mathbf{v}$ for the momentum, where m is the free electron mass. For free electrons either replacement would have given correct results, but in this section we shall argue that $\hbar\mathbf{k}$ is the right choice for the general case of an electron in a periodic potential. The mass of an electron is more problematic, and depends on the shape of the band structure $E(\mathbf{k})$. One choice of mass allows electrons to have negative, or even infinite masses. In avoiding the conceptual problems of particles with negative mass we will be led to the idea of a *hole* in a band.

2.3.1 Electron velocities

When we are interested in the transport of electrons in solids, it is convenient to use not Bloch waves, with a definite value of the wave vector $\mathbf{k}$, but wave packets, which are superpositions of Bloch waves with a small range Δk of wave vectors. A wave packet may be fairly well localized in space, with an uncertainty in position given by Heisenberg's relation $\Delta x \sim 1/\Delta k$. For the wave vector to be sharply defined compared with the size of the Brillouin zone we must have $\Delta k \ll 1/a$, so that the wave packet spreads over many unit cells of the solid, $\Delta x \gg a$. If we are interested only in the motion of the electron over distances that are much greater than Δx, we can effectively treat the wave packet as a point particle and investigate its equation of motion. This is the semiclassical limit of electron dynamics.

The wave packet moves through the solid with a *group velocity* whose components $v_i = \partial\omega/\partial k_i$ are given by

$$v_x = \frac{1}{\hbar}\frac{\partial E}{\partial k_x}, \quad v_y = \frac{1}{\hbar}\frac{\partial E}{\partial k_y}, \quad v_z = \frac{1}{\hbar}\frac{\partial E}{\partial k_z}, \tag{2.23}$$

where we have used the quantum-mechanical expression $\omega = E/\hbar$ for the angular frequency. [Note that because the zero of energy can be chosen arbitrarily, the *phase velocity* of the electron, $\omega/k = E/\hbar k$, is a much less useful quantity.] The three expressions (2.23) are often abbreviated to

$$\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\frac{\partial E(\mathbf{k})}{\partial \mathbf{k}} \quad \text{or} \quad \mathbf{v}(\mathbf{k}) = \hbar^{-1}\nabla_{\mathbf{k}} E(\mathbf{k}).$$

Now the state of motion of a Bloch electron is specified by the wave vector $\mathbf{k}$, which is constant in the absence of any applied fields; so its velocity $\mathbf{v}$, given by (2.23), is also constant. Hence if an electric current is set up in a perfect crystalline metal, the electron velocities will never change and the current will remain constant, even in the absence of an applied electric field. The conductivity of such a metal would be infinite despite the electrons' interaction with the ions of the lattice. We recall that the conductivity of pure copper became very large at low temperatures, when the thermal motion of the ions was expected to be small.

2.3.2 Motion in an applied field

An electron in an applied electric field[7] $\mathcal{E}_x$ gains energy at a rate

$$\dot{E} = [\text{force}] \cdot [\text{velocity}] = -e\mathcal{E}_x v_x,$$

which can be justified in the semiclassical picture as the rate of change of potential energy $-e\mathcal{E}_x x$ of an electron wave packet localized near x. Assuming that the electron does not make a transition between bands,[8] the rate of change of its energy can also be written

$$\dot{E} = \dot{k}_x \partial E/\partial k_x = \hbar \dot{k}_x v_x,$$

[7] The electric field strength is denoted by $\boldsymbol{\mathcal{E}}$ to avoid confusion with the energy E.

[8] This requires the applied field to be "weak", $e\mathcal{E}_x a \ll E_g$ according to perturbation theory, where E_g is the energy gap between different bands. In this context a "weak" field may in fact be extremely large in laboratory terms, as you may see by setting $a \sim 1$ Å and $E_g \sim 1$ eV.

using (2.23) for the group velocity. Comparison of the right-hand sides of the last two expressions shows that

$$\hbar \dot{k}_x = -e\mathcal{E}_x.$$

For simplicity we have imagined the motion to be restricted to 1D, but the argument can be extended to 3D to give the semiclassical equation of motion

$$\hbar \dot{\mathbf{k}} = -e\boldsymbol{\mathcal{E}}. \tag{2.24}$$

It is worth emphasizing here that, despite the formal resemblance between (2.24) and Newton's second law, $\hbar\mathbf{k}$ is still not the true momentum of the electron. The rate of change of the true momentum would have to include the forces due to the ions of the crystal lattice.

Bloch oscillations:

Like any other physical property of the electron, its group velocity is an oscillatory function of the wave vector. In the case of a constant electric field the equation of motion (2.24) can be integrated at once to give a wave vector which varies linearly with time, so that the velocity of a band electron will oscillate in time. These oscillations were "predicted" long ago by Bloch, though in an ordinary solid they are unobservable: the period of oscillation is $2\pi\hbar/e\mathcal{E}_x a$, which is much greater than the typical mean free time of an electron. Very recently, however, it has become possible to observe the Bloch oscillations directly in layered semiconductor structures engineered so as to have long spatial periods of a few hundred angstrom. For these structures the period of oscillation is sufficiently short (in the microwave region) for resonance to be observed at very low temperature, where the scattering rate due to phonons is greatly reduced.

2.3.3 Effective mass of an electron

Equation (2.24) is an equation of motion for the wave vector of an electron in a crystal, whereas normally in mechanics we take the position $\mathbf{r}$ of the particle as fundamental. It is therefore worth rewriting (2.24) in a form resembling [acceleration] = [force]/[mass].

The acceleration $\dot{v}_x$ is given, in one-dimensional motion, by

$$\dot{v}_x(k_x) = \dot{k}_x \frac{\partial v_x}{\partial k_x} = -e\mathcal{E}_x \times \frac{1}{\hbar^2}\frac{\partial^2 E(k_x)}{\partial {k_x}^2},$$

where we have used (2.23) and the equation of motion for k_x. This is now of the form [acceleration] = [force]/[mass] with the mass replaced by the *electron effective mass*

$$m_e(k_x) = \frac{\hbar^2}{\partial^2 E/\partial {k_x}^2}.$$

Note that the effective mass depends on the wave vector of the electron and on the shape of the band. Near the band discontinuity induced at a zone boundary by a weak periodic potential, the curvature of the band can be large because the change in $E(k_x)$ all occurs in a narrow range of k_x. The effective mass may be small there compared with the free electron mass.

For well separated atoms described by the tight-binding model, the electron energies [given, for example, by (2.21)] vary little with k_x, so that the corresponding effective masses may be large compared with the mass of a free electron. This agrees with intuition, as in an applied field the electrons will find it difficult to tunnel from one atom to the next if the overlap of the atomic wave functions is small.

Exercise 2.9:

> Show that the effective mass of an electron with $E(k)$ given by the tight-binding result (2.21) is $m_e = \hbar^2/(2Ba^2)$ at $k = 0$. Note that m_e is large if the overlap integral B is small.

Effective mass can be infinite at a point of inflexion of $E(k_x)$, or negative near the top of a band. If the effective mass is negative, the electron will accelerate in the opposite direction to an applied external force. At first sight this appears paradoxical, but remember that the external force is not the only one acting on an electron: there is also the periodic electric field due to the ions. In practice this latter force is much larger than any externally applied field. Its effect on the electron dynamics is hidden away in the form of $E(k_x)$, which must be worked out using Schrödinger's equation for the electrons.

Holes in a nearly-filled band

In cases where the Fermi energy falls near the top of a band, the electrons most important for conduction will have negative effective masses. It is more convenient then to describe the nearly-filled band using the properties of its unoccupied states.

A band with just one missing electron is said to contain one *hole*. The properties of the hole are simply taken to be the properties of the nearly-filled band. For example, the wave vector of the hole is given by

$$\mathbf{k}_h = -\mathbf{k}_e,$$

which is the *change* in the total wave vector of the band when an electron is removed from a state with wave vector $\mathbf{k}_e$. Similarly, the energy of the hole will be the change in the total energy when an electron is removed,

$$E_h(\mathbf{k}_h) = -E_e(\mathbf{k}_e) = -E_e(-\mathbf{k}_h).$$

With these definitions it is not difficult to show that the group velocity of the hole is $\mathbf{v}_h(\mathbf{k}_h) = \mathbf{v}_e(\mathbf{k}_e)$, and that the hole effective mass satisfies $m_h = -m_e$. The semiclassical equation of motion (2.24) becomes

$$\hbar\dot{\mathbf{k}}_h = +e\boldsymbol{\mathcal{E}},$$

so that the dynamics of a nearly-filled band of electrons with negative charge $-e$ are formally identical to the dynamics of a much smaller number of holes with positive charge $+e$ and positive effective mass.

2.4 Free-electron bands and crystal structure

So far we have seen that crystal structure can affect the form of the bands directly via the spacing of atoms in the tight-binding picture, or by perturbation of the free-electron dispersion relation near zone boundaries. We looked in detail at a divalent metal [with one atom per primitive cell] and suggested that, if the crystal potential was big enough, the Fermi sphere might be drawn back into the first Brillouin zone, turning the material into an insulator. In reality there are no divalent insulators of this kind, but we can use a similar picture to correlate the free-electron bands with the diamond structure of the tetravalent elements silicon and germanium. Only the simplest, qualitative ideas will be used here.

Since the binding energy of a crystal of neutral atoms is due mainly to the valence electrons, we can turn the problem around and ask how the electrons in a band affect the stability of a crystal structure. Again we shall use a free-electron picture, this time arguing that in bismuth the atoms have rearranged so as to minimize the energy of the electrons. The stability of alloys can be investigated in a similar way.

Although the crystal structures of diamond and of bismuth are not the same, each is closely related to the face-centered cubic lattice. We digress briefly to construct the lattice reciprocal to FCC.

2.4.1 Construction of the reciprocal lattice for FCC

The FCC lattice is sketched in Fig. 2.9. The whole lattice can be generated by the three vectors joining a corner of the conventional unit cell to the midpoints of the adjoining faces: $\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$, where

$$\mathbf{a}_1 = [0,1,1]a/2, \quad \mathbf{a}_2 = [1,0,1]a/2, \quad \mathbf{a}_3 = [1,1,0]a/2.$$

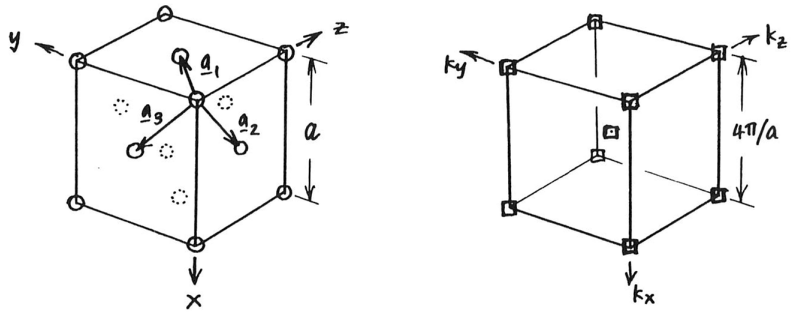

Figure 2.9: The FCC lattice (left) and its reciprocal (right), which is BCC.

Using the equations (1.10) satisfied by the reciprocal lattice vectors **G**, we find the following relations between the Cartesian components,

$$G_y + G_z = 4\pi p_1/a, \quad G_z + G_x = 4\pi p_2/a, \quad G_x + G_y = 4\pi p_3/a,$$

where the p_i can be any integers. Solving for G_x we find

$$G_x = 2\pi(p_2 + p_3 - p_1)/a,$$

and similar expressions for G_y and G_z found by cyclic permutation of p_1, p_2 and p_3. Now whichever signs are chosen, the quantities $\pm p_1 \pm p_2 \pm p_3$ are either all even or all odd integers, so that

$$G_i = 2\pi q_i/a, \quad \text{with } q_i \text{ either all even or all odd.}$$

For example, $(200)2\pi/a$ and $(111)2\pi/a$ are vectors of the reciprocal lattice, while $(100)2\pi/a$ and $(3\bar{3}2)2\pi/a$ are not. When we sketch the allowed values we find that the reciprocal lattice is body-centered cubic with cubic lattice parameter $4\pi/a$.

2.4.2 Group IV elements: Jones theory

The tetravalent elements carbon (diamond), silicon, germanium and grey tin (the form stable at low temperatures) are all insulating in the pure state at low temperatures. They adopt a structure with two atoms per lattice point of the FCC structure. Each atom is covalently bonded to 4 equally spaced neighbours, arranged tetrahedrally around it. Although this arrangement can be explained in terms of directed valence,[9] this alone does not explain why the materials are insulators. For example, molten germanium is a good conductor, and carbon in the form of graphite (a 2D layered structure) is a rather poor metal.

Here we try to explain the insulating behaviour starting from a nearly-free electron picture: the periodic crystal potential induces gaps (discontinuities) in the free-electron $E(\mathbf{k})$ on planes (zone boundaries, or Bragg planes) bisecting reciprocal lattice vectors. States just outside the zone boundary will have higher energies than those just inside, so that the total energy of the electrons can be reduced by transferring some of them to states within the zone. If this picture is tenable for the Group IV elements, the energy gap must be big enough for the Fermi surface to have disappeared into a set of zone boundaries enclosing the same volume as the Fermi sphere. Hence we look for possible zone boundaries whose distances from $\mathbf{k} = \mathbf{0}$ are slightly less than k_F.

The first few reciprocal lattice vectors of the FCC structure and the distances from $\mathbf{k} = \mathbf{0}$ of the planes bisecting them are listed in Table 2.1.

[9] For our purposes valence equals the number of weakly bound electrons. Here this includes only the s and p electrons. The d electron wave functions in Ge and Sn decrease rapidly with distance, and are best described (like other states of low energy) as a filled—and therefore electrically inert—tight-binding band.

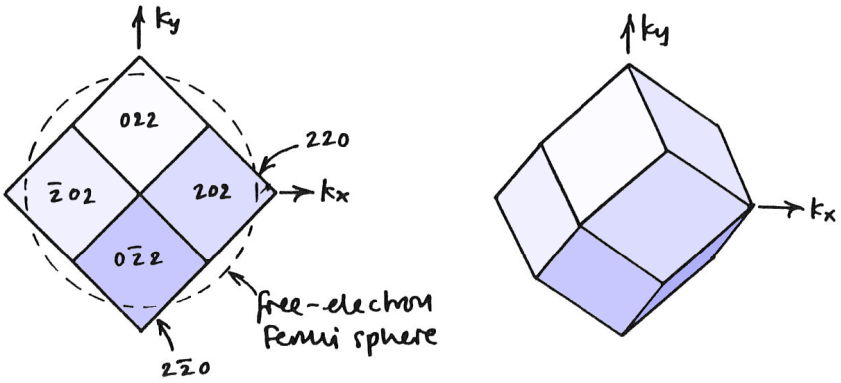

Figure 2.10: Jones zone for the diamond structure. *Left:* Viewed along the k_z-axis, looking towards the (k_x, k_y) plane; the zone boundaries have been labelled by the reciprocal lattice vectors they bisect. *Right:* Perspective view.

Table 2.1: Shortest reciprocal lattice vectors **G** for the FCC lattice

G:	(111)	(200)	(220)	(311)	(222)	$\times 2\pi/a$
distance $\frac{1}{2}\lvert\mathbf{G}\rvert$:	0.866	1	1.414	1.658	1.732	$\times 2\pi/a$

We use (2.1) to calculate k_F: there are 32 valence electrons in the cubic unit cell of volume a^3 [i.e., 4 valence electrons for each of 8 atoms]; hence

$$k_F{}^3 = 3\pi^2 \times \frac{32}{a^3}, \qquad \text{or} \qquad k_F = 1.563 \times 2\pi/a.$$

This just exceeds 1.414, so the free-electron Fermi sphere would slightly overlap the 12 zone boundaries bisecting (220), $(\bar{2}20)$, $(2\bar{2}0)$, $(\bar{2}0\bar{2})$, ... In Fig. 2.10 the zone boundaries are labelled by the reciprocal lattice vectors that they bisect. The (220) zone boundaries enclose a so-called *Jones zone* whose volume is $16 \times (2\pi/a)^3$, and which accommodates the required 4 electrons per atom.

As a bonus we might expect the occupied states of highest energy [top of the valence band] to lie at the furthest corners of the Jones zone, where the kinetic energy is greatest. One such corner is at $\mathbf{k} = (200)$, which is equivalent, after translation by $\mathbf{G} = (\bar{2}00)$, to the centre of the *Brillouin* zone, $\mathbf{k} = \mathbf{0}$. This result from the nearly-free electron model happens to be correct, but predictions for the conduction-band minimum are not, in general.

2.4.3 Binding energy of metals

To investigate the stability of different crystal structures it is helpful to estimate the different contributions to the total energy of the solid, relative to the energy of well-separated ions and valence electrons. We shall do this in the free-electron picture, expressing the result—the binding energy per ion—in terms of the Fermi wave vector of the gas.

Kinetic energy:

The free electron kinetic energy has been worked out in PHYS20252 and again in PHYS30151. Here we simply quote the result

$$E_{\text{kin}} = \tfrac{3}{5} N_e E_F = \tfrac{3}{5} N_e \cdot \hbar^2 {k_F}^2 / 2m.$$

Since each of the N_i ions has contributed Z electrons to the gas, the total number of electrons is $N_e = N_i Z$, giving

$$E_{\text{kin}} / N_i = \tfrac{3}{10} Z \hbar^2 {k_F}^2 / m \tag{2.25}$$

as the electron kinetic energy per ion.

Electrostatic energy:

We make a rather drastic approximation that the ions can be regarded as point charges. The calculation of the electrostatic energy of point ions in a uniform electron gas is fairly involved, and analogous to the calculation of the electrostatic energy of ionic solids. We shall not go into details, but will quote the result in the conventional form

$$E_{\text{Coul}} / N_i = -\frac{1}{2} \frac{e^2}{4\pi\epsilon_0} \frac{Z^2 \alpha}{r_0}, \tag{2.26}$$

where r_0 is the radius of a so-called *atomic sphere*, which encloses the point ion and Z electrons. The dimensionless constant α is characteristic of the particular crystal structure, like the Madelung constant for ionic solids. The radius of the atomic sphere is related (see Problem 2.4) to k_F by

$$r_0 = (9\pi Z/4)^{1/3} / k_F. \tag{2.27}$$

Most simple metals adopt one of the three structures FCC, BCC or HCP (hexagonal close-packed), for which $\alpha = 1.79$; the atoms are close together, giving the lowest (i.e., most negative) electrostatic energy. The more open simple-cubic structure ($\alpha = 1.76$) and diamond structure ($\alpha = 1.67$) are not favoured; the latter is stabilized by the kind of "zone effect" that is discussed below for the pentavalent semi-metals.

Total electronic energy:

Using (2.25), (2.26) and (2.27) we find the total energy per ion to be

$$E_{\text{tot}} / N_i = \frac{3Z\hbar^2 {k_F}^2}{10\, m} - \frac{e^2}{4\pi\epsilon_0} \frac{\alpha Z^2 k_F}{(18\pi Z)^{1/3}}.$$

The value of k_F that minimizes the total energy can be found by differentiating the last expression; the final results are

$$k_F^* = \frac{5}{3} \frac{\alpha Z^{2/3}}{(18\pi)^{1/3}} \frac{m}{\hbar^2} \frac{e^2}{4\pi\epsilon_0} \qquad \text{and} \qquad E_{\text{tot}}^* / N_i = -\frac{3Z\hbar^2 {k_F^*}^2}{10\, m}$$

for the optimum Fermi wave vector k_F^* and the corresponding minimum energy E_{tot}^*.

	Z	k_F^* [Å^{-1}]	$-E_{\text{tot}}^*/N_i$ [eV]
sodium	1	1.47 (0.91)	4.9 (6.3)
magnesium	2	2.33 (1.37)	24.8 (24.4)
aluminium	3	3.05 (1.75)	63.8 (56.3)

Predictions for the binding energy are not too far from the experimental values (given in brackets), but the Fermi wave vectors are much less accurately estimated. Note also that our results depend only on valence, so they cannot represent the variation among elements of the same group [column of the periodic

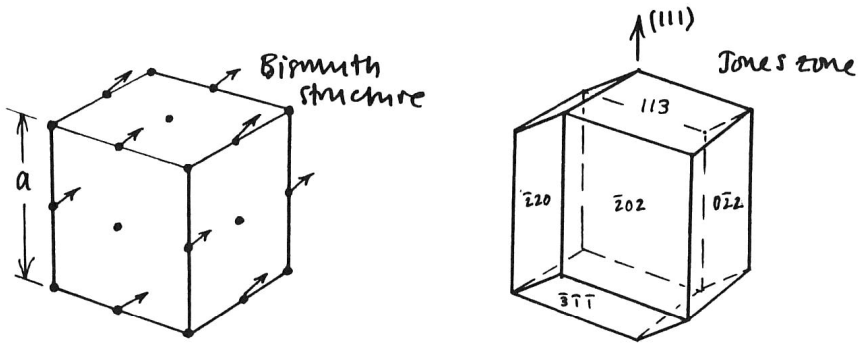

Figure 2.11: Crystal structure (left) and Jones zone (right) for the pentavalent semimetals

table]: in reality k_F varies by about a factor of two between the alkali metals lithium and caesium. Much of the discrepancy with experiment is caused by our approximating the positive ions by point charges. Valence electrons are largely excluded from the ion cores, which would raise the electron kinetic energy if the solid did not expand. The expansion lowers the mean density of electrons, and so leads to a smaller k_F.

2.4.4 Jones theory of Group V elements

Arsenic, antimony and bismuth are pentavalent metals, but their physical properties at low temperature [e.g., resistivity and heat capacity] suggest that the number of carriers is exceedingly small. In Bi the number densities of electrons and holes are $n_e = n_h = 3 \times 10^{23}\,\mathrm{m}^{-3}$, about 5 orders of magnitude less than in ordinary metals. Materials of this kind are called *semi-metals*. They lie on the borderline of being insulators, so that their electronic structure should correspond to a nearly filled zone.

The crystal structure is interesting, being *not quite* simple cubic; see Fig. 2.11. Alternate atoms are shifted a little along the [111] direction, making the structure FCC with two atoms per lattice point. (We neglect a secondary distortion which in Bi reduces the angle between the FCC primitive vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ from the ideal 60° to about 57°.) Without the distortion the materials would be good metals, having 5 electrons per simple-cubic cell. Unlike the case of diamond, it is hard to explain this structure in terms of directed valence bonds, but we can still attempt an explanation based on zone effects.

There are 40 electrons in the FCC unit cell, so that

$$k_F{}^3 = 3\pi^2 \times \frac{40}{a^3}, \qquad \text{or} \qquad k_F = 1.684 \times 2\pi/a.$$

Referring Table 2.1, we find that the free-electron sphere would protrude from the (220) and (311) zone boundaries. Remarkably, these planes define a zone that holds the required 5 electrons per atom; this is *not* at all easy to verify. The zone is shaped like the cell of a honeycomb: planes bisecting $(2\bar{2}0)$, $(0\bar{2}2)$, $(\bar{2}02)$, $(\bar{2}20)$, $(20\bar{2})$, $(02\bar{2})$ form a hexagonal tube which is capped at either end by bisectors of (311), (131), (113) and $(\bar{3}\bar{1}\bar{1})$, $(\bar{1}\bar{3}\bar{1})$, $(\bar{1}\bar{1}\bar{3})$. The distortion away from simple cubic is a striking example of a structural change induced by electrons. For if Bi was simple cubic with edge $a/2$ (half the edge of the FCC cell), the reciprocal lattice vectors would be the usual

$$\mathbf{G} = (p_1, p_2, p_3) \times \frac{2\pi}{a/2} = (2p_1, 2p_2, 2p_3) \times 2\pi/a,$$

corresponding only to the even-numbered **G**s of the FCC lattice, such as (220). The odd-numbered **G**s, such as (311), occur only because of the distortion towards FCC. Electron energies are lowered in the neighbourhood of the (311) zone boundaries, so that the distortion is favoured by the reduction in the total electron energy.

2.4.5 Structure of alloys

As we have seen for the case of bismuth, the total electronic energy can be reduced if the free-electron Fermi surface lies at a zone boundary. The stability of certain metal alloy structures is generally attributed to zone effects of this kind [H. Jones 1934]. Experimentally it has been found, for example, that an alloy of copper (valence 1) with zinc (valence 2) adopts the BCC structure (so-called β-brass) only if the ratio of electrons to atoms is in the neighbourhood of 1.5. The same ratio is found for the BCC structure of a number of other systems, including the alloys of copper with aluminium (valence 3) and with silicon (valence 4) [W. Hume-Rothery, 1930s].

It is worth understanding the reasoning that gives a qualitative explanation of this observation; the discussion in lectures tends to be quite brief.

Suppose that an alloy with a given number of electrons per atom can adopt one of two crystal structures, A and B. The alloy will adopt the structure that leads to the lower total energy, and we want to argue from the nearly-free electron model that the energies of different structures can be different, depending on how the (fictitious) free-electron Fermi surface lies in relation to the first Brillouin zone boundary. Bearing that in mind, we imagine that in A the Fermi sphere only just touches the first Brillouin zone boundary, so that all electrons are in the first band. In the alternative structure B [which has a differently-shaped Brillouin zone], we suppose that the Fermi sphere has a portion that protrudes from some part of the first Brillouin zone boundary; if the crystal potential is not too large there will be electrons in the second band. In A, electrons at the portion of the Fermi surface near the zone boundary have their energy lowered by the crystal potential: the total energy of the electron gas is therefore reduced by the crystal potential. In B there is a similar energy-lowering effect for electrons in the *first* band, but there are also electrons in the *second* band that have their energy *raised* by the crystal potential. This increase in their energy tends to cancel the reduction in total energy of electrons in the first band, so the net energy-lowering effect is smaller for structure B than for the structure A. Other things being equal,[10] structure A is the one with the lower energy, and hence is the one that should be observed.

Now we return to the specific case of the BCC alloys. The "preferred" ratio of electrons to atoms in the BCC structure (i.e., the ratio that leads to the lowest energy, when compared with other structures) is calculated approximately by finding the electron density needed to make the free-electron Fermi sphere touch the zone boundary. The reciprocal of the BCC lattice with cubic lattice constant a is an FCC lattice with cubic lattice constant $4\pi/a$; see Problem 1.5, or follow an approach similar to that used for the FCC lattice in Sec. 2.4.2. The shortest reciprocal lattice vectors join a corner of the FCC cell to the midpoint of one of the adjoining faces of the cell: $(011)2\pi/a$ is one of these, and there are 11 more vectors related to this by cubic symmetry. The bisectors of these 12 reciprocal lattice vectors are the faces of the first Brillouin zone, which is a rhombic dodecahedron.[11] Measured from $\mathbf{k} = \mathbf{0}$, the perpendicular distance to each of these zone boundaries is $\sqrt{2}\,\pi/a$; so, if the Fermi sphere just touches them, $k_F = \sqrt{2}\,\pi/a$. We can use this to calculate the number-density of electrons,

$$n = \frac{k_F^3}{3\pi^2} = \frac{2\sqrt{2}\,\pi}{3a^3},$$

which corresponds to $2\sqrt{2}\,\pi/3 \simeq 2.96$ electrons per cubic unit cell of side a, or 1.48 electrons per atom, since there are 2 atoms per cubic unit cell of the BCC structure. The figure of 1.48 electrons per atom is in good agreement with the experimental observation. For other, more complex, structures the numerical agreement is not always so good.[12]

[10] Perhaps the surprising conclusion to be drawn from the observations on real alloy structures is that "other things" quite often *are* more-or-less equal.

[11] This is the same *shape* as for the Jones zone of the diamond structure, illustrated on the right of Fig. 2.10. The *volume*, of course, is different—by a factor of 8, if you take $2\pi/a$ to have the same value in each case.

[12] If you are interested, many more details can be found in N. F. Mott and H. Jones, *Theory of the Properties of Metals and Alloys*.

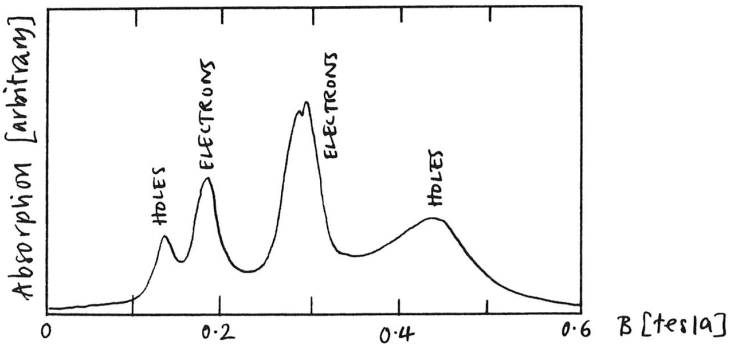

Figure 2.12: Cyclotron resonance in silicon at 24 GHz and 4 K; after Dresselhaus et al., 1955.

2.5 Cyclotron resonance

In a uniform magnetic field B a classical free electron moves in a circle (or a helix) with angular frequency ω_c. If the radius of the orbit is a, the speed will be $\omega_c a$ and the Lorentz force of magnitude $e\omega_c aB$ will give rise to the centripetal acceleration. Hence[13]

$$m_0\omega_c^2 a = e\omega_c aB, \qquad \text{or} \qquad \omega_c = eB/m_0.$$

When an electromagnetic (e.g., microwave) field of fixed frequency is applied, energy will be absorbed by the electron if its orbital frequency matches that of the applied field. By varying the uniform magnetic field and monitoring the microwave absorption it would be possible to calculate m_0 from the position of the peak in the absorption plotted as a function of B.

The basic principle is the same for cyclotron resonance measurements of effective masses of carriers in semiconductors. In general, a cyclotron effective mass m_c can be defined from the resonance frequency and the applied field, $m_c = eB/\omega_c$.

To obtain sharp resonance peaks the carriers must complete at least a few orbits before being scattered. Pure samples and low temperatures (about 4 kelvin) must therefore be used to reduce scattering by impurities and phonons. But unlike free space, crystals are not isotropic, so that orbits are not necessarily circular and the analysis is usually more complicated. Extra information is needed, such as the positions of the resonances as the direction of the magnetic field is varied. An example of this kind of analysis is given in Problem 2.5, but you should read the next two sections before attempting it.

2.5.1 Effective mass tensor

First we need to investigate the form of the electron dispersion relation near a minimum of the conduction band. Although it is not easy to calculate $E(\mathbf{k})$ for particular semiconductors, some general properties depend only on the symmetry of the solid, and we consider these below. Similar considerations apply to holes near the top of the valence band.

To simplify things we assume that the atomic arrangement in the semiconductor is simple cubic. The first Brillouin zone is then a cube of side $2\pi/a$, where a is the lattice constant. We only need to consider one such cube because the energy is periodic in $\mathbf{k}$: $E(\mathbf{k}) = E(\mathbf{k}+\mathbf{G})$, where $\mathbf{G}$ is a reciprocal lattice vector.

The precise form of $E(\mathbf{k})$ depends on the position of the minimum in the zone; see Fig. 2.13. Suppose that it occurs at the zone centre, $\mathbf{k} = \mathbf{0}$. We expand the energy in a power series with respect to $\mathbf{k}$ about this point. The first non-zero terms are

$$E(\mathbf{k}) = E(\mathbf{0}) + \frac{\hbar^2}{2m_e}\left(k_x^2 + k_y^2 + k_z^2\right) + \text{higher powers of } k, \tag{2.28}$$

[13] From now on we make clear when the free-electron mass (previously denoted by m) is being used by giving it a subscript zero, $m_0 \equiv m = 9.1 \times 10^{-31}$ kg.

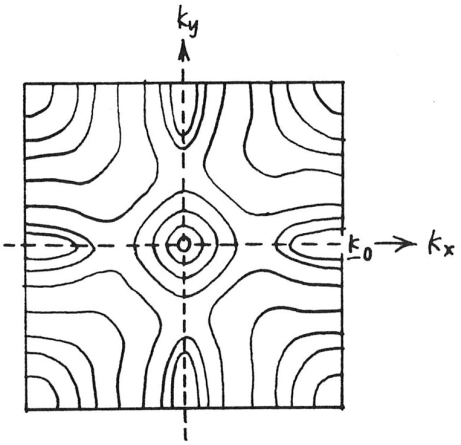

Figure 2.13: Contours of constant energy in the first Brillouin zone of a hypothetical simple-cubic semiconductor.

where the coefficient has been written as $\hbar^2/2m_e$, by analogy with the energy of a free electron, $E = \hbar^2k^2/2m_0$. No other second-order terms are possible. Products like k_xk_y cannot occur because they would change sign under a reflection such as $(k_x, k_y, k_z) \to (-k_x, k_y, k_z)$. The coefficients of k_x^2, k_y^2 and k_z^2 must all be the same if $E(\mathbf{k})$ is to be unchanged by rotations of 90° about k_x, k_y, k_z.

Exercise 2.10:

Explain why the zone centre must always be an extremum of $E(\mathbf{k})$; i.e., why terms linear in $\mathbf{k}$ cannot occur.

Close to $\mathbf{k} = \mathbf{0}$, the surfaces of constant energy are spherical. In the diagram they are represented by circles. As we move out from the centre, the higher powers of $\mathbf{k}$ become more important and the surfaces are no longer spheres, though they still have the same symmetry as the cubic Brillouin zone.

Suppose there is another local minimum in $E(\mathbf{k})$ at the centre of the zone face, $\mathbf{k} = \mathbf{k}_0 = (\pi/a, 0, 0)$. We can expand the energy in powers of $\Delta\mathbf{k} = (\mathbf{k} - \mathbf{k}_0)$, the "distance" from the minimum,

$$E(\mathbf{k}) \simeq E(\mathbf{k}_0) + \frac{\hbar^2}{2}\left(\frac{\Delta {k_x}^2}{m_l} + \frac{\Delta {k_y}^2 + \Delta {k_z}^2}{m_t}\right). \tag{2.29}$$

Here there is no need for the coefficient of $\Delta {k_x}^2$ to be the same as that of $\Delta {k_y}^2$ or $\Delta {k_z}^2$, though the latter two must have the same coefficient to allow rotation of 90° about k_x.

Exercise 2.11:

Repeat the last exercise for the point $\mathbf{k}_0$ in the Brillouin zone.

The surfaces of constant energy are ellipsoids of revolution about k_x.[14] By symmetry, there must be similar ellipsoids too on the k_y and k_z zone faces. At moderate temperatures (or in a doped semiconductor), electrons will be present near the conduction band minima (the states of lowest energy), and we often speak of *ellipsoidal pockets* of electrons at these points in the zone; see Fig. 2.14.

Both cases (2.28) and (2.29) can be represented by the same general expression

$$E(\mathbf{k}) \simeq \text{const.} + \frac{\hbar^2}{2}\sum_{i,j=1}^{3}\left[\mathbf{M}^{-1}\right]_{ij}\Delta k_i\,\Delta k_j, \tag{2.30}$$

[14] $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ is the equation of an ellipsoid with semi-axes a, b and c. If $b = c$, as in (2.29), we have an ellipsoid of revolution about the x-axis.

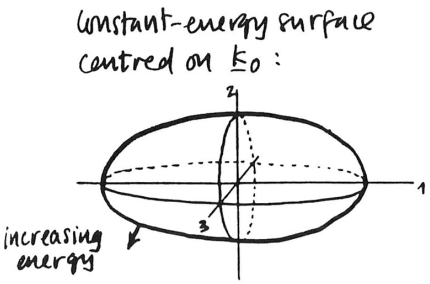

Figure 2.14: Ellipsoidal surface of constant energy.

where $\mathbf{M}$ is a symmetric 3×3 matrix, the *effective mass tensor*. For example, in (2.28) the mass tensor is just the unit matrix multiplied by m_e; while in (2.29)

$$\mathbf{M} = \begin{pmatrix} m_l & 0 & 0 \\ 0 & m_t & 0 \\ 0 & 0 & m_t \end{pmatrix}. \tag{2.31}$$

Earlier, in Sec. 2.3.3, we have related the effective mass of an electron to the reciprocal of the curvature of E: $m_e = \hbar^2/[\partial^2 E/\partial {k_x}^2]$. The geometrical meaning of (2.31) is that the curvature of the band can be different in different directions. One *physical* consequence is that an electron can have different accelerations if a force of a given magnitude is applied in different directions. In this context m_l and m_t are sometimes called the *longitudinal* and *transverse* effective masses of the electron, because they determine the electron's acceleration when the force is respectively along the axis of the ellipsoid or perpendicular to it.

We can work out the relation between the electron group velocity $\mathbf{v}$ and crystal momentum implied by (2.30). For the ith component,

$$v_i = \frac{1}{\hbar}\frac{\partial E}{\partial k_i} = \hbar \sum_{j=1}^{3} \left[\mathbf{M}^{-1}\right]_{ij} \Delta k_j.$$

In vector notation this can be written

$$\mathbf{v} = \hbar \mathbf{M}^{-1} \Delta \mathbf{k}, \qquad \text{or} \qquad \hbar \Delta \mathbf{k} = \mathbf{M}\,\mathbf{v}.$$

Note that *matrix* multiplication is implied, so that the velocity and the crystal momentum need not be parallel.

2.5.2 Calculation of the cyclotron frequency

The Lorentz force on an electron in a magnetic field is

$$\mathbf{F} = \hbar \dot{\mathbf{k}} = -e\mathbf{v}(\mathbf{k}) \times \mathbf{B}. \tag{2.32}$$

No work is done by the $\mathbf{B}$-field, as $\dot{E} = \mathbf{F} \cdot \mathbf{v} = 0$, so the motion in $\mathbf{k}$-space is along contours of constant energy perpendicular to the magnetic field $\mathbf{B}$. An electron on an ellipsoidal energy surface will therefore move on an elliptical orbit in $\mathbf{k}$-space. Integrating (2.32) once with respect to time gives

$$\hbar \mathbf{k} = -e\mathbf{r} \times \mathbf{B} + \text{const.},$$

so that the orbit in ordinary space is also elliptical, but rotated through 90°.

The conduction band minima of silicon are very like those described for the simple cubic zone face. In the resonance curves plotted for silicon, up to three peaks can occur for electrons. This fact can be explained immediately: for a general orientation of the field $\mathbf{B}$ there are *three* possible orbits for electrons—one for each ellipsoid; see Fig. 2.15.

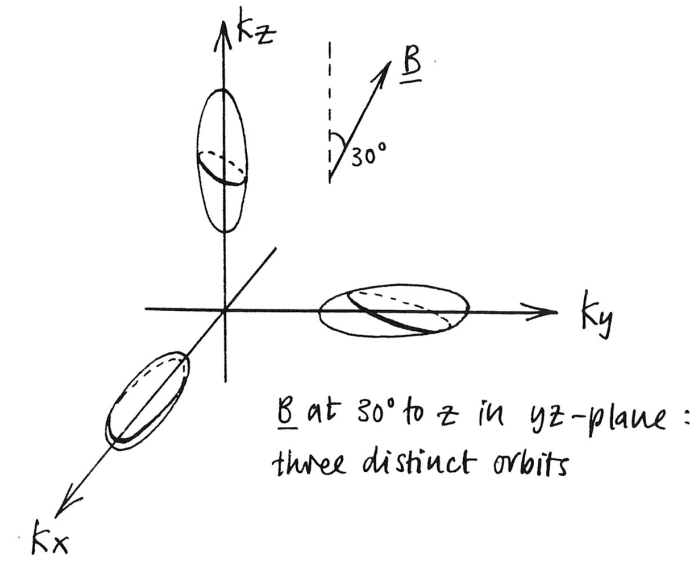

Figure 2.15: **B** in y, z plane at 30° to z axis: three distinct orbits/resonances.

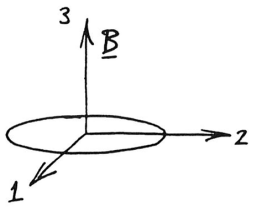

Figure 2.16: Elliptical contour of constant E, perpendicular to **B**.

Each of these orbits is an ellipse with a different axial ratio, unless **B** happens to lie in a special direction; say, along one axis, or midway between two axes. What we need is to relate the cyclotron frequency to the shape of the electron orbit.

We start from the equation of motion (2.32). On a plane perpendicular to the magnetic field the energy is given by

$$E = \frac{\hbar^2}{2m_1}k_1^2 + \frac{\hbar^2}{2m_2}k_2^2 + \text{const.},$$

where k_1 and k_2 are components along the axes of the ellipse, measured from the band minimum at $\mathbf{k}_0$; see Fig. 2.16. The components of velocity are given by

$$\hbar k_1 = m_1 v_1 \quad \text{and} \quad \hbar k_2 = m_2 v_2,$$

which can be substituted into (2.32) to give

$$m_1 \dot{v}_1 = -ev_2 B \qquad \text{and} \qquad m_2 \dot{v}_2 = +ev_1 B.$$

Eliminating v_2 from the first of these equations by means of the second we obtain the equation of simple harmonic motion,

$$\ddot{v}_1 = -e^2 B^2 v_1 / m_1 m_2 = -\omega_c^2 v_1,$$

with frequency $\omega_c = eB/\sqrt{m_1 m_2}$.

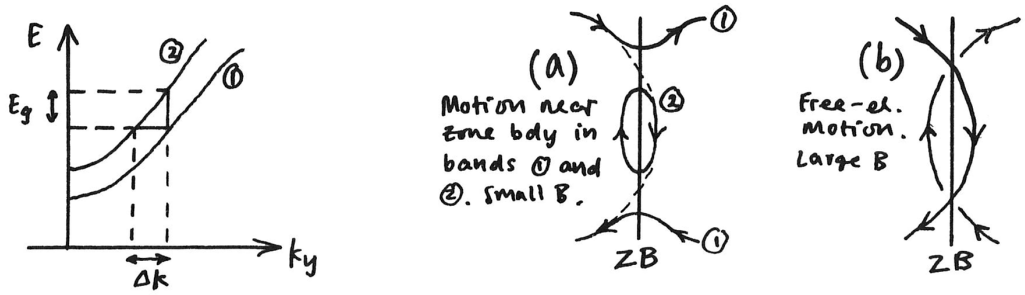

Figure 2.17: Overlapping bands (left) for a simple-cubic divalent metal, plotted as a function of k_y for fixed $k_x = \pi/a$, $k_z = 0$. Cyclotron orbits in k-space for (a) a weak magnetic field and (b) a field strong enough to allow magnetic breakthrough.

2.5.3 Cyclotron resonance in metals

Cyclotron motion is the basis of a number of techniques for obtaining information about the band structure near the Fermi surface in metals. Here we mention just one, due to Azbel' and Kaner [1956].

A feature of the use of cyclotron resonance in metals is that the exciting microwave field can penetrate only a short distance, the skin depth δ, into the metal, where δ is much shorter than the mean free path of the electrons. Hence, if the magnetic field is applied parallel to the metal surface, electrons on their closed [more generally, helical] orbits see the oscillating microwave field only once per cycle, when they are within distance δ of the surface. Resonance can occur provided the frequency ω_c of the cyclotron motion is a sub-multiple of the frequency ω of the applied field, so that the electrons always see an electric field of the same phase when they arrive back at the surface. We have

$$\omega = n\omega_c = neB/m_c, \quad \text{or} \quad \frac{1}{B} = \frac{ne}{\omega m_c} \quad \text{with} \quad n = 1, 2, 3, \ldots;$$

for fixed ω, each cyclotron orbit gives a number of different resonance peaks, equally spaced in $1/B$.

Since the electrons must absorb energy in cyclotron resonance, the Pauli principle allows only those at the Fermi surface to participate, as only they have unoccupied states to move to. Fermi surfaces can have complicated shapes (Problem 2.6 (b) is complicated enough), so that, unlike the case of semiconductors, it can be quite difficult to disentangle which resonances correspond to which closed orbits.

2.5.4 Magnetic breakthrough: failure of the semiclassical approximation

In the semiclassical picture an electron or hole always has a definite position and crystal momentum, and never makes transitions between bands. This approximation can be expected to fail in a strong enough field: the effect of the applied field will eventually exceed that of the periodic crystal potential. One might guess that this would happen when the quantum of energy $\hbar\omega_c$, associated with the cyclotron motion, exceeds the energy splitting $E_g = 2|V_\mathbf{G}|$ at a zone boundary. Actually, even for a small splitting[15] of 0.01 eV this would require a practically unattainable field of order 100 tesla. Nevertheless, magnetic-field–induced transitions between bands are observed even for moderate field strengths. The phenomenon is called *magnetic breakthrough* (or *breakdown*), and occurs in metals for

$$\hbar\omega_c > E_g^2/E_F, \tag{2.33}$$

where E_F is the Fermi energy. Compared with our earlier guess, this is greatly reduced by the factor E_g/E_F; for $E_g = 0.01\,\text{eV}$ and $E_F = 10\,\text{eV}$, the necessary field is about 0.1 tesla.

[15] Small energy splittings, in the range 0.001–0.01 eV, are found in certain HCP metals, such as magnesium. The splitting is due in these cases to the spin–orbit interaction, a small relativistic correction to the bands.

It is not too hard to understand the criterion (2.33) in a qualitative way, using the uncertainty principle of quantum mechanics. We have found in Sec. 2.5.2 that cyclotron orbits in **r**-space are geometrically similar to the contours of constant energy followed by the carrier in **k**-space. The orbit is rotated by 90° and scaled by a factor depending on the field so that, for example, $\hbar k_y = eBx + \text{const}$. Hence the usual uncertainty principle $\Delta k_x \, \Delta x > 1$ becomes an uncertainty principle relating the two components of the wave vector,

$$\Delta k_x \, \Delta k_y > eB/\hbar.$$

In the presence of a magnetic field, therefore, the wave vector cannot be defined with a precision greater than about $\Delta k_c = \sqrt{eB/\hbar}$.

Now we envisage a case similar to that illustrated on the left of Fig. 2.17: two overlapping nearly-free electron bands separated in energy by E_g. As shown in the figure, states of the same energy in the two bands occur at different wave vectors, the separation being Δk. If Δk is less than the uncertainty in **k** it is an over-refinement to say that the electron is in a definite band [case (a) in Fig. 2.17, for small B]; instead, we might expect to find cyclotron motion (and resonance spectra) characteristic of free electrons in a magnetic field [case (b), large B]. Quantitatively this requires

$$\Delta k_c > \Delta k = \frac{E_g}{\partial E/\partial k}.$$

At the Fermi energy the velocity is $v_F = \hbar^{-1} \partial E/\partial k$. Using this and the expression for Δk_c in the last inequality we find

$$\sqrt{eB/\hbar} > E_g/\hbar v_F \qquad \text{or} \qquad eB/\hbar > \left(E_g/\hbar v_F\right)^2 \sim m_0 E_g^2/\hbar^2 E_F.$$

When making a rough estimate we should not be too worried about being wrong by a factor of 2, so we have replaced ${v_F}^2$ by E_F/m_0 to get the last expression; multiplying each side by $\hbar^2/m_0$ finally gives the criterion (2.33).

Chapter 3

Magnetism

The magnetic properties of matter include its response to an applied magnetic field. Since this is governed largely by the response of the electrons (effects due to nuclear magnetic moments are at least six orders of magnitude smaller) it is useful to study the motion of charged particles in a magnetic field.

3.1 Electrons in a magnetic field

The classical Newtonian description of the motion of charges is easy,

$$m\dot{\mathbf{v}} = Q(\boldsymbol{\mathcal{E}} + \mathbf{v} \times \mathbf{B}) \,, \tag{3.1}$$

where $\boldsymbol{\mathcal{E}}$ and $\mathbf{B}$ are applied electric and magnetic fields and m and Q are the mass and electric charge of the particle. The problem is how to include the effect of the velocity-dependent Lorentz force in a Schrödinger equation

$$\hat{H}\psi(\mathbf{r}) = E\,\psi(\mathbf{r}) \,.$$

Usually we simply take $\hat{H}$ to be the classical expression for the energy of the particle and replace the kinetic energy by the operator $\hat{\mathbf{p}}^2/2m$. But a time-independent magnetic field does no work on the electron ($\mathbf{v} \times \mathbf{B}$ is perpendicular to $\mathbf{v}$) and so has no direct effect on the energy

$$E = \frac{mv^2}{2} + Q\phi(\mathbf{r}) \,, \tag{3.2}$$

where $\phi(\mathbf{r})$ is the scalar potential of the electric field, $\boldsymbol{\mathcal{E}} = -\nabla\phi$. It is worth recalling here that $\hat{H}$ plays a dual rôle in quantum mechanics: besides being the energy, it also determines the time development of the system via the time-dependent Schrödinger equation. Before we attempt any quantum mechanics we will show how the energy also determines the time development in *classical* mechanics.[1]

3.1.1 Hamiltonians in classical mechanics

A wide range of classical mechanical systems can be described using *Hamilton's equations*, which we write down for the simplest case of one particle:

$$\dot{\mathbf{r}} = \frac{\partial H}{\partial \mathbf{p}} \tag{3.3}$$

$$\dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{r}} \,. \tag{3.4}$$

[1] For this course, the only important result in Sections 3.1.1 and 3.1.2 is Eq. (3.5), the Hamiltonian of a charge in an electromagnetic field. Some students have already met this, but I think it is useful for everyone to have an idea of where this mysterious-looking result comes from.

The Hamiltonian energy function $H(\mathbf{p},\mathbf{r})$ introduced here depends on the momentum and position of the particle; $\partial H/\partial\mathbf{r}$ is just an alternative notation for ∇H, and $\partial H/\partial\mathbf{p}$ similarly means the derivative with respect to components of $\mathbf{p}$.

We verify that a particle in a potential $V(\mathbf{r})$ can be described by (3.3)–(3.4). In this case the Hamiltonian is $p^2/2m + V(\mathbf{r})$, and Hamilton's equations give

$$\dot{\mathbf{r}} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{p}/m \qquad \text{and} \qquad \dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{r}} = -\nabla V\,;$$

the first of these simply says that the momentum is $m\mathbf{v}$; the second is Newton's second law. We can think of the complete Hamiltonian as a generalization of the potential, $V(\mathbf{r})$: as with V, the derivative of H with respect to $\mathbf{r}$ gives the rate of change of $\mathbf{p}$; but in addition the derivative with respect to $\mathbf{p}$ give the rate of change of $\mathbf{r}$. This symmetry in the treatment of position and momentum is one of the attractive features of Hamilton's method.

Not *all* systems can be described by a Hamiltonian, just as they cannot all be described by a Lagrangian; frictional forces, for example, cannot be treated in this way. Nevertheless, if we *can* write our equations of motion in the *canonical form* (3.3)–(3.4) it is generally possible to go over to the quantum case by replacing $\mathbf{p}$ by $-i\hbar\nabla$.

3.1.2 Classical Hamiltonian of a charge in a magnetic field

We need to find a Hamiltonian that describes velocity-dependent forces, such as those in (3.1). The energy (3.2), expressed in terms of $\mathbf{r}$ and $\mathbf{v}$, does not involve the magnetic field, while $H(\mathbf{p},\mathbf{r})$ (which is also supposed to be the energy) must involve $\mathbf{B}$, since this appears in the equation of motion. It seems likely, therefore, that the magnetic field should enter via the dependence of $\mathbf{p}$ on $\mathbf{v}$. We write

$$\mathbf{p} = m\mathbf{v} + Q\mathbf{A}\,,$$

where $Q\mathbf{A}$ is expected to vanish if $\mathbf{B} = \mathbf{0}$, or if the particle has no charge. When this form is substituted in the left-hand side of (3.4) time derivatives of $\mathbf{A}$ appear, in addition to $m\dot{\mathbf{v}}$. These extra terms (when transferred to the other side of the equation) are contributions to the force on the electron. If $\mathbf{A}$ depends on $\mathbf{v}$ this force will depend on acceleration, in contradiction to (3.1); hence we expect $\mathbf{A}$ to depend only on position. From (3.3) we then find

$$\frac{\partial H}{\partial \mathbf{p}} = \dot{\mathbf{r}} = \big(\mathbf{p} - Q\mathbf{A}(\mathbf{r})\big)/m\,,$$

which can be integrated with respect to $\mathbf{p}$ to obtain

$$H = \frac{(\mathbf{p} - Q\mathbf{A})^2}{2m} + V(\mathbf{r})\,, \tag{3.5}$$

where $V(\mathbf{r})$ is an (as yet) undetermined function of position.

We finally verify that (3.5) leads to the Lorentz force equation when substituted into (3.4). A straightforward calculation of the derivatives gives

$$\begin{aligned}\dot{\mathbf{p}} = m\dot{\mathbf{v}} + Q(\mathbf{v}\cdot\nabla)\mathbf{A} &= -\frac{\partial H}{\partial \mathbf{r}} \\ &= -\nabla V + Q\sum_{i=1}^{3}(p_i - QA_i)\,\nabla A_i \\ &\equiv -\nabla V + Q\sum_{i=1}^{3} v_i \nabla A_i\,,\end{aligned}$$

so that, after taking the term $Q(\mathbf{v}\cdot\nabla)\mathbf{A}$ to the other side of the equation,

$$\begin{aligned} m\dot{\mathbf{v}} &= -\nabla V + Q\sum_{i=1}^{3} v_i \nabla A_i - Q(\mathbf{v}\cdot\nabla)\mathbf{A} \\ &= -\nabla V + Q\mathbf{v}\times[\nabla\times\mathbf{A}] \\ &\equiv Q\boldsymbol{\mathcal{E}} + Q\mathbf{v}\times\mathbf{B}\,; \end{aligned}$$

the second line is obtained by inspection of the formula for the vector triple product, $\mathbf{P}\times[\mathbf{R}\times\mathbf{S}] = (\mathbf{P}\cdot\mathbf{S})\mathbf{R} - (\mathbf{P}\cdot\mathbf{R})\mathbf{S} = \sum_i P_i\mathbf{R}S_i - (\mathbf{P}\cdot\mathbf{R})\mathbf{S}$. The second and third lines show that $Q\boldsymbol{\mathcal{E}} = -\nabla V$ and $\mathbf{B} = \nabla\times\mathbf{A}$: we can now identify $V/Q = \phi$ with the scalar potential of the electric field and $\mathbf{A}$ with the magnetic vector potential.

Note that, because of the modified relation between $\mathbf{p}$ and $\mathbf{v}$, the Hamiltonian (3.5) is precisely the anticipated energy function (3.2), but expressed in different variables.

Exercise 3.1:

> Extend the treatment to allow for time-varying fields, and hence show that $\boldsymbol{\mathcal{E}} = -\nabla\phi - \partial\mathbf{A}/\partial t$ in general.

3.1.3 No magnetism in classical physics

The result from the last section allows us to calculate the *additional* magnetic field generated by a system in an applied magnetic field $\mathbf{B}$. We might expect some such effect: free charges in a magnetic field move in circular (or helical) orbits, and so should behave like little current loops. According to the Biot–Savart law, the magnetic field induced at $\mathbf{r}'$ by a charge Q moving with velocity $\mathbf{v}$ at $\mathbf{r}$ is given by

$$\mathbf{B}_{\text{ind}}(\mathbf{r}',\mathbf{r},\mathbf{v}) = \frac{\mu_0}{4\pi}\frac{Q\mathbf{v}\times(\mathbf{r}'-\mathbf{r})}{|\mathbf{r}'-\mathbf{r}|^3}\,.$$

More generally, the induced magnetic field will be the sum of many such terms, taken over all the electrons of the system. In practice we are interested not in the instantaneous value of this sum, which would be a very rapidly fluctuating quantity, but in its value $\langle\mathbf{B}_{\text{ind}}\rangle$ averaged over the motion of the electrons at thermal equilibrium. We might anticipate one of the following three possibilities:

(a) $\langle\mathbf{B}_{\text{ind}}\rangle$ parallel to the applied field, and vanishing as $B\to 0$. This case is called *paramagnetism* and is found in many simple metals, such as sodium, and in salts of the transition metals.

(b) $\langle\mathbf{B}_{\text{ind}}\rangle$ *anti*-parallel to the applied field, and vanishing for $B\to 0$. This is called *diamagnetism* and it is found experimentally in rare-gas solids (frozen neon, say), covalently-bonded insulators, semimetals such as bismuth, and very often in gases of carriers in semiconductors.

(c) $\langle\mathbf{B}_{\text{ind}}\rangle$ parallel to the applied field, but *not* vanishing for small fields: the material is said to have a *spontaneous magnetization*, or to be *ferromagnetic*. The archetypal ferromagnet is iron, but there are other ferromagnetic transition metals and rare-earths, and many alloys.

In fact *none* of the above occur in classical physics, as we can see by forming the average $\langle\mathbf{B}_{\text{ind}}\rangle$ using the Boltzmann distribution. We have

$$\begin{aligned} \langle\mathbf{B}_{\text{ind}}\rangle &= \frac{\iint \mathbf{B}_{\text{ind}}(\mathbf{r}',\mathbf{r},\mathbf{v})\,\exp[-H(\mathbf{p},\mathbf{r})/k_BT]\,d^3\mathbf{p}\,d^3\mathbf{r}/h^3}{\iint \exp[-H(\mathbf{p},\mathbf{r})/k_BT]\,d^3\mathbf{p}\,d^3\mathbf{r}/h^3} \\ &= \frac{\iint \mathbf{B}_{\text{ind}}(\mathbf{r}',\mathbf{r},\mathbf{v})\,\exp\!\left[-\left(\tfrac{1}{2}mv^2+V(\mathbf{r})\right)/k_BT\right]d^3\mathbf{v}\,d^3\mathbf{r}}{\iint \exp\!\left[-\left(\tfrac{1}{2}mv^2+V(\mathbf{r})\right)/k_BT\right]d^3\mathbf{v}\,d^3\mathbf{r}}\,, \end{aligned}$$

where $H(\mathbf{p},\mathbf{r}) = (\mathbf{p} - Q\mathbf{A})^2/2m + V(\mathbf{r})$ is the Hamiltonian of an electron, modified to include the effect of an applied magnetic field. The denominator (a partition function) is required to normalize the Boltzmann distribution, so that the sum of probabilities is unity. The second line has been obtained from the first by a change of variable, replacing the integral over momentum by an integral over velocity: for given $\mathbf{r}$ this is simply a shift in the origin, $\mathbf{p} = m\mathbf{v} + Q\mathbf{A}(\mathbf{r})$, and the only change in the element of integration is $d^3\mathbf{p} \to m^3 d^3\mathbf{v}$; the factors of m cancel from the numerator and denominator. But $\mathbf{B}_{\text{ind}}(\mathbf{r}', \mathbf{r}, \mathbf{v})$ is an odd function of $\mathbf{v}$, whereas the Boltzmann factor is an even function, and so the integral in the numerator vanishes: the average induced field is exactly zero. The same argument can be used for any number of electrons—there are no macroscopic magnetic effects in a classical system at thermal equilibrium [N. Bohr, 1911, and J. H. van Leeuwen, 1919].

The last, surprising, result tells us that it is futile to attempt to explain magnetism using classical physics: quantum effects must be introduced from the outset.

When a magnet is thrust into a wire loop, a current is induced in the loop, and the resulting magnetic field tends to oppose the field of the magnet. This might have been supposed to be a qualitative classical explanation of diamagnetism, but the resistance of the wire quickly brings the induced current to zero once the relative motion of magnet and loop has ceased, so any "diamagnetic" effect disappears as the system reaches thermal equilibrium. But perhaps it is unsatisfactory to invoke the macroscopic concept of resistance here: What of the *microscopic* motion of the electrons in the material? In a pure sample at low temperatures we expect an electron to make a few orbits in the applied magnetic field before being scattered into a new cyclotron orbit. An electron in a closed orbit certainly has a magnetic moment,[2] so why is there no resulting diamagnetism? Actually we must consider the motion of electrons throughout the *whole* solid: the ones at the surface as well as those in the bulk. Although there are relatively few electrons near the surface, their cyclotron orbits are disproportionately large and have a large magnetic moment in the opposite sense to that of orbits in the bulk—the result of repeated reflection at the surface; see Fig. 3.1. The resulting induced fields should tend to cancel.

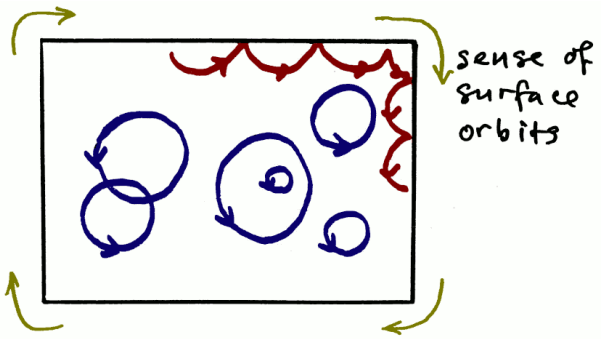

Figure 3.1: Motion of classical electrons in a magnetic field. The field is directed out of the page. Highlighted in red is part of the orbit of an electron near the surface. The electron orbits clockwise around the surface, owing to repeated reflection; its orbital magnetic moment is directed out of the page, in the opposite direction to the orbital moments due to electrons in the bulk.

The remarkable conclusion of the Bohr–van-Leeuwen theorem is that the magnetic fields due to bulk and surface electrons always cancel *exactly*, regardless of the precise shape of the sample.

Exercise 3.2:

> The motion of charged particles within the sun generates a large magnetic field which has persisted over thousands of millions of years. Assuming that the relevant equations of plasma physics are classical, why does this *not* violate the Bohr–van-Leeuwen theorem?

[2] Recall that the magnetic moment of a current loop is the product of the current and the (vector) area enclosed.

3.1.4 Quantum Hamiltonian of an electron in a magnetic field

There are two contributions to the interaction of an electron with a magnetic field. One of these, the part due to the electron's orbital motion, is the quantized version of (3.5),

$$\hat{H}_{\rm orb} = (\hat{\mathbf{p}} - Q\mathbf{A})^2/2m_e,$$

where $\hat{\mathbf{p}} = -i\hbar\nabla$ is the momentum operator, and $Q = -e$ is the (negative) electronic charge; for simplicity we have assumed an isotropic effective mass m_e for the electron, and this need not equal the free-electron mass m_0. The second, purely quantum, effect is the intrinsic magnetic moment associated with the electron's spin $\hat{\mathbf{s}}$, and this interacts directly with the applied field,

$$\hat{H}_{\rm spin} = -\hat{\mathbf{m}}_{\rm spin}\cdot\mathbf{B}, \quad \text{where} \quad \hat{\mathbf{m}}_{\rm spin} = -2\mu_B\hat{\mathbf{s}}/\hbar = -\mu_B\boldsymbol{\sigma}.$$

Here $\mu_B = e\hbar/2m_0$ is the Bohr magneton, numerically equal to $9.3\times10^{-24}\,\mathrm{J\,T^{-1}}$, and $\boldsymbol{\sigma} = (\sigma_x,\sigma_y,\sigma_z)$ are the Pauli spin matrices. At this point all we need to know about the spin matrices is that they have eigenvalues ±1, so that the energy of a spin in a uniform magnetic field B is $\mp\mu_B B$.

3.2 Magnetic quantities in thermodynamics

For later use we give some definitions of magnetic quantities, starting at the microscopic level and building up to the relations of macroscopic thermodynamics.

The orbital magnetic moment of a charge Q moving with velocity $\mathbf{v}$ can be defined as

$$\mathbf{m}_{\rm orb} = \tfrac{1}{2}Q\,\mathbf{r}\times\mathbf{v}\,. \tag{3.6}$$

This is consistent with Ampère's definition (footnote 2) because $\mathbf{r}\times\delta\mathbf{r}/2$ is the vector area of the triangle spanned by $\mathbf{r}$ and $\delta\mathbf{r} = \mathbf{v}\,\delta t$, so when averaged over a closed orbit (3.6) gives the required product of current and area.

Exercise 3.3:

> Verify this by forming the time average $\int \mathbf{m}_{\rm orb}\,dt/T$ for a closed orbit of period T. Note that Q/T is the mean current.

Now consider the change in the Hamiltonian $H(\mathbf{p},\mathbf{r})$ when the externally applied uniform magnetic field $\mathbf{B}$ changes by $\delta\mathbf{B}$. Only the kinetic part of H depends on $\mathbf{B}$, via the vector potential,

$$\delta H = \delta(\mathbf{p}-Q\mathbf{A})^2/2m = -Q\,\delta\mathbf{A}\cdot(\mathbf{p}-Q\mathbf{A})/m \equiv -Q\,\delta\mathbf{A}\cdot\mathbf{v}\,.$$

For a uniform field, $\mathbf{A}$ can be taken as $\frac{1}{2}\mathbf{B}\times\mathbf{r}$, so that

$$\delta H = -\tfrac{1}{2}Q\,\delta\mathbf{B}\times\mathbf{r}\cdot\mathbf{v} = -\tfrac{1}{2}Q\,\mathbf{r}\times\mathbf{v}\cdot\delta\mathbf{B} = -\mathbf{m}_{\rm orb}\cdot\delta\mathbf{B}\,,$$

where the cyclic property of the scalar triple product has been used.

When averaged over the motion of the charge, δH gives the energy change, and so must equal the "magnetic work" done on the charge. [Physically, the work is done by the *electric* field $\boldsymbol{\mathcal{E}} = -\partial\mathbf{A}/\partial t$, induced when the magnetic field is changing; as noted before, the magnetic field itself does no work.] The work done on a body is

$$\delta W_{\rm mag} = -\mathbf{M}\cdot\delta\mathbf{B} = -M\,\delta B\,,$$

where $\mathbf{M}$ is its total magnetic moment, i.e., the sum of the moments of all the individual charges; to simplify the notation in the final expression we have assumed that $\mathbf{M}$, $\mathbf{B}$ and $\delta\mathbf{B}$ are all in the same direction. For a quantum mechanical system at zero temperature the work done must equal the change in the ground-state energy $E_0(B)$, so we have

$$M = -\frac{\partial E_0}{\partial B}.$$

The magnetic susceptibility χ of a *linear* material (as most are at small fields) is defined by $\mathbf{M} = \chi\mathbf{H}$. However, for materials that are only *weakly* magnetic the field intensity $\mathbf{H}$ is very nearly equal to $\mathbf{B}/\mu_0$, so we can write

$$\chi \simeq \mu_0 \frac{M}{B} = \mu_0 \frac{\partial M}{\partial B} = -\mu_0 \frac{\partial^2 E_0}{\partial B^2}.$$

A paramagnetic substance has $\chi > 0$, a diamagnetic one has $\chi < 0$. If M denotes the *magnetization*, i.e., the magnetic dipole moment *per unit volume* ($\mathbf{M} = \mathbf{B}/\mu_0 - \mathbf{H}$), χ will be dimensionless, but it is also common practice to refer M and χ to different quantities of matter, e.g., to one mole or one kilogram, rather than to one cubic metre of material; for example, if $\mathbf{M}$ is referred to one mole, its units will be $\mathrm{A\,m^2\,mol^{-1}}$ and the units of χ will be $\mathrm{m^3\,mol^{-1}}$.

The variation of the energy with B gives a way of measuring the magnetic moment of a sample. In a field which varies slowly with position x,

$$[\text{force on sample}] = -\frac{\partial E_0}{\partial x} = -\frac{\partial E_0}{\partial B}\frac{\partial B}{\partial x} = M\frac{\partial B}{\partial x}.$$

Exercise 3.4:

Show that a paramagnetic substance tends to be drawn into regions of greater magnetic field, while a diamagnet is expelled.

The expression for the magnetic work is not restricted to zero temperature, so we can find how these relations are generalized to $T \neq 0$. From the fundamental theorem of thermodynamics, the change in the energy of a body in a magnetic field is

$$dE = đW_{\mathrm{mag}} + T\,dS = -M\,dB + T\,dS,$$

where E and S (the entropy) refer to the same amount of substance as M. Hence in a change at constant temperature $M\,dB = -dE + T\,dS = -d(E - TS)$, or

$$M = -\left(\frac{\partial F}{\partial B}\right)_T, \tag{3.7}$$

where $F \equiv E - TS$ is the Helmholtz free energy of the system. The earlier expressions for χ and for the force on a body are generalized simply by replacing E_0 by F.

3.3 Magnetism of a gas of free electrons

A general calculation of the magnetic properties of a gas of free electrons is difficult. To keep the treatment simple, we suppose that the gas is at high temperature or low density. In this limit, the Fermi–Dirac distribution function reduces to the Maxwell–Boltzmann form,

$$f(E_i) \simeq \exp[(\mu - E_i)/k_B T],$$

where μ is the chemical potential of the electrons.[3] The Maxwell-Boltzmann limit is appropriate to carriers in a lightly-doped semiconductor, but not (for example) to electrons in a metal.

Leaving aside the application to semiconductors, the reason it is helpful to consider this limit is that E_i, just like the Hamiltonian, can be separated into spin and orbital contributions, $E_i = E_{i,\mathrm{spin}} + E_{i,\mathrm{orb}}$. When the energy is a sum of terms, the Maxwell–Boltzmann distribution can be written as a product of factors

$$f(E_i) \propto e^{-(E_{i,\mathrm{spin}}+E_{i,\mathrm{orb}})/k_B T} = e^{-E_{i,\mathrm{spin}}/k_B T} \times e^{-E_{i,\mathrm{orb}}/k_B T};$$

i.e., the spin and orbital degrees of freedom are statistically independent, and so can be considered separately. This is what we do in the next two sections.

[3] However, for the purposes of *this* course, you can just regard the factor $\exp[\mu/k_B T]$ as a normalization factor, chosen so that the sum of the $f(E_i)$, taken over all states, equals the number of electrons in the system.

3.3.1 Pauli spin paramagnetism of an electron gas

Here we calculate the spin magnetic moment of an electron in a low-density electron gas, by using the Boltzmann distribution. The spin contribution to its energy is $\pm\mu_B B$, so the thermal average of the electron's magnetic moment is

$$M_{\text{spin}} = \frac{\mu_B e^{\mu_B B/k_B T} - \mu_B e^{-\mu_B B/k_B T}}{e^{\mu_B B/k_B T} + e^{-\mu_B B/k_B T}} = \mu_B \tanh[\mu_B B/k_B T].$$

For sufficiently large fields ($\mu_B B \gg k_B T$), the tanh function approaches the value 1; as might have been expected, the electron's magnetic moment aligns with the applied field. Because of the small size of μ_B we more usually have the opposite case $\mu_B B \ll k_B T$, so the tanh can be approximated by the first term of its Taylor series, $\tanh x \simeq x$, giving

$$M_{\text{spin}} \simeq \mu_B^2 B/k_B T = \left(\frac{e\hbar}{2m_0}\right)^2 \frac{B}{k_B T}. \tag{3.8}$$

The magnetization is in the same direction as the applied field, so the spins are paramagnetic. Note that the result is proportional to $1/m_0^2$, which explains why the contribution from the much more massive *nuclear* spins can usually be neglected.

3.3.2 Landau orbital diamagnetism of an electron gas

The classical cyclotron motion of electrons in a uniform magnetic field consists of free motion parallel to the field and circular motion of frequency $\omega_c = eB/m_e$ perpendicular to the field. If the field is in the z direction we might expect the energy levels to have the form

$$E_n(k_z) = \left(n + \tfrac{1}{2}\right)\hbar\omega_c + \hbar^2 k_z^2/2m_e,$$

where $n = 0, 1, 2, \ldots$, numbers the energy levels of the harmonic motion, called, in this context, the *Landau levels*. This guess for the energy levels can be confirmed by straightforward calculation [Problem 3.x, maybe]. A crucial result from the full quantum treatment is that each Landau level is highly degenerate: there are $G = eBA/h$ states corresponding to any given values of n and k_z.[4] Accordingly the *entropy* of an electron with given n and k_z is given by Boltzmann's formula

$$S_n = k_B \ln G = k_B \ln[eBA/h].$$

The orbital magnetic moment of an electron in the nth Landau level can be found from (3.7)

$$M_n = -\frac{\partial E_n}{\partial B} + T\frac{\partial S_n}{\partial B} = -\left(n + \tfrac{1}{2}\right)\frac{e\hbar}{m_e} + \frac{k_B T}{B}.$$

It is relatively simple to average this magnetization with respect to the Boltzmann distribution, which is valid at low density or high temperature: $\langle n\rangle$, the mean number of quanta in a harmonic oscillator, is given by the Bose distribution, and the remaining terms are independent of the Landau level,

$$M_{\text{orb}} = -\left(\langle n\rangle + \tfrac{1}{2}\right)\frac{e\hbar}{m_e} + \frac{k_B T}{B} = -\left\{\frac{1}{e^{\hbar\omega_c/k_B T} - 1} + \frac{1}{2}\right\}\frac{e\hbar}{m_e} + \frac{k_B T}{B}.$$

The expression in curly brackets is identical to $\frac{1}{2}\coth[\hbar\omega_c/2k_B T]$, which for small fields (small ω_c) can be expanded in the series $\coth x \simeq x^{-1} + \frac{1}{3}x$. The term in the expansion proportional to $1/B$ is cancelled exactly by the entropy term $k_B T/B$, and we are left with

$$M_{\text{orb}} \simeq -\frac{1}{3}\left(\frac{e\hbar}{2m_e}\right)^2 \frac{B}{k_B T}. \tag{3.9}$$

The magnetization is in the opposite direction to B, corresponding to diamagnetism.

[4] We might expect the "number of places" we can put a cyclotron orbit to be proportional to the area A. The field dependence is the result of an uncertainty principle for the position (x_0, y_0) of the orbit centre: $\Delta x_0\, \Delta y_0 \sim h/eB$.

3.3.3 Total magnetic response of the electron gas

If we compare the expressions (3.8) and (3.9), we notice that they are of the same form, but with different prefactors:

$$M_{\rm orb} = -\frac{1}{3}\left(\frac{m_0}{m_e}\right)^2 M_{\rm spin}. \tag{3.10}$$

It turns out that this relation between the orbital and spin magnetic moment for weak fields is a general one for free electrons with an isotropic effective mass, and does *not* depend on the assumption of low density used here;[5] however, the proportionality to $1/k_BT$ that appears in (3.8) and (3.9) is specific to the case of low density and high temperature.

The relation (3.10) implies that the orbital susceptibility $\chi_{\rm orb} = -\frac{1}{3}(m_0/m_e)^2\chi_{\rm spin}$, so that the total susceptibility can be written as

$$\chi = \chi_{\rm spin} + \chi_{\rm orb} = \left[1 - \tfrac{1}{3}(m_0/m_e)^2\right]\chi_{\rm spin}. \tag{3.11}$$

In a simple metal such as sodium $m_e \simeq m_0$, so that the total susceptibility is positive, and equal to $\frac{2}{3}\chi_{\rm spin}$. This agrees, very roughly, with the observed paramagnetism of simple metals. On the other hand, in doped semiconductors and semi-metals such as bismuth the electron effective mass is small, typically in the range $0.01\,m_0$ to $0.1\,m_0$, so that experimentally the orbital diamagnetism completely overwhelms the paramagnetic contribution from the spins.

3.4 Magnetism of ions

In Section 3.1.4 we have written down the orbital and spin contributions to the energy of an electron in a magnetic field. The contribution to the energy of an ion can be obtained simply by summing over all its electrons, so that the spin term is

$$\hat{H}_{\rm spin} = \sum_i 2\mu_B \hat{\mathbf{s}}_i \cdot \mathbf{B}/\hbar \equiv 2\mu_B \hat{\mathbf{S}} \cdot \mathbf{B}/\hbar,$$

where $\hat{\mathbf{S}}$ is the total electronic spin of the ion. In a uniform magnetic field the vector potential can be taken to be $\mathbf{A} = \frac{1}{2}\mathbf{B} \times \mathbf{r}$ and the $\mathbf{B}$-dependent part of $\hat{H}_{\rm orb}$ is

$$\begin{aligned}
\Delta\hat{H}_{\rm orb} &= \sum_i \left\{ \frac{(\hat{\mathbf{p}}_i - Q\mathbf{A})^2}{2m_0} - \frac{\hat{\mathbf{p}}_i^2}{2m_0} \right\} \\
&= -\frac{Q}{m_0}\sum_i \mathbf{A}(\mathbf{r}_i)\cdot\hat{\mathbf{p}}_i + \frac{Q^2}{2m_0}\sum_i \mathbf{A}(\mathbf{r}_i)^2 \\
&= \frac{e}{m_0}\sum_i \tfrac{1}{2}\mathbf{B}\times\mathbf{r}_i\cdot\hat{\mathbf{p}}_i + \frac{e^2}{2m_0}\sum_i \frac{B^2(x_i^2+y_i^2)}{4} \\
&= \frac{e}{2m_0}\sum_i (\mathbf{r}_i\times\hat{\mathbf{p}}_i)\cdot\mathbf{B} + \frac{e^2B^2}{8m_0}\sum_i \left(x_i^2+y_i^2\right),
\end{aligned}$$

where, in the last two lines, Q has been replaced by the electronic charge $-e$ and it has been assumed that $\mathbf{B}$ is in the z direction, so that $\mathbf{A} = \mathbf{B}\times\mathbf{r}/2 = (-y, x, 0)\,B/2$. Using also the fact that $\sum \mathbf{r}_i \times \hat{\mathbf{p}}_i = \hat{\mathbf{L}}$, the total orbital angular momentum operator for the electrons, the B dependence of the energy of the ion becomes

$$\begin{aligned}
\Delta\hat{H}_{\rm mag} &= \hat{H}_{\rm spin} + \Delta\hat{H}_{\rm orb} \\
&= \mu_B(\hat{\mathbf{L}} + 2\hat{\mathbf{S}})\cdot\mathbf{B}/\hbar + \frac{e^2B^2}{8m_0}\sum_i \left(x_i^2+y_i^2\right).
\end{aligned} \tag{3.12}$$

[5] The proof is not easy, but is given in full by Peierls in *Quantum Theory of Solids* and by Landau and Lifshitz in *Statistical Physics*, Part 1.

The first term has the form $-\hat{\boldsymbol{\mu}} \cdot \mathbf{B}$, so that the operator $\hat{\boldsymbol{\mu}} = -\mu_B(\hat{\mathbf{L}} + 2\hat{\mathbf{S}})/\hbar$ might be interpreted as the "intrinsic" magnetic moment of the ion, which it would have even in the absence of the field. [Note that the orbital contribution to $\hat{\boldsymbol{\mu}}$ is not quite the same as that previously defined in (3.6), because the momentum $\hat{\mathbf{p}}_i$ equals $m_0\mathbf{v}_i$ only in the limit of vanishing field.]

3.4.1 Hund's rules

The state of an ion in free space can be classified by its total orbital angular momentum and spin, L and S. In most cases the state of lowest energy of an ion with a partially-filled shell of electrons can be determined using *Hund's rules*, which were first discovered empirically by study of atomic spectra. We give them here for reference:

(H1) electrons are arranged so as to have the greatest total spin S consistent with the Pauli exclusion principle: no two electrons should have the same quantum numbers

(H2) for this spin configuration, the electrons are arranged among states of different l_z so that the total L is also maximized

(H3) the total angular momentum J is given by

$$J = \begin{cases} |L-S| & \text{if the shell is no more than half full} \\ L+S & \text{if the shell is at least half full.} \end{cases}$$

The last of these rules arises from the relativistic spin–orbit interaction, which will not normally be considered in this course. The other two rules minimize the total electrostatic potential energy of the electrons, though this is by no means obvious: we shall discuss rule H1 later in the course.

Given the values of J, J_z, L and S the ion's intrinsic magnetic moment can be obtained from[6]

$$\mu_z = -g(JLS)\mu_B J_z, \quad \text{where} \quad g(JLS) = \frac{3}{2} + \frac{S(S+1) - L(L+1)}{2J(J+1)}$$

is the Landé g-factor.

To give a specific example, a Cr^{3+} ion has the electronic configuration [Ar]$3d^3$, so there are three electrons in states with $l = 2$. To maximize S we suppose that these electrons all have $s_z = +\frac{1}{2}$, which (by the exclusion principle) requires them to be in states with different values of l_z. To maximize L we occupy the states of greatest l_z, namely $l_z = 2, 1, 0$. Hence $S = \frac{3}{2}$ and $L = 3$ [which is denoted by the spectroscopic symbol F]. The d-shell will accommodate ten electrons, so it is less than half full. Rule H3 gives $J = |L - S| = \frac{3}{2}$. Spectroscopists would denote this ground-state "term" by the symbol $^{2S+1}[L]_J = {}^4F_{3/2}$.

With shells that are more than half full it is sometimes convenient, though not essential, to consider the states of the smaller number of "holes" in the shell. For example, the rare-earth ion Tm^{3+} has configuration [Xe]$4f^{12}$. Since the f-shell can accommodate 14 electrons, this is equivalent to having two holes. Putting these in states with $s_z = +\frac{1}{2}$ and $l_z = 3$ and 2 (the largest possible values) gives $S = 1$ and $L = 5$. As the shell is more than half full of electrons, these are combined to give $J = L + S = 6$. The spectroscopic symbol would be ${}^3H_6$.

Exercise 3.5:

Use Hund's rules to show that the ground state of Fe^{2+} is ${}^5D_4$. [The electronic configuration of the ion is [Ar]$3d^6$.]

[6] Notationally, J_z without a hat should be an eigenvalue of $\hat{J}_z$, so strictly speaking it takes the values $-J\hbar$ to $J\hbar$, in steps of $\hbar$. However, in this chapter we make heavy use of angular momentum, and it is tiresome to write in every factor of $\hbar$. So, just as the total angular momentum quantum number J is taken to be a pure number, we do the same with J_z [and similarly later for s_z and l_z], inserting or (more usually) deleting factors of $\hbar$ to keep the units right. In other courses, such as PHYS30101, the quantum number M_J is identical to our J_z, but the letter M is already heavily overworked.

3.4.2 Diamagnetism of closed-shell systems

When a shell is completely full of electrons we have a special case of Hund's rules, $J = L = S = 0$. It arises in rare-gas atoms, and in solids (such as NaCl) in which all the ions have a rare-gas configuration of electrons. The ion has no *intrinsic* magnetic moment (the g-factor is zero), so that all the effect of the magnetic field on the energy of the ion must come from the second term in (3.12). Using first-order perturbation theory the energy shift is

$$\Delta E_{\text{mag}} = \frac{e^2 B^2}{8m_0} \sum_i \langle x_i^2 + y_i^2 \rangle$$

per ion. Since the energy is raised in the presence of B, flux tends to be expelled from the material, so that the effect is *diamagnetic*, but in practice rather small: $\chi \simeq -10^{-5}$ [dimensionless] in rare gas solids. An expression for the magnetic susceptibility is given in Section 3.2: the value of the susceptibility *per ion* is

$$\chi = -\mu_0 \frac{\partial^2 \Delta E_{\text{mag}}}{\partial B^2} = -\frac{\mu_0 e^2}{4m_0} \sum_i \langle x_i^2 + y_i^2 \rangle = -\frac{\mu_0 e^2}{6m_0} \sum_i \langle r_i^2 \rangle,$$

where $\frac{2}{3}\langle r_i^2 \rangle$ has been written for $\langle x_i^2 + y_i^2 \rangle$, on the assumption that the ion is spherically symmetrical, or occupies a site of cubic symmetry in a solid.

3.4.3 Paramagnetism of ions with partially filled shells

Ions with partially filled shells of electrons have a net intrinsic magnetic moment which tends to align with an applied magnetic field. The resulting paramagnetism is analogous to the Pauli spin paramagnetism of the low-density electron gas. For ions with $L = 0$ and $J = S = \frac{1}{2}$ the g-factor is 2 (as for a free electron) so that as in the Pauli case analysed in Section 3.3.1.

$$\langle \mu_z \rangle = \mu_B \tanh[\mu_B B / k_B T] \to \mu_B^2 B / k_B T$$

for small fields or high temperature. The high-temperature susceptibility is

$$\chi \simeq \mu_0 \mu_B^2 / k_B T;$$

the dependence on $1/T$ is known as *Curie's law*.

The case of general J, L and S is very similar. Although we don't give all the details here, the exact expression for $\langle \mu_z \rangle$ can be worked out and is qualitatively very similar to the function $\tanh[\mu_B B / k_B T]$ that applies in the case $J = \frac{1}{2}$. The main differences appear in the low- and high-temperature limits,

$$\langle \mu_z \rangle \to \begin{cases} gJ\mu_B & \text{for} \quad k_B T \ll \mu_B B \\ \\ \frac{1}{3} g^2 \mu_B^2 J(J+1) B / k_B T & \text{for} \quad k_B T \gg \mu_B B \,. \end{cases} \tag{3.13}$$

The low-temperature [or high-field] limit is intuitively obvious: the magnetic moment is completely aligned with the field in this case. The high-temperature [or low-field] limit is Curie's law again, but with a different prefactor.

Exercise 3.6:

Check that these low- and high-temperature limits agree with the earlier results when $L = 0$, $J = S = \frac{1}{2}$.

The high-temperature case can be worked out relatively easily, though the details are not examinable. We start from the allowed energies of the ion, which are given by the formula for the Zeeman splitting, $\Delta E_{\rm mag} = -\mu_z B = g(JLS)\mu_B J_z B$, where $J_z = -J, -J+1, \ldots, J-1, J$; see footnote 6. At high temperature the average of μ_z with respect to the Boltzmann distribution gives

$$\begin{aligned}\langle \mu_z \rangle &= \frac{\sum(-g\mu_B J_z)\exp[-g\mu_B J_z B/k_B T]}{\sum \exp[-g\mu_B J_z B/k_B T]} \\ &\simeq \frac{\sum(-g\mu_B J_z)(1 - g\mu_B J_z B/k_B T + \ldots)}{2J+1} \\ &= \frac{g^2\mu_B^2}{3}\frac{J(J+1)}{k_B T}B,\end{aligned}$$

The key steps were to expand the exponential function in powers of the small quantity $g\mu_B J_z B/k_B T$ and use the identities

$$\sum_{J_z=-J}^{J} J_z = 0 \quad \text{and} \quad \sum_{J_z=-J}^{J} J_z^2 = \frac{J(J+1)(2J+1)}{3}.$$

Comparison with experiment: "quenching" of the orbital angular momentum

In an experiment, the quantity $p = g\sqrt{J(J+1)}$ (sometimes called the effective number of Bohr magnetons) can be extracted from the measured Curie's-law magnetization at high temperature [in practice, near *room* temperature], which follows the second of Eqs. (3.13). For solids containing ions of rare-earth metals there is typically good agreement between experiment and the theoretical value of p; see Table 3.1. There are, however, large discrepancies for Sm^{3+} and Eu^{3+}. In each of these cases, there is a J multiplet that is close in energy to the ground state of the ion, and this invalidates the simple analysis that leads to (3.13). [First-order perturbation theory no longer gives accurate results for the energy levels in an applied field, and one also cannot neglect thermal excitation into states of higher-lying J multiplets.]

Table 3.1: Theoretical and experimental values of p for rare earth ions.

Ion	configuration	term	J, L, S	$p_{\rm th} = g\sqrt{J(J+1)}$	$p_{\rm expt}$
Ce^{3+}	$4f^1$	$^2F_{5/2}$	$\frac{5}{2}, 3, \frac{1}{2}$	2.54	2.4
Pr^{3+}	$4f^2$	$^3H_4$	$4, 5, 1$	3.58	3.5
Nd^{3+}	$4f^3$	$^4I_{9/2}$	$\frac{9}{2}, 6, \frac{3}{2}$	3.62	3.5
Sm^{3+}	$4f^5$	$^6H_{5/2}$	$\frac{5}{2}, 5, \frac{5}{2}$	0.85	1.5
Eu^{3+}	$4f^6$	$^7F_0$	$0, 3, 3$	0.00	3.4
Gd^{3+}	$4f^7$	$^8S_{7/2}$	$\frac{7}{2}, 0, \frac{7}{2}$	7.94	8.0
Tb^{3+}	$4f^8$	$^7F_6$	$6, 3, 3$	9.72	9.5
Dy^{3+}	$4f^9$	$^6H_{15/2}$	$\frac{15}{2}, 5, \frac{5}{2}$	10.65	10.6
Ho^{3+}	$4f^{10}$	$^5I_8$	$8, 6, 2$	10.61	10.4
Er^{3+}	$4f^{11}$	$^4I_{15/2}$	$\frac{15}{2}, 6, \frac{3}{2}$	9.58	9.5
Tm^{3+}	$4f^{12}$	$^3H_6$	$6, 5, 1$	7.56	7.3
Yb^{3+}	$4f^{13}$	$^2F_{7/2}$	$\frac{7}{2}, 3, \frac{1}{2}$	4.54	4.5

Agreement can also be obtained for salts of transition metals if L is taken to be zero in the calculation of J and $g(JLS)$; see Table 3.2. This empirical result (which contradicts Hund's rules) is known as *quenching* of the orbital angular momentum. It arises because, in a crystal, an ion has other ions nearby, so the

Table 3.2: Theoretical and experimental values of p for first-row transition-metal ions. The calculated values obtained by setting $L = 0$, i.e., $p_{\text{th}} = 2\sqrt{S(S+1)}$, are in better agreement with experiment than those obtained from $p_{\text{th}} = g\sqrt{J(J+1)}$, with J determined by Hund's third rule.

Ion	configuration	term	J, L, S	$p_{\text{th}} = g\sqrt{J(J+1)}$	$2\sqrt{S(S+1)}$	p_{expt}
V^{4+}	$3d^1$	$^2D_{3/2}$	$\frac{3}{2}, 2, \frac{1}{2}$	1.55	1.73	1.8
V^{3+}	$3d^2$	$^3F_2$	$2, 3, 1$	1.63	2.83	2.8
V^{2+}	$3d^3$	$^4F_{3/2}$	$\frac{3}{2}, 3, \frac{3}{2}$	0.77	3.87	3.8
Mn^{4+}	$3d^3$	$^4F_{3/2}$	$\frac{3}{2}, 3, \frac{3}{2}$	0.77	3.87	4.0
Cr^{2+}	$3d^4$	$^5D_0$	$0, 2, 2$	0.00	4.90	4.8
Mn^{3+}	$3d^4$	$^5D_0$	$0, 2, 2$	0.00	4.90	5.0
Mn^{2+}, Fe^{3+}	$3d^5$	$^6S_{5/2}$	$\frac{5}{2}, 0, \frac{5}{2}$	5.92	5.92	5.9
Fe^{2+}	$3d^6$	$^5D_4$	$4, 2, 2$	6.71	4.90	5.4
Co^{2+}	$3d^7$	$^4F_{9/2}$	$\frac{9}{2}, 3, \frac{3}{2}$	6.63	3.87	4.8
Ni^{2+}	$3d^8$	$^3F_4$	$4, 3, 1$	5.59	2.83	3.2
Cu^{2+}	$3d^9$	$^2D_{5/2}$	$\frac{5}{2}, 2, \frac{1}{2}$	3.55	1.73	1.9

potential is not spherically symmetric; and in the absence of spherical symmetry the components of $\hat{\mathbf{L}}$ will not, in general, be conserved. The effect is to break the degeneracy of the L multiplet. [For example, at an atomic site with cubic symmetry the "crystal field" splits the d-orbitals into two distinct sets, $\{d_{xy}, d_{yz}, d_{zx}\}$ and $\{d_{x^2-y^2}, d_{3z^2-r^2}\}$, which have different energies.] In the first-row transition elements, the energy splitting due to the crystal field is greater than the spin–orbit interaction that couples L and S to form a definite J. The result is that $\langle\hat{\mathbf{L}}\rangle = \mathbf{0}$ for an ion at a site of sufficiently low symmetry, leading to $\langle\hat{\boldsymbol{\mu}}\rangle = -\mu_B\langle\hat{\mathbf{L}} + 2\hat{\mathbf{S}}\rangle/\hbar = -2\mu_B\langle\hat{\mathbf{S}}\rangle/\hbar$; the effective g-factor is therefore simply 2, and J is replaced by S.

Crystal field effects are not so important in the rare-earth materials. The f-orbitals extend less far from the ions, and so are less influenced by their environment. Additionally, the spin–orbit interaction is greater in these heavier ions, so the net effect is for L and S to remain coupled to form a single conserved angular momentum J. As a result, the $4f$-shell has properties very similar to those of a free ion, as confirmed by the results in Table 3.1.

3.5 Ordered magnetic states

So far we have discussed only dia- and paramagnetic effects due to independent electrons, atoms or ions, but the more interesting phenomena of ferro- and antiferromagnetism are examples of *cooperative* behaviour. Ultimately they are the result of short-ranged interactions between pairs of electrons in the solid, but large numbers of electrons must act in concert to produce an ordered magnetic state.

To simplify the picture we assume that the elementary magnetic moments are localized on particular atoms or ions in the solid. This is the case for many magnetic insulators and rare-earth materials. Owing to the forces between spins (see later) the directions of the moments are correlated, so that a pair of adjacent moments tend to point in the same (or opposite) directions. This can be true at any temperature. But if the temperature is low enough, so that the thermally-induced disorder is reduced, *all* of the moments in the solid may tend to point—on the average—in some definite direction. We speak of a *spontaneous magnetization* of the solid.

The microscopic arrangement of magnetic moments in a solid can be investigated experimentally via neutron scattering, because neutrons themselves have a magnetic moment which interacts with the moments in the solid.

If adjacent spins tend to be parallel, the solid acquires a macroscopic magnetization $\mathbf{M}$, like a bar magnet, which is the case of *ferromagnetism*. More commonly adjacent moments are directed opposite

to one another, and the configuration is said to be *antiferromagnetic*: the sum of the moments in a unit cell averages to zero. In both cases it is found that the magnetic order is greatest at low temperatures and decreases continuously to zero at a critical temperature T_c, sometimes called the Curie temperature (ferromagnets) or Néel temperature (antiferromagnets). Above this temperature there is no net magnetization, even though the directions of adjacent moments are still correlated: the magnetization characterizes the correlations between widely-separated moments. In some cases the critical temperatures are surprisingly high, so that the forces between spins must be fairly large.

Table 3.3: Curie temperatures of some transition metals and rare-earths

	T_c [K]
Fe	1043
Co	1388
Gd	293
Dy	85

Other, more exotic kinds of magnetism are possible in which different ions in the unit cell have different sizes of magnetic moment. In other cases (such as iron) the magnetic moment is distributed in space in the gas of conduction electrons, and sometimes (as in chromium) this distributed moment may take the form of a wave whose period is unrelated to that of the crystal lattice. These and many other possibilities are still the subject of intensive research.

In these notes we look briefly at the forces tending to align the elementary magnetic moments in a solid, studying in most detail the states of just *two* electrons. We then examine the simplest theory of the critical point, which shows how magnetic moments with short-range forces between them can cooperate at low temperatures to give an ordered magnetic state. Corrections to this picture include *magnons*, which are wave-like disturbances of the magnetization.

3.5.1 Dipolar interaction between spins

Electrons have a magnetic moment $\hat{\mathbf{m}} = -2\mu_B\hat{\mathbf{s}}$ associated with their intrinsic spin angular momentum $s = \frac{1}{2}$. Our first guess would be that magnetism was due to dipolar interaction of the magnetic moments: the dipoles have an associated magnetic field, so we might expect each moment to align in the magnetic field of its neighbours. However, an estimate of this energy will show that it is much too small to explain transition temperatures on the order of 10^3 K.

The interaction energy of two magnetic moments $\mathbf{m}_1$ and $\mathbf{m}_2$ separated by distance $\mathbf{r}$ is[7]

$$U = \frac{\mu_0}{4\pi r^3}[\mathbf{m}_1 \cdot \mathbf{m}_2 - 3(\mathbf{m}_1 \cdot \hat{\mathbf{r}})(\mathbf{m}_2 \cdot \hat{\mathbf{r}})].$$

Inserting orders of magnitude for two dipoles separated by about 1 Å we find

$$U \sim \frac{\mu_0 \mu_B^2}{4\pi r^3} \sim \frac{10^{-7} \times (10^{-23})^2}{(10^{-10})^3} \sim 10^{-23}\,\text{J}.$$

The spins should be disordered if the thermal energy $k_B T$ is comparable with the energy U required to reverse a spin. This would suggest a transition temperature of around 1 K, which in the cases of iron and cobalt is too small by three orders of magnitude.

[7] You could prove this using the expression $\psi = \mu_0 \mathbf{m}_1 \cdot \hat{\mathbf{r}}/4\pi r^2$ for the magnetic *scalar* potential for the field around $\mathbf{m}_1$. The energy of the second dipole in this field is $\mathbf{m}_2 \cdot \nabla\psi$.

3.5.2 Exchange interaction

In fact, the mechanism for spin alignment in magnets comes about indirectly as a result of *electrostatic* interactions and the antisymmetry of electron wave functions $\Phi(\mathbf{r}_1\sigma_1, \mathbf{r}_2\sigma_2, \ldots) = -\Phi(\mathbf{r}_2\sigma_2, \mathbf{r}_1\sigma_1, \ldots)$, which is a fundamental property of all fermion wave functions. This is the general rule that leads to Pauli's exclusion principle. For if two electrons are known to be in states α and β, the only antisymmetric combination of these is

$$\Phi(1,2) = \alpha(1)\beta(2) - \alpha(2)\beta(1),$$

where the symbols (1) and (2) are just a shorthand for $(\mathbf{r}_1\sigma_1)$ and $(\mathbf{r}_2\sigma_2)$, or whatever variables the states are supposed to depend on. If the two states α and β are the same $\Phi(1,2)$ vanishes, so two electrons cannot occupy the same state.

To the extent that spin–orbit coupling can be neglected, the interaction between two electrons is simply the Coulomb interaction $e^2/4\pi\epsilon_0 r_{12}$, and can be included in the Hamiltonian $\hat{H}(1,2)$ along with the kinetic and single-particle potential energies. Because the energy is symmetrical and does not depend on the spin variables, solutions of the purely *spatial* Schrödinger equation,

$$\hat{H}\psi(\mathbf{r}_1, \mathbf{r}_2) = E\psi(\mathbf{r}_1, \mathbf{r}_2),$$

which determine the energy levels of the two electrons, are themselves either symmetric or antisymmetric: $\psi(1,2) = \pm\psi(2,1)$. [This is easy to see for a non-degenerate level as $\psi(1,2)$ and $\psi(2,1)$ each solve the same Schrödinger equation. If the level is non-degenerate, the two functions must be identical apart from a factor, $\psi(2,1) = C\psi(1,2)$. Repeating the argument gives $\psi(1,2) = C^2\psi(1,2)$, so that $C = \pm 1$.]

This is consistent with the Pauli principle because the *complete* wave function

$$\Phi(1,2) = \psi(\mathbf{r}_1, \mathbf{r}_2)\,\chi(\sigma_1, \sigma_2) \tag{3.14}$$

is the product of space- and spin-dependent functions, so if ψ is symmetric, χ will be antisymmetric, and vice versa.

If we are interested in only a few low-energy states of a system that can be specified uniquely by their spin configuration it is sometimes convenient to reverse the argument. The spin state determines the symmetry of the spatial state, and this in turn determines the energy, via Schrödinger's equation. This dependence of the energy on the spin state is known as the *exchange interaction*.

We can see how the exchange interaction leads to Hund's first rule for electrons in spatial states a and b, assumed to be degenerate in energy when electron interactions are neglected. (The functions a and b might be two atomic p orbitals, as in an atom of carbon.) There is only one antisymmetric spin wave function for two electrons,

$$\chi_s = (\uparrow_1\downarrow_2 - \downarrow_1\uparrow_2)/\sqrt{2}\,.$$

This "singlet" state has total $S_z = 0$, and so must correspond to total $S = 0$. The corresponding symmetric spatial wave function is

$$\psi_s = (a(1)b(2) + b(1)a(2))/\sqrt{2}\,. \tag{3.15}$$

On the other hand there is a "triplet" of states with symmetrical spin wave functions

$$\chi_t = \begin{cases} \uparrow_1\uparrow_2 \\ (\uparrow_1\downarrow_2 + \downarrow_1\uparrow_2)/\sqrt{2} \\ \downarrow_1\downarrow_2\,. \end{cases} \tag{3.16}$$

They have, respectively, $S_z = 1$, 0 and -1, and so correspond to $S = 1$. For all three, the antisymmetric spatial wave function is

$$\psi_t = (a(1)b(2) - b(1)a(2))/\sqrt{2}\,, \tag{3.17}$$

which vanishes when $\mathbf{r}_1 = \mathbf{r}_2$. Thus, the probability of finding the two electrons close together is smaller in the triplet state than in the singlet; the expectation value of the electron–electron interaction is reduced,

so that the triplet has lower energy than the singlet. This is the origin of Hund's first rule for the coupling of spins, as it prefers $S = 1$ over $S = 0$.

Suppose that $E_s > E_t$ are the energies of the singlet and triplet states. In this case we can (in a purely ad hoc way) write the energy as a function of the total spin S,

$$E_{\text{eff}}(S) = E_s - \tfrac{1}{2}(E_s - E_t)S(S+1) = \begin{cases} E_s & \text{for} \quad S = 0 \\ E_t & \text{for} \quad S = 1. \end{cases} \tag{3.18}$$

The function $E(S)$ has been written in terms of $S(S+1)$ for a reason: $S(S+1)\hbar^2$ are the eigenvalues of the square of the operator for the total spin of the two electrons, so that (3.18) shows that E_s and E_t are the eigenvalues of an effective spin-Hamiltonian,

$$\hat{H}_{\text{eff}} = E_s - \tfrac{1}{2}(E_s - E_t)\hat{\mathbf{S}}^2/\hbar^2 \,.$$

The expression for $\hat{H}_{\text{eff}}$ can be written explicitly in terms of the operators for the individual spins,

$$\begin{aligned} \hat{\mathbf{S}}^2 &= (\hat{\mathbf{s}}_1 + \hat{\mathbf{s}}_2)^2 \\ &= \hat{\mathbf{s}}_1^2 + \hat{\mathbf{s}}_2^2 + 2\hat{\mathbf{s}}_1 \cdot \hat{\mathbf{s}}_2 \\ &= \tfrac{3}{2}\hbar^2 + 2\hat{\mathbf{s}}_1 \cdot \hat{\mathbf{s}}_2 \end{aligned}$$

where to get the last result we have used $\hat{\mathbf{s}}_1^2 = \hat{\mathbf{s}}_2^2 = \frac{3}{4}\hbar^2$. Putting these results together we find that the effective Hamiltonian can be written in the form

$$\hat{H}_{\text{eff}}(\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2) = \text{const.} + J\hat{\mathbf{s}}_1 \cdot \hat{\mathbf{s}}_2/\hbar^2, \tag{3.19}$$

where the coupling constant, $J = (E_t - E_s) < 0$, is the singlet–triplet energy difference [Heisenberg 1926, Dirac 1929]. Within an atom, J is on the order of one electron volt, which is much greater than the dipolar interaction energy.

Exercise 3.7:

Verify the result in (3.19) and find the value of the additive constant in terms of E_s and E_t.

Given the form of the singlet and triplet spatial wave functions (3.15) and (3.17) we can use first-order perturbation theory to estimate the energy-splitting of the singlet and triplet states. The results are [in atomic units, $e^2/4\pi\epsilon_0 = 1$]

$$\begin{aligned} \left\langle \frac{1}{r_{12}} \right\rangle_s &= \iint \frac{|a(1)|^2|b(2)|^2}{r_{12}}\, d^3\mathbf{r}_1\, d^3\mathbf{r}_2 + \iint \frac{a^*(1)b^*(2)a(2)b(1)}{r_{12}}\, d^3\mathbf{r}_1\, d^3\mathbf{r}_2 \\ &\equiv U_D[a,b] + U_X[a,b] \end{aligned}$$

$$\left\langle \frac{1}{r_{12}} \right\rangle_t = U_D[a,b] - U_X[a,b] \,.$$

U_D and U_X are known, respectively, as the direct and exchange energies. The direct energy can be interpreted as the electrostatic interaction energy of two charge distributions $-e|a(\mathbf{r})|^2$ and $-e|b(\mathbf{r})|^2$. The difference in sign of the exchange energy in the singlet and triplet states can be traced directly to the symmetry and antisymmetry of the spatial wave functions.

Exercise 3.8:

Optional. Show that U_X is always positive. [Regard U_X as the energy of interaction of two *complex conjugate* charge distributions $\rho(1) = a(1)b^*(1)$ and $\rho^*(2) = a^*(2)b(2)$ and separate ρ, ρ^* and U_X into real and imaginary parts. The imaginary part of U_X vanishes, and the real part can be recognised as the sum of the (positive) electrostatic energies of $\text{Re}\,\rho$ and $\text{Im}\,\rho$.]

3.5.3 Exchange interaction between ions

So far we have seen that the exchange interaction makes it favourable for electron spins to align within an atom, and so leads to Hund's rule (H1). Interactions of the same general kind are responsible for the interaction between the spins on different ions in a magnetic solid.

The expression for the exchange energy U_X depends on the product $a^*(\mathbf{r})b(\mathbf{r})$, and so is small if the orbitals $a(\mathbf{r})$ and $b(\mathbf{r})$ overlap only slightly, as is the case for electrons belonging to different ions. Compared with the interactions between electron spins within an ion, the interactions between the total spins of different ions should correspond to smaller values of exchange coupling constant J. A critical temperature of 10^3 K in a magnetic solid would suggest $|J| \sim 0.1$ eV, which is indeed substantially less than in the intraatomic $|J| \sim 1$ eV responsible for Hund's first rule.

3.5.4 Why aren't all magnets FM?

The picture we have just described cannot be complete, otherwise all magnets would be ferromagnetic. There are several complications to the picture.

Indirect exchange interaction

In the case of rare earth elements (the lanthanides) the picture of localized spins is correct (they are located in the f-orbitals), but the orbitals are too tightly bound to the atoms to allow a simple exchange interaction. The interaction in this case occurs indirectly, via the conduction electrons of the solid.

Roughly speaking, there is a simple exchange interaction between the conduction electrons and the f electrons in the rare-earth ion. Up- and down-spin conduction electrons are attracted by different amounts to the ion. The resulting shift in phase between the wave functions of up and down electrons causes one or the other to predominate at any given point in space. It can be shown (using second-order perturbation theory) that the difference $n_\uparrow - n_\downarrow$ oscillates as a function of position, where the oscillations have wave vector $2k_F$; k_F is the Fermi wave vector of the gas of conduction electrons. As a result, the energy of a second rare-earth ion immersed in the conduction-electron sea has an oscillatory dependence on its distance R from the first ion,

$$E(\mathbf{S}_1, \mathbf{S}_2) = J\mathbf{S}_1 \cdot \mathbf{S}_2 \quad \text{with} \quad J \sim \text{const.} \times \frac{\cos(2k_F R)}{R^3}.$$

This is the *Rudermann–Kittel interaction*, or *indirect exchange*.

Whether the solid is ferro- or antiferromagnetic depends on the sign of the coupling constant J, which in turn depends on the separation of the magnetic ions, and on the conduction electron density via k_F. Both kinds of magnetism (and more) have been observed among the rare-earth metals and their alloys.

Superexchange interaction

In the electrically-insulating oxides of transition metals, such as manganese oxide, a slightly different mechanism is reponsible for the alignment of the spins of the transition-metal ions.

Manganese oxide is antiferromagnetic at temperatures below 116 K. In the crystal, Mn^{2+} and O^{2-} ions occupy alternate sites of a simple-cubic lattice. [The lattice type is therefore cubic-F, with one Mn and one O ion per lattice point.] The manganese ions have a magnetic moment, due to the electrons in the half-filled $3d$ shell, but the ions are too far apart for the $3d$ wave functions to have any significant overlap, so the normal exchange interaction would be too small to explain the ordering of the spins; moreover, it would be of the wrong sign to give an antiferromagnetic interaction. Instead, the interaction arises via the exchange interaction with the electrons on the magnetically-inert oxygen ions. An oxygen p-electron can make virtual transitions into states of one of its neighbouring Mn ions, leaving an electron behind that can interact (via the exchange interaction) with a $3d$ electron of another Mn ion.[8]

[8] An appeal to "virtual transitions" is usually a metaphor for an effect that can be understood by using second-order perturbation theory. The theory was worked out in detail by P. W. Anderson [1950], though it was not the work that won him a Nobel prize.

Indirect interactions of this kind, via the electrons of non-magnetic ions, are usually called *superexchange*; they can be either ferromagnetic or (more commonly) antiferromagnetic, as in MnO. The three possibilities for exchange discussed above are illustrated in Figure.

3.5.5 The Heisenberg Hamiltonian

Regardless of the details of the exchange mechanism, the low-energy states of a magnetic material with localized spins are often described by the *Heisenberg Hamiltonian*,

$$\hat{H}_{\text{Heis}} = \sum_{\langle i,j \rangle} J_{ij}\, \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j / \hbar^2 , \tag{3.20}$$

where $\hat{\mathbf{S}}_i$ denotes the operator for the *total* spin of the ith ion (which need not be $\frac{1}{2}$) and the notation $\langle i, j \rangle$ means that the sum is taken over distinct *pairs* of ions i and j.

Although the Heisenberg Hamiltonian has often been used as the starting point for theoretical discussions of magnetism, it should not be regarded as fundamental. It can be derived by perturbation theory in some cases [footnote 8], but the approximations involved are rarely under "good control": there is not always a suitable small parameter to expand in. Equation (3.20) is perhaps better regarded as the simplest Hamiltonian that respects the fact that the ionic spins are quantum-mechanical and can interact via one of the exchange mechanisms. We shall use it here without further discussion.

3.6 Ferromagnetic groundstate and excitations

Besides being used to characterize the magnetic order in a solid, neutron scattering can also be used to study magnetic excitations. The neutrons are coupled to the magnetization density via their own magnetic moments, and are scattered with loss (or gain) of energy corresponding to creation (or absorption) of these excitations. Measurement of the energy change as a function of the change in the neutron wave vector gives the dispersion relation of the excitations: crystal momentum and energy are conserved in the interactions.

The excitations of lowest energy are called *magnons*, and correspond classically to propagating wave-like disturbances of the magnetization in the solid, rather as *phonons* correspond classically to sound waves. A classical derivation of the spin-wave spectrum is surprisingly involved [the details are given by Kittel], so instead we take the easier route of looking at the excited states of a one-dimensional ferromagnetic chain of quantum-mechanical spins with $s = \frac{1}{2}$, each of which is coupled to its neighbours by the exchange interaction. The Heisenberg model Hamiltonian is

$$\hat{H} = -\frac{|J|}{\hbar^2} \sum_n \hat{\mathbf{S}}_n \cdot \hat{\mathbf{S}}_{n+1},$$

where n runs over all N spins in the chain, and the exchange coupling constant has been written as $-|J|$ to emphasize the fact that the interaction is ferromagnetic, and so tends to align the spins.

3.6.1 Groundstate energy

All the spins are parallel in a state of lowest energy; one state of this kind is

$$\chi^{(0)} = \uparrow_1 \uparrow_2 \ldots \uparrow_{N-1} \uparrow_N;$$

it has total $S_z = \frac{1}{2}N$, the largest possible value for N spins-$\frac{1}{2}$.

To work out the energy, we first apply any one of the terms $\hat{\mathbf{S}}_n \cdot \hat{\mathbf{S}}_{n+1}$ to $\chi^{(0)}$; for example, the term with $n = 1$. Now, $\hat{\mathbf{S}}_1 \cdot \hat{\mathbf{S}}_2$ can be rewritten[9]

$$\hat{\mathbf{S}}_1 \cdot \hat{\mathbf{S}}_2 = \tfrac{1}{2}\{\hat{\mathbf{S}}_{12}^2 - \hat{\mathbf{S}}_1^2 - \hat{\mathbf{S}}_2^2\},$$

[9] Recall the "trick" of replacing $\hat{\mathbf{L}} \cdot \hat{\mathbf{S}}$ by $\frac{1}{2}\{\hat{\mathbf{J}}^2 - \hat{\mathbf{L}}^2 - \hat{\mathbf{S}}^2\}$ to work out its eigenvalues in terms of the quantum numbers J, L and S.

where $\hat{\mathbf{S}}_{12} \equiv \hat{\mathbf{S}}_1 + \hat{\mathbf{S}}_2$. Like the square of any angular momentum operator, the eigenvalues of $\hat{\mathbf{S}}_{12}^2$ are $S_{12}(S_{12}+1)\hbar^2$, where $S_{12} = 1$ in the state $\chi^{(0)}$ because the two spins are parallel [see Eq. (3.16), first line]. On the other hand,

$$\hat{\mathbf{S}}_1^2 = S_1(S_1+1)\,\hbar^2 = \tfrac{3}{4}\,\hbar^2,$$

because $S_1 = \frac{1}{2}$; similarly, we can replace $\hat{\mathbf{S}}_2^2$ by $\frac{3}{4}\,\hbar^2$.

Putting the last results together, we find

$$\hat{\mathbf{S}}_1 \cdot \hat{\mathbf{S}}_2\,\chi^{(0)} = \tfrac{1}{2}\{1 \times (1+1)\hbar^2 - 2 \times \tfrac{3}{4}\,\hbar^2\}\chi^{(0)} = \tfrac{1}{4}\,\hbar^2\chi^{(0)},$$

and we can argue in the same way for any pair of spins in the chain. Finally we obtain

$$\hat{H}\chi^{(0)} = -\frac{|J|}{\hbar^2}\sum_{n=1}^{N}\hat{\mathbf{S}}_n \cdot \hat{\mathbf{S}}_{n+1}\,\chi^{(0)} = -\tfrac{1}{4}N|J|\chi^{(0)}, \tag{3.21}$$

which shows that $\chi^{(0)}$ is an eigenstate of $\hat{H}$ with energy eigenvalue $E^{(0)} = -\frac{1}{4}N|J|$.

It is worth pointing out that the ground "state" is actually highly degenerate: the total spin $S = \frac{1}{2}\,N$, so the degeneracy is $2S+1 = N+1$, corresponding to the possible values of $S_z = -S, -S+1, \ldots, S-1, S$. A closely-related point should also be understood: the Hamiltonian has full rotational symmetry (it depends only on the scalar products $\hat{\mathbf{S}}_n \cdot \hat{\mathbf{S}}_{n+1}$), but the state $\chi^{(0)}$ lacks this symmetry, because all the spins are pointing along the z-axis. In this and similar cases we say that the groundstate has *broken symmetry*.

3.6.2 Spin-flip excitations and magnons

The basic excitation can be taken to be a "spin-flip", i.e., a state which differs from the ground state by the reversal of one spin on the nth site,

$$\chi_n = \uparrow_1 \ldots \uparrow_{n-1}\downarrow_n\uparrow_{n+1}\ldots\uparrow_N .$$

The total S_z in this state is 1 less than in the ground state. By itself the spin-flip is *not* a solution of Schrödinger's equation. We can understand this in much the same way that we worked out the ground state energy in the last section,[10] by considering the result of applying $\hat{\mathbf{S}}_m \cdot \hat{\mathbf{S}}_{m+1}$ to χ_n.

Provided $m \neq n-1$ or n, $\hat{\mathbf{S}}_m$ and $\hat{\mathbf{S}}_{m+1}$ will both be acting on up-spins, so the result is the same as in the groundstate:

$$\hat{\mathbf{S}}_m \cdot \hat{\mathbf{S}}_{m+1}\chi_n = \tfrac{1}{4}\,\hbar^2\chi_n \qquad \text{for } m \neq n-1 \text{ or } n; \tag{3.22}$$

but if $m = n$ we must consider the effect of $\hat{\mathbf{S}}_n \cdot \hat{\mathbf{S}}_{n+1}$ on $\downarrow_n\uparrow_{n+1}$. To simplify the calculation, we note that

$$\downarrow_n\uparrow_{n+1} = \tfrac{1}{2}\left(\downarrow_n\uparrow_{n+1} - \uparrow_n\downarrow_{n+1}\right) + \tfrac{1}{2}\left(\downarrow_n\uparrow_{n+1} + \uparrow_n\downarrow_{n+1}\right)$$

is a linear combination of the singlet state (spin 0) with one of the triplet states (spin 1). Just as in the calculation of the groundstate energy, we rewrite $\hat{\mathbf{S}}_n \cdot \hat{\mathbf{S}}_{n+1}$ in terms of its square, so that

$$\begin{aligned}
\hat{\mathbf{S}}_n \cdot \hat{\mathbf{S}}_{n+1}\downarrow_n\uparrow_{n+1} &= \tfrac{1}{2}\left\{\left(\hat{\mathbf{S}}_n + \hat{\mathbf{S}}_{n+1}\right)^2 - \tfrac{3}{2}\,\hbar^2\right\}\left[\tfrac{1}{2}\left(\downarrow_n\uparrow_{n+1} - \uparrow_n\downarrow_{n+1}\right) + \tfrac{1}{2}\left(\downarrow_n\uparrow_{n+1} + \uparrow_n\downarrow_{n+1}\right)\right] \\
&= \tfrac{1}{4}\left[\{0 \times 1\,\hbar^2 - \tfrac{3}{2}\,\hbar^2\}\left(\downarrow_n\uparrow_{n+1} - \uparrow_n\downarrow_{n+1}\right) + \{1 \times 2\,\hbar^2 - \tfrac{3}{2}\,\hbar^2\}\left(\downarrow_n\uparrow_{n+1} + \uparrow_n\downarrow_{n+1}\right)\right] \\
&= -\tfrac{1}{4}\,\hbar^2\downarrow_n\uparrow_{n+1} + \tfrac{1}{2}\,\hbar^2\uparrow_n\downarrow_{n+1},
\end{aligned}$$

which gives, after inserting the states $\uparrow_m$ for the remaining $N-2$ spins,

$$\hat{\mathbf{S}}_n \cdot \hat{\mathbf{S}}_{n+1}\,\chi_n = -\tfrac{1}{4}\,\hbar^2\,\chi_n + \tfrac{1}{2}\,\hbar^2\,\chi_{n+1}. \tag{3.23}$$

[10] The key result in this subsection is (3.25); the steps that lead up to it are *not* examinable. Nevertheless, if you are interested, a more sophisticated (and algebraically simpler) approach to this problem is given in Aschroft & Mermin, *Solid State Physics*, equation (33.19) onwards; also in Lifshitz & Pitaevskii, *Statistical Physics*, Part 2, Sec. 72.

In a very similar way one can show that

$$\hat{\mathbf{S}}_{n-1} \cdot \hat{\mathbf{S}}_n \, \chi_n = -\tfrac{1}{4}\hbar^2 \chi_n + \tfrac{1}{2}\hbar^2 \chi_{n-1}. \tag{3.24}$$

After putting the results of (3.22), (3.23) and (3.24) together and summing over m, we find

$$\begin{aligned} \hat{H}\chi_n &= -\frac{|J|}{\hbar^2} \sum_{m=1}^{N} \hat{\mathbf{S}}_m \cdot \hat{\mathbf{S}}_{m+1} \, \chi_n \\ &= E^{(0)}\chi_n + |J|\big(\chi_n - \tfrac{1}{2}\chi_{n+1} - \tfrac{1}{2}\chi_{n-1}\big). \end{aligned} \tag{3.25}$$

In words, the effect of $\hat{H}$ is to transfer a spin-flip on site n to the neighbouring sites $n \pm 1$. This sounds very similar to the description of an *electron* tunnelling between atoms in the tight-binding model, though it should be emphasized that it is a spin-flip, rather than an actual particle, which is in motion here.

Magnon dispersion relation

The close analogy with the tight-binding model makes it easy to find a wave function that describes a propagating spin-flip. First we construct a Bloch function ϕ_k from the states χ_n,

$$\phi_k = \sum_{m=1}^{N} e^{ikma} \, \chi_m \, .$$

We apply $\hat{H}$ to ϕ_k, which (after using (3.25)) gives

$$\hat{H}\phi_k = \big(E^{(0)} + |J|\big)\phi_k - \tfrac{1}{2}|J| \sum_{m=1}^{N} e^{ikma}(\chi_{m+1} + \chi_{m-1}) \, . \tag{3.26}$$

Provided the chain of spins is closed (so that $\chi_{N+1} \equiv \chi_1$), the first sum in (3.26) can be completed by making a change of variable $n = m + 1$:

$$\sum_{m=1}^{N} e^{ikma}\chi_{m+1} = \sum_{n=2}^{N+1} e^{ik(n-1)a}\chi_n \equiv e^{-ika}\phi_k \, .$$

Similarly, by making the change of variable $n = m - 1$,

$$\sum_{m=1}^{N} e^{ikma}\chi_{m-1} \equiv e^{ika}\phi_k \, .$$

Using these last two results in (3.26) finally gives

$$\begin{aligned} \hat{H}\phi_k &= \big[E^{(0)} + |J| - \tfrac{1}{2}|J|e^{-ika} - \tfrac{1}{2}|J|e^{ika}\big] \, \phi_k \\ &= \big[E^{(0)} + |J| \, (1 - \cos ka)\big] \, \phi_k \, , \end{aligned}$$

showing that ϕ_k is a solution of Schrödinger's equation with energy slightly greater than the groundstate energy $E^{(0)}$. We can interpret the *excess* energy as the energy

$$\epsilon_k = |J|(1 - \cos ka) \tag{3.27}$$

of a magnon with crystal momentum $\hbar k$. Note that the result for ϵ_k is very similar to the energy $E(k)$ of an electron in a tight-binding model with hopping between nearest neighbours; see (2.21).

Exercise 3.9:

Check that you understand (and can reproduce) all the steps leading from (3.25) to (3.27).

For long wavelengths ($k \to 0$) the dispersion relation of the magnon is similar to that of a free particle, $\epsilon_k \simeq \frac{1}{2}|J|k^2a^2$. Neutron scattering data from iron, cobalt and nickel suggest a magnon dispersion of the same form Ak^2, with $A \simeq 0.3$–$0.5\,\mathrm{eV\,\AA^2}$. This would be consistent with an exchange coupling constant $|J|$ in the range 0.1–0.2 eV, as would also be required for a critical temperature in excess of 1000 K. Nevertheless, it should be borne in mind that the model is not precisely comparable with experiment here. The magnetization density in these metals is distributed in the gas of conduction electrons, rather than being located solely on the ions as required by the model.

In antiferromagnets the long-wavelength dispersion relation is found experimentally to be linear rather than quadratic, $E \propto k$. An analysis starting from Schrödinger's equation is more difficult here, and we do not attempt it.[11] In fact, it is a far from trivial task to work out even the ground state wave function of the 1D Heisenberg antiferromagnet: it is not the "obvious" state $\ldots\uparrow\downarrow\uparrow\downarrow\ldots$ that one might have guessed by analogy with the ferromagnet.

Effect of magnons on the magnetization at low temperature

Experimentally, the magnetization, $M(T)$, of a ferromagnet is found to follow a power law at low temperatures,

$$[M(0) - M(T)] \propto T^{3/2}.$$

The reason for the power law dependence on T is the presence of low-energy magnetic excitations, even at the lowest temperatures. These excitations are the long-wavelength magnons, for which $\epsilon_k \simeq Ak^2$. Each magnon has $s_z = -1$ (it corresponds to a single delocalized spin-flip), so that as the temperature rises the magnetization decreases in proportion to the number of magnons present.

Now, we expect a given spin-wave mode of the solid to be thermally excited if the magnon energy is less than about k_BT, so here we expect all modes to be excited up to $k_{\max}$, where $Ak_{\max}{}^2 \simeq k_BT$, or $k_{\max} \propto T^{1/2}$. The total number of excited modes is proportional to the volume of reciprocal space with $k < k_{\max}$; this volume is in turn proportional to $k_{\max}{}^3$, or $T^{3/2}$. The power-law dependence on temperature is usually known as *Bloch's* $T^{3/2}$ *law*.[12]

Exercise 3.10:

Optional: see Q. 3 on the final examples sheet for PHYS30151. The above argument for the $T^{3/2}$ law takes no account of the fact that different modes contain different numbers of magnons. Correct the argument on the assumption that magnons are non-interacting, non-conserved *bosons*, so that the average number with wave vector $\mathbf{k}$ follows the Bose–Einstein distribution law,

$$\langle n_{\mathbf{k}} \rangle = \frac{1}{e^{\epsilon_{\mathbf{k}}/k_BT} - 1},$$

with chemical potential $\mu = 0$.

3.7 Mean-field theory of the critical point

So far we have discussed the magnetization only at low temperature. As the temperature rises towards T_c, the magnetization steadily decreases to zero as illustrated in Fig., and it is zero at all temperatures above T_c. This general behaviour is not intuitively obvious, and it requires an explanation: why shouldn't the magnetization remain non-zero for *all* temperatures, tending to zero only for $T \to \infty$? Alternatively, why

[11] A classical treatment is given in Kittel's textbook, if you are interested.

[12] Not to be confused with his T^5 law for the phonon-induced electrical resistivity in metals.

shouldn't it jump *discontinuously* to zero at the critical temperature? These things, at least, we will be able to explain. We should also like (over-optimistically, it turns out) to understand some of the details of the temperature dependence. For example, just below the critical temperature it is found experimentally that

$$M(T) \propto (T_c - T)^{\beta},$$

with $\beta \simeq 0.33$ typically; the power β is a particular case of a *critical exponent*. In large, homogeneous samples the magnetic susceptibility of a ferromagnet diverges as

$$\chi(T) \propto |T_c - T|^{-\gamma},$$

defining another critical exponent γ which typically lies in the range 1.2–1.3. Indeed, *all* of the macroscopic, thermodynamic quantities show some kind of unusual behaviour at the critical point. Our aim is to understand this at the simplest level of approximation, so we should not be too concerned if some of the details turn out to be wrong.

Making the mean-field approximation

To get a simple picture of the ferromagnetic phase transition we can make a *mean-field* approximation, in which the effect of the neighbours of a given spin $\mathbf{s}_0$ is replaced by an average.

In the magnetic state, the average magnetic moment of a spin is non-zero:

$$\langle \boldsymbol{\mu}_i \rangle = -g\mu_B \langle \mathbf{s}_i \rangle / \hbar \neq \mathbf{0}\,, \tag{3.28}$$

where the angled brackets denote the usual thermal average. If all the spins have the same atomic environment this average will be the same for *every* spin: $\langle \boldsymbol{\mu}_i \rangle \equiv \langle \boldsymbol{\mu} \rangle$, independent of i.

Any given spin tends, on average, to point in the same direction as its neighbours; this is the result of the ferromagnetic Heisenberg exchange interaction, which favours the alignment of the spins. Focus, for the moment, on a particular spin, $\mathbf{s}_0$, with n nearest neighbours; its energy is

$$\begin{aligned} E = -\frac{|J|}{\hbar^2} \sum_{j=1}^{n} \mathbf{s}_0 \cdot \mathbf{s}_j &= -\frac{n|J|}{\hbar^2}\, \mathbf{s}_0 \cdot \left[\frac{\sum_{j=1}^{n} \mathbf{s}_j}{n} \right] \\ &\equiv -\frac{n|J|}{g^2\mu_B^2}\, \boldsymbol{\mu}_0 \cdot \left[\frac{\sum_{j=1}^{n} \boldsymbol{\mu}_j}{n} \right] \\ &\simeq -\frac{n|J|}{g^2\mu_B^2}\, \boldsymbol{\mu}_0 \cdot \langle \boldsymbol{\mu}_j \rangle\,, \end{aligned} \tag{3.29}$$

where, in the last line, we have replaced the average over the n nearest neighbours by the average over the whole solid: this replacement is *the mean-field approximation*. It certainly cannot be *exact*, but we might hope that it would be a reasonably good approximation when n is large. Typical values of n are 8 or 12, for the BCC and FCC structures, respectively.

We can re-write (3.29) in the same form as the interaction of a spin with a effective magnetic field, $\mathbf{B}_{\text{eff}}$,

$$E \simeq -\boldsymbol{\mu}_0 \cdot \left(\frac{n|J|}{g^2\mu_B^2} \langle \boldsymbol{\mu}_j \rangle \right) \equiv -\boldsymbol{\mu}_0 \cdot \mathbf{B}_{\text{eff}}\,. \tag{3.30}$$

From this point onward, we shall assume, for simplicity, that the spins have $s = \frac{1}{2}$, so that $g = 2$; we also assume that $\langle \boldsymbol{\mu} \rangle$ and $\mathbf{B}_{\text{eff}}$ are in the z-direction. With these assumptions, the magnitude of the effective field will be

$$B_{\text{eff}} = \frac{n|J|}{4\mu_B^2} \langle \mu_z \rangle\,. \tag{3.31}$$

The mean-field magnetization

The energy of a spin in a magnetic field **B** in the z-direction is $E = \mp\mu_B B$, the two values corresponding to $\mu_z = \pm\mu_B$. In Section 3.3.1 we found the thermal average of the magnetic moment to be given by

$$\langle \mu_z \rangle = \mu_B \tanh[\mu_B B / k_B T] \,; \tag{3.32}$$

we now simply replace B by B_{eff} [defined in (3.31] and introduce a temperature scale T_c, defined by $k_B T_c \equiv n|J|/4$. [The notation anticipates the result, of course: T_c will turn out to be the critical temperature in the mean-field approximation.] After making these replacements in (3.32) we find

$$\langle \mu_z \rangle / \mu_B = \tanh\left[\frac{T_c \,\langle \mu_z \rangle / \mu_B}{T}\right]. \tag{3.33}$$

The quantity $\langle \mu_z \rangle / \mu_B \equiv m$ is a scaled magnetic moment, which takes values between -1 and 1. In terms of m, the last equation can be written

$$m = \tanh[T_c \, m / T] \,, \tag{3.34}$$

which is the basic equation of the mean-field approximation for a spin-half ferromagnet. In principle, we can solve it to obtain $m(T)$.

Temperature dependence of m

We can predict the general behaviour of $m(T)$ by a graphical method. First we rearrange (3.34) as

$$\operatorname{arctanh} m = m \, T_c / T \,. \tag{3.35}$$

Each side of the equation can be sketched as a function of m; the solution $m(T)$ is given by the intersection of the two curves.

If T is too large, the straight line $y = m \, T_c / T$ falls below the curve $y = \operatorname{arctanh} m$, and the only solution of (3.35) is the trivial one, $m = 0$. In this case (high temperature) the expected magnetization is zero.

On the other hand, if T is small the two curves intersect three times. The new solutions with $m \neq 0$ correspond to a spontaneous magnetization that increases with decreasing T. There is a definite critical temperature T_c at which this magnetization appears, and we can calculate the behaviour of the magnetization when it is small, at temperatures just below T_c. Using the series expansion $\operatorname{arctanh} m \simeq m + \frac{1}{3}m^3$, valid for small m, we find

$$m + \tfrac{1}{3} \, m^3 \simeq m \, T_c / T \,.$$

After dividing through by m (we ignore the trivial solution $m = 0$) and rearranging a little, we find

$$\tfrac{1}{3} \, m^2 \simeq \frac{T_c}{T} - 1 \qquad \text{or} \qquad m^2 \simeq 3(T_c - T)/T \,,$$

which is zero when $T = T_c$. Just below the transition, therefore, the magnetization varies as $(T_c - T)^{1/2}$. This is a definite prediction of the mean-field theory: the critical exponent for the magnetization is $\beta = \frac{1}{2}$. It does not agree very well with the value $\beta \simeq 0.33$ seen in experiment.

The observed deviations from mean-field theory at the critical point are significant and a great deal of experimental and theoretical work has been done to clarify their origin. This is a useful thing to do because the critical behaviour is fairly insensitive to the details of the interactions between magnetic moments, so that many different magnetic solids have the *same* critical exponents. It is difficult, unfortunately, to give a short, qualitative account of precisely *how* the theoretical description of the critical point has been put into agreement with experiment.

The real value of the mean-field approximation is its simplicity and generality, rather than its accuracy, so that it can provide a qualitative understanding of even quite complicated magnetic systems. Nevertheless, the mean field theory has been found to be *quantitatively* accurate for at least *one* material, $HoRh_4B_4$.

It is an alloy containing rare-earth ionic spins, whose interactions with one another is probably best described by the Rudermann–Kittel interaction. The coupling constant J is thought to take the same sign for the interaction of a spin with a large number of its neighbours, not just the ones nearest to it in the crystal lattice. In this case the sum $\sum_{j=1}^{n} \boldsymbol{\mu}_j$ is unusually well approximated by $n\langle\boldsymbol{\mu}\rangle$, because the interaction samples many more than 8 or 12 spins. Even so, we should expect more accurate experiments to show deviations from the mean-field behaviour at temperatures very close to the critical point.

Other weaknesses of mean-field theory

Mean-field theory also makes an incorrect prediction for the behaviour of the magnetization at temperatures well below T_c. In this régime (3.34) reduces to

$$\langle m(T)\rangle \simeq 1 - 2\,e^{-2T_c/T}, \quad \text{or} \quad [M(0) - M(T)] \propto e^{-2T_c/T},$$

so that the curve of magnetization as a function of temperature can be expected to be very flat in the region $T \ll T_c$. But as we have discussed before, the low temperature magnetization is actually observed to follow a power law,

$$[M(0) - M(T)] \propto T^{3/2},$$

which is entirely different from the mean-field prediction, but which we have explained by the presence of low-energy excitations (the long-wavelength magnons) at low temperature. These excitations are completely ignored by the mean-field approximation.

> The situation just described is reminiscent of what happens with the heat capacity of a non-magnetic insulating solid at low temperature. The Debye theory of the specific heat shows—and experiments agree—that $C(T)$ varies as T^3, because at low temperatures the thermal properties are dominated by long-wavelength phonons, which are the only low-energy excitations of an insulator. Like mean-field theory, the simpler Einstein model of a solid incorrectly predicts that $C(T)$ vanishes as $\exp[-\Theta/T]$ at low temperatures, where Θ is a characteristic temperature. You could, in fact, think of Einstein's model as *being* a mean-field approximation in which each atom vibrates independently in an averaged, harmonic potential due to its neighbours. This potential is the same for every atom, so the vibrational frequency ω_E is the same for each atom; the characteristic temperature is then related to this frequency by $\Theta \sim \hbar\omega_E/k_B$.

Total energy and heat capacity in the mean-field approximation

This was not covered in the lectures in 2017, so the topic is not examinable this year.