MP464: Solid Sate Physics

Brian Dolan

1. Introduction

Broadly speaking there are three common states of matter: solid, liquid and gas, though plasmas and other more exotic states can also be legitimately called different states of matter. Thermodynamics studies all states of matter in general terms while fluid dynamics deals with properties specific to liquids and gases. Solid State physics describes the properties of solids.

Examples of solids at room temperature are: rocks, metals (except mercury), ice, glass and wood. This course will deal exclusively with one type of solid — crystals (rocks and metals are made up of crystals, glass and wood are not crystals). The regular structure of crystals makes it easier to construct realistic mathematical models of them, the cellular structure of wood is much more complicated at a microscopic level than a crystal. While this restriction to crystals may seem rather narrow it is in fact more general than one might think: metals and rocks are in fact made up of an agglomeration of large numbers of small crystals. While the crystal structure is obvious in some rocks, such as the sample of Iron Pyrites shown below, it is not obvious in metals where the crystals are usually too small to see without a microscope.

Some crystals can be very large, metres across like the ones shown here from a mine in Mexico[1]

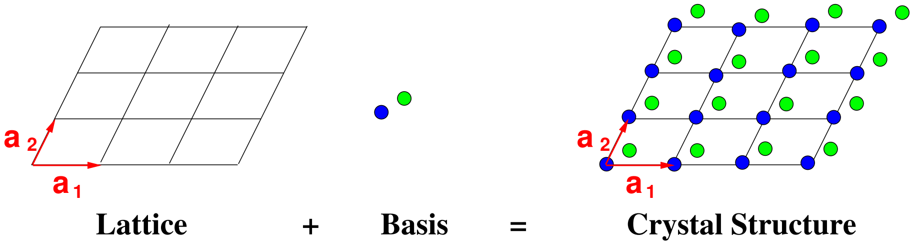
[1] It has even been suggested by some geologists, based on analysis of seismic data and computer modelling of the quantum mechanical properties of iron at high pressure, that the inner core of the Earth might be a single crystal of iron more than 2400 km in size — but this is speculative.

2. Lattices and crystals

A **crystal** is a periodic array of atoms or molecules in a regular lattice structure. Mathematically a **lattice** is a rigid, periodic array of points that looks exactly the same from every point and is infinite in extent. Putting an atom, a group of atoms or a molecule (a **basis**) at every point of a lattice gives a crystal structure.

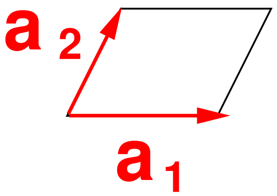
Crystal structure = Lattice + Basis.

Below is a two dimensional representation of this concept. The blue and green dots represent atoms, *e.g.* Zn and S for a crystal of Zinc Sulphide. A lattice is an abstract mathematical structure that is completely determined by a set of basis vectors, $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ below, which, when combined with the basis, gives a representation of a crystal,[2]



A lattice is defined by a set of **primitive lattice vectors**, such as $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ in the two dimensional example. The definition of a set of primitive lattice vectors is that any lattice vector $\mathbf{L}$ can be expressed as a linear combination of primitive lattice vectors, $\mathbf{L} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2$, with integer co-efficients. Primitive lattice vectors describe a **primitive cell** of the lattice, a parallelogram in this case,

[2] Real crystals do not have infinite extent, of course, but even small crystals of a milligramme can have 10^{20} atoms in them so it not unreasonable to model them with a lattice of infinite extent.



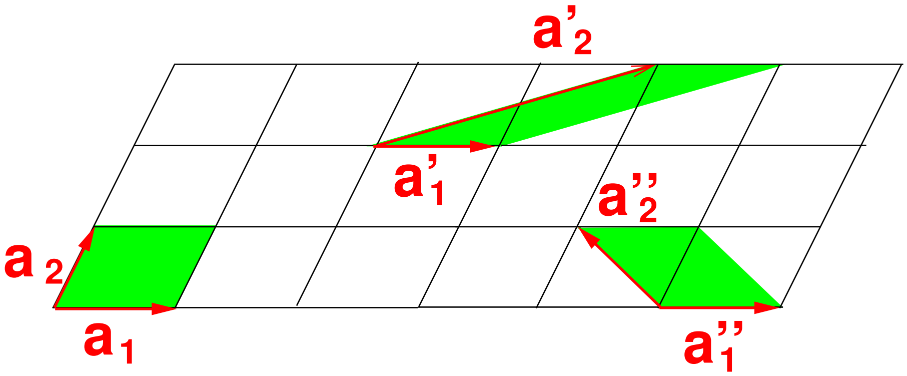
It may be useful to think of a two-dimensional lattice as a tiling of the two-dimensional plane by primitive cells. A primitive cell need not be a parallelogram. By definition a primitive cell contains one complete lattice point and only one complete lattice point.

A general point in a two dimensional lattice is described by a **lattice vector**

$$\mathbf{L} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2$$

defined by two integers n_1 and n_2.

Primitive lattice vectors and primitive cells are not unique, the pairs $(\boldsymbol{a}_1, \boldsymbol{a}_2)$, $(\boldsymbol{a}_1', \boldsymbol{a}_2')$ and $(\boldsymbol{a}_1'', \boldsymbol{a}_2'')$ in the figure below are all primitive lattice vectors and the green shapes are all possible primitive cells,



The three green shapes in the figure above all have the same area,

$$|\boldsymbol{a}_1 \times \boldsymbol{a}_2| = |\boldsymbol{a}_1' \times \boldsymbol{a}_2'| = |\boldsymbol{a}_1'' \times \boldsymbol{a}_2''|.$$

A three-dimensional lattice is described by three primitive lattice vectors $(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$, lattice vectors are defined by three integers, n_1, n_2 and n_3,

$$\mathbf{L} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3,$$

and all primitive three dimensional cells have the same volume

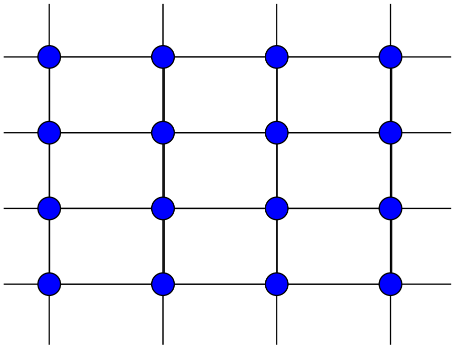
$$V_c = |\boldsymbol{a}_1.(\boldsymbol{a}_2 \times \boldsymbol{a}_3)|.$$

Symmetries

The set of all possible lattices can be classified by their symmetries:

- All lattices are symmetric under translations by any lattice vector (all lattice points move under such a translation);
- Symmetries leaving at least one lattice point fixed are called **point symmetries** — the set of all point symmetries is called the **point group** of the lattice. Point symmetries are: rotations about a lattice point; reflections in lines or planes containing a lattice point and inversion about a lattice point (any given lattice might not have all of these symmetries).
- The combination of all lattice translations and the point group of the lattice is called the **space group** of the lattice.

As an example in 2-dimensions, consider the pattern below and imagine it to be infinitely extended in both directions:



When extended this rectangular pattern is symmetric under rotations through π about any point and, of course, rotations though 2π which just brings the pattern back to its original orientation. The pattern is also symmetric under reflections about any of the marked horizontal lines, we shall represent such reflections by the symbol M_1 (M for mirror), and reflections about any of the vertical lines, which we shall represent by M_2. The rotations leave precisely one point fixed while the reflections leave an entire line of points fixed, these operations are part of the point group. Combining any two symmetry operations that leave the same point fixed should also be a symmetry of the point group: for example we could perform M_1 followed by a rotation through π, this does not give a new symmetry operation because it is completely equivalent to M_2 (convince yourself of this).

To understand the point group in more detail it is useful to draw up a table that shows the result of combining any two symmetry operations, this is called a *group multiplication table.* Denote a clockwise rotation though an angle θ by θ itself and the result of doing nothing at all (or rotating through 2π) by **1** then the table below shows the result obtained by first applying the operation in the top row and then applying the operation in the first column. We get a 4×4 table because we must include **1** in order to complete the table.

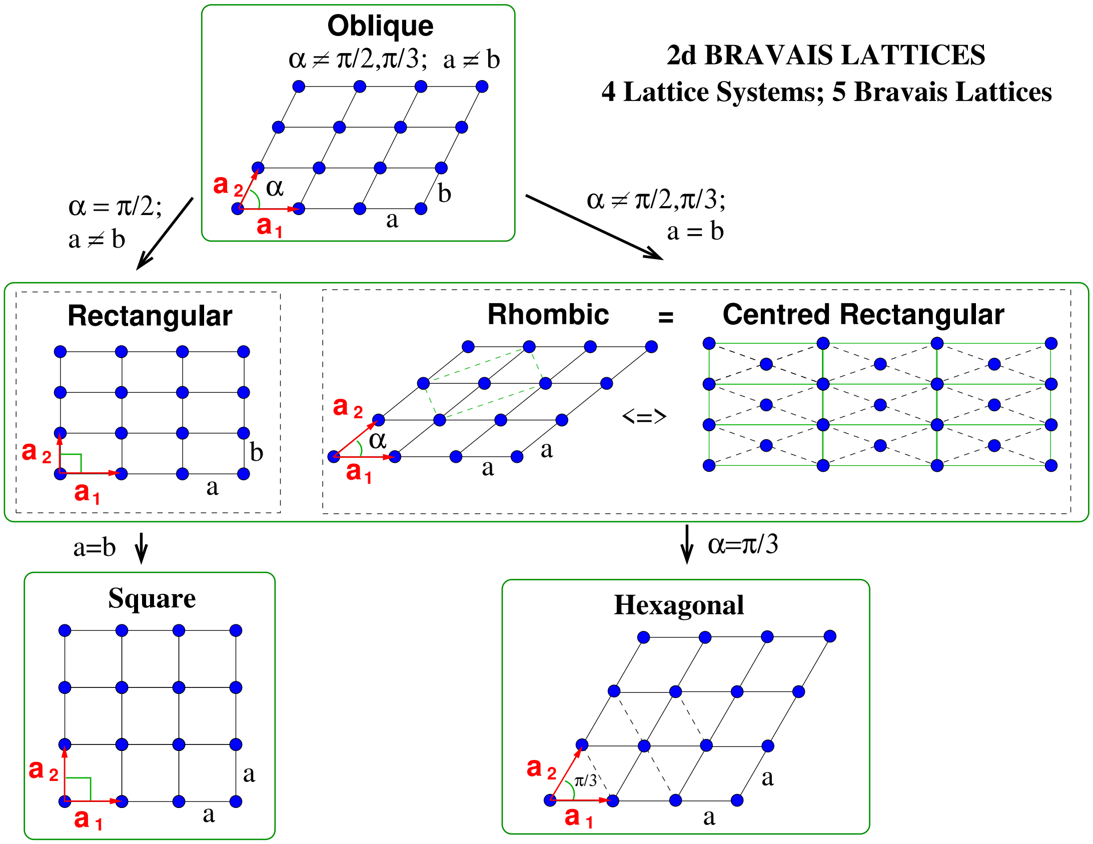
	$\mathbf{1}$	π	M_1	M_2
$\mathbf{1}$	$\mathbf{1}$	π	M_1	M_2
π	π	$\mathbf{1}$	M_2	M_1
M_1	M_1	M_2	$\mathbf{1}$	π
M_2	M_2	M_1	π	$\mathbf{1}$

Note that any symmetry multiplied by **1** just reproduces the symmetry itself, so **1** is called the *identity* operation. Also each row and each column contains a **1**, any two operations that combine to produce a **1** are called inverses of each other and every entry has an inverse. The requirement that applying any two symmetry operation must produce another symmetry and that every operation has an inverse in the multiplication table puts very strong restrictions on the number of consistent multiplication tables that can be constructed. All possible point groups have a finite number of elements and have been classified and listed by mathematicians.

Space groups have a (countably) infinite number of elements, because there are an infinite number of lattice vectors available for translations, but nevertheless all possible space groups can also be classified and listed. This means that all possible lattice structures can be classified and in three dimensions this was first achieved by the French physicist Bravais in 1850. For this reason these lattices are called Bravais lattices. Sometimes there is more than one space group with the same point group as we shall see below.

Two dimensional lattices

For simplicity we start with two dimensional lattices. In two dimensions there are 4 possible point groups (giving rise to 4 lattice systems) and 5 possible space groups (giving rise to 5 inequivalent lattices). The possibilities are shown below (lattice points are indicated by blue dots for clarity):



In two dimensions the only possible point symmetries are:

i) Rotations by $\frac{\pi}{3}$, $\frac{\pi}{2}$ and multiples of these, namely $\frac{2\pi}{3}$, π, $\frac{4\pi}{3}$, $\frac{3\pi}{2}$ and $\frac{5\pi}{3}$.

ii) Reflection in a line.

All two-dimensional lattices have rotations by π as part of their space group, the complete set of possibilities is:

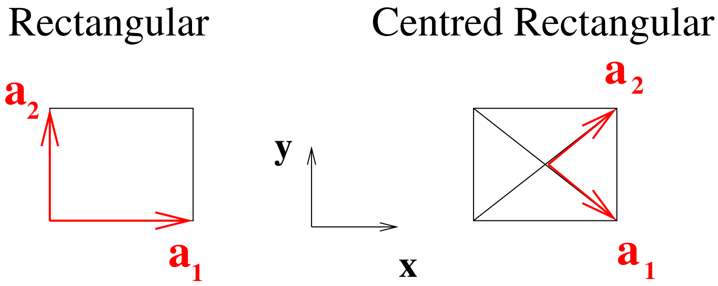
$$\begin{array}{l} \text{4 lattice systems} \\ \text{(point groups)} \end{array} \left\{ \begin{array}{ll} \pi \text{ only} & \text{Oblique} \\ \pi + \text{reflections} & \left\{ \begin{array}{l} \text{Rectangular} \\ \text{Centred Rectangular} \end{array} \right. \\ \text{multiples of } \frac{\pi}{2} + \text{reflections} & \text{Square} \\ \text{multiples of } \frac{\pi}{3} + \text{reflections} & \text{Hexagonal} \end{array} \right\} \quad \text{5 Bravais lattices.}$$

Although the rectangular and centred rectangular lattices share the same point group they are different because they have different space groups, as can be seen by combining reflections with translations. If M_1 represents reflection in the x-axis and M_2 reflection in the y-axis then, for the rectangular lattice

$$M_1 : \begin{cases} \boldsymbol{a}_1 \to \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \to -\boldsymbol{a}_2 \end{cases} \qquad M_2 : \begin{cases} \boldsymbol{a}_1 \to -\boldsymbol{a}_1 \\ \boldsymbol{a}_2 \to \boldsymbol{a}_2, \end{cases}$$

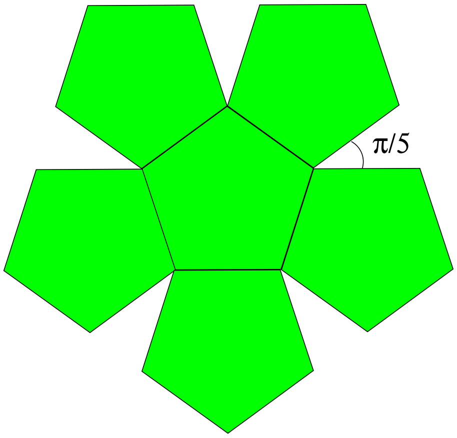
while, for the centred rectangular lattice

$$M_1 : \begin{cases} \boldsymbol{a}_1 \to \boldsymbol{a}_2 \\ \boldsymbol{a}_2 \to \boldsymbol{a}_1 \end{cases} \qquad M_2 : \begin{cases} \boldsymbol{a}_1 \to -\boldsymbol{a}_2 \\ \boldsymbol{a}_2 \to -\boldsymbol{a}_1. \end{cases}$$

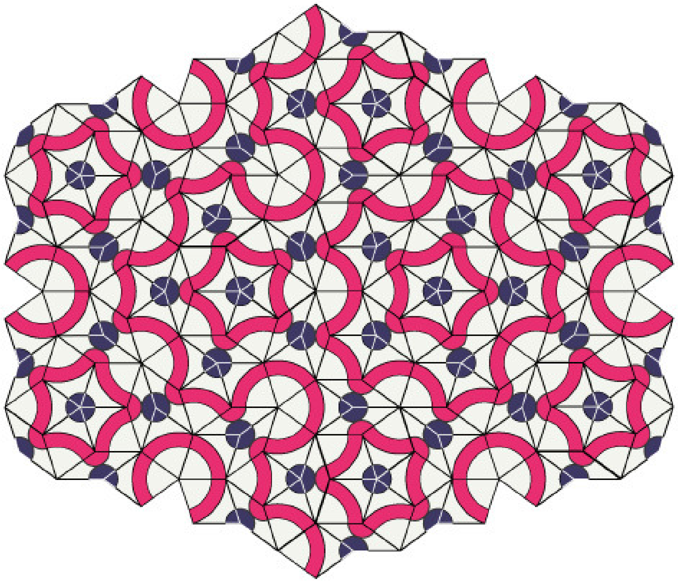


Thus M_1 interchanges $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ for the centred rectangular lattice, and this is a symmetry. There is no such symmetry for a general rectangular lattice, unless $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ have the same length in which case the a lattice is square and has a different space group with more rotational symmetries.

Note that rotations by $\frac{2\pi}{5}$ is not a possibility — it is not possible to tile a two dimensional plane with a single shape with 5-fold symmetry, the figure below shows the kind of thing that goes wrong if we try to do so,

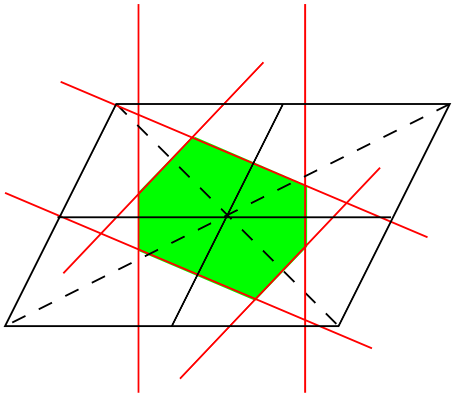


Curiously it is possible to tile the two dimensional plane with a 5-fold symmetric pattern (point group consisting of rotations by $\frac{2\pi}{5}$) but which has *no* translational symmetries at all: the pattern *never* repeats, and so does not fall into the category of crystals by our definition. This pattern requires two different rhombic tiles for its construction and is called a **Penrose tiling**,



Structures similar to this have been seen in Nature, they are called quasi-crystals, but we shall not be describing these any further in this course.

Before going on to describe the classification of three-dimensional lattices we first describe the construction of a special primitive cell, called a **Wigner-Seitz cell**. To construct a Wigner-Seitz cell first pick any lattice point and draw lines connecting it to all its neighbours. Bisect these lines at right-angles and the bisectors enclose a Wigner-Seitz cell.

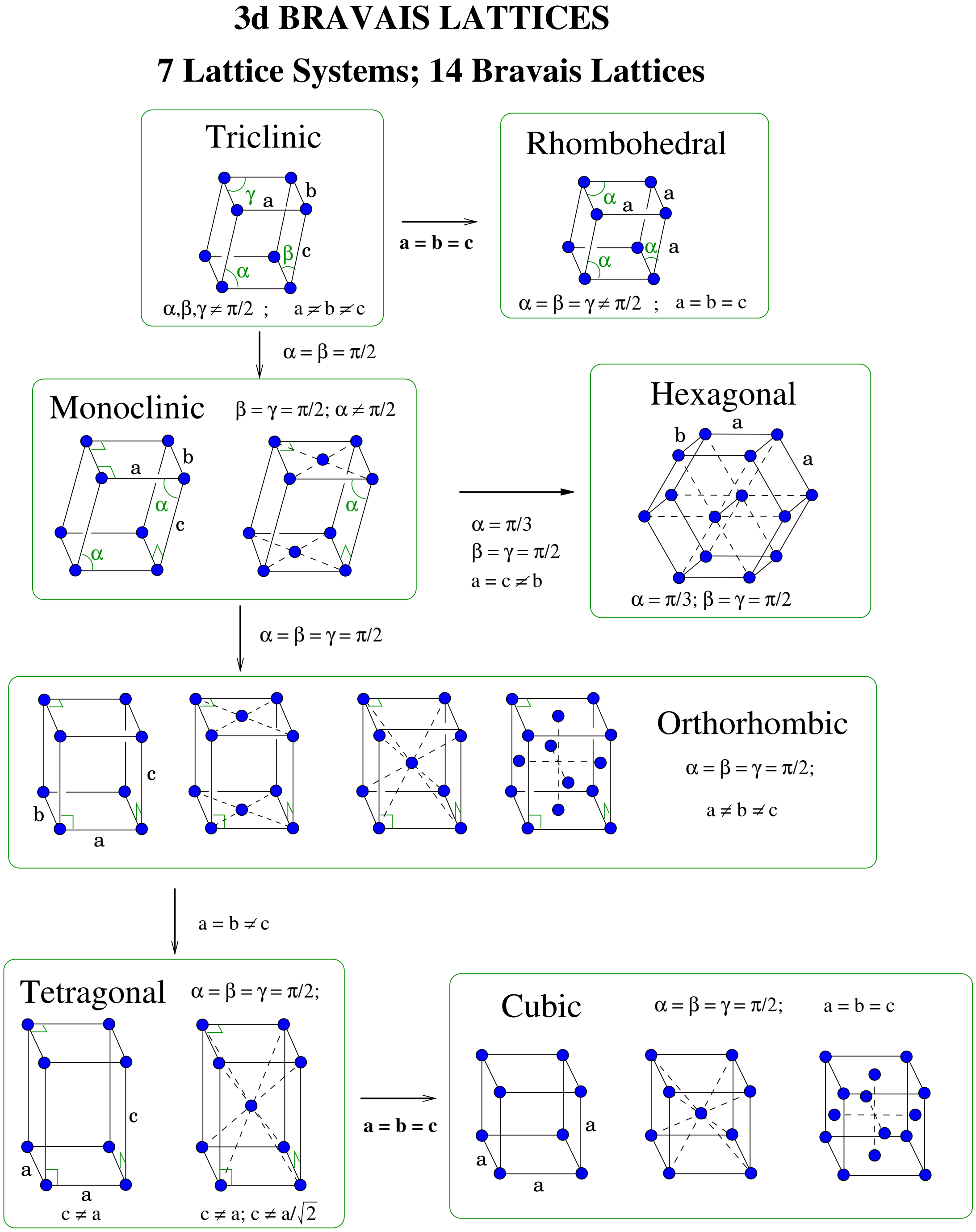


In the figure above solid black lines enclose primitive cells, the parallelograms described earlier, and dotted black lines link other neighbours to the chosen lattice point, at the centre of the green shape. Red lines represent perpendicular bisectors of all the black lines, both solid and dashed. The red lines enclose the six-sided green shape, which is a Wigner-Seitz cell for this lattice — it has the same area as one of the parallelograms.

Three dimensional lattices

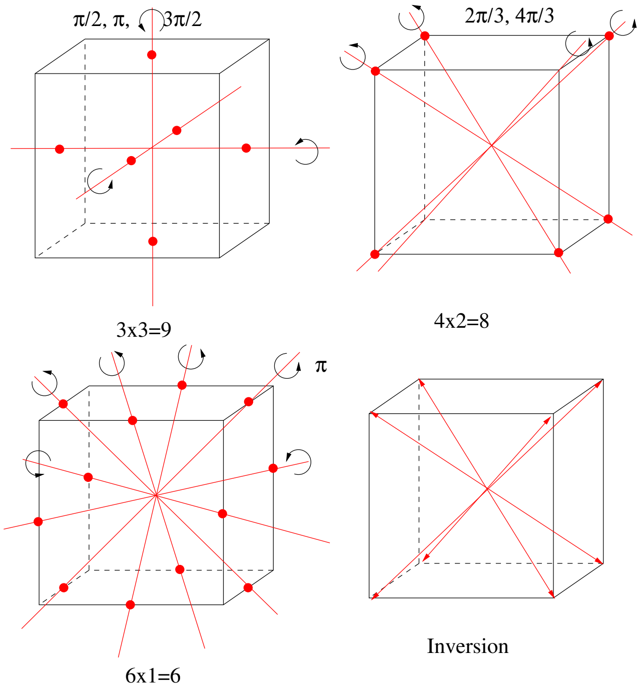
In three dimensions the only possible allowed rotations of a crystal are the same set as in 2-dimensions, but around any one of three axes. There can be up to three reflection planes and inversion in an origin corresponds to a reflection plus a rotation of π radians (in 2-dimensions reflection in the origin is completely equivalent to a rotation through π).

There are 7 possible point groups in 3-dimensions, giving different 7 lattice systems, with 14 different space groups and hence 14 inequivalent Bravais lattices:



We shall consider four of the simpler cases in more detail. Firstly the three cubic lattices all have space groups which are the symmetries of a cube, which include rotations,

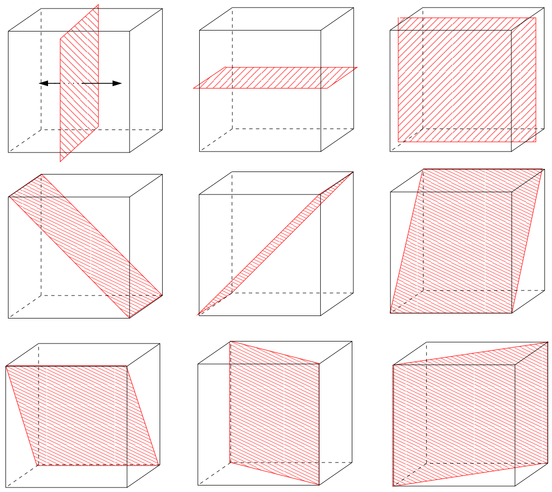
Symmetries of a cube (Octahedral group)



Including the identity gives 24 proper (chiral) operations;
Including inversion gives 24 achiral operations = 48 in total.

and reflections in various planes,

Examples of mirror (achiral) symmetries of a cube, reflection in a plane



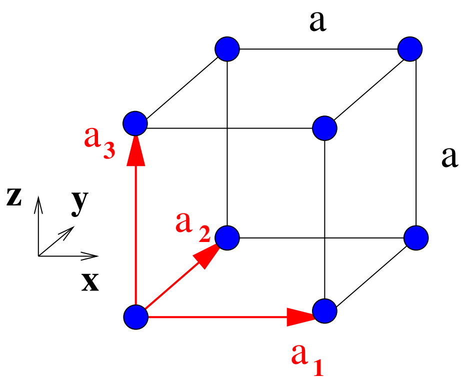
1. Simple cubic lattice

For the simple cubic lattice we can choose primitive lattice vectors to be

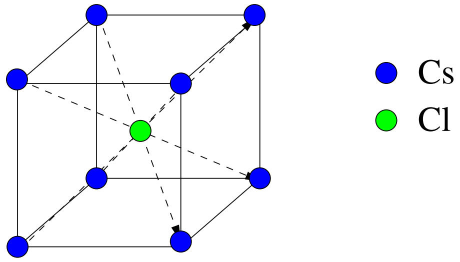
$$\boldsymbol{a}_1 = a\hat{\mathbf{x}}, \qquad \boldsymbol{a}_2 = a\hat{\mathbf{y}}, \qquad \boldsymbol{a}_3 = a\hat{\mathbf{z}}.$$

The volume of a primitive cell is

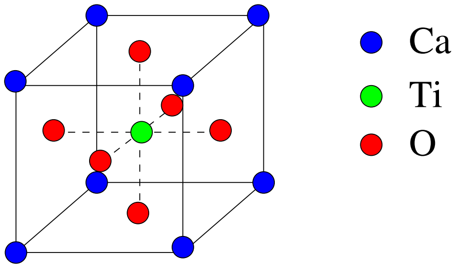
$$V_c = |\boldsymbol{a}_1.(\boldsymbol{a}_2 \times \boldsymbol{a}_3)| = a^3.$$



Examples of materials that crystallise in simple cubic form are Nitrogen (at 20° K), Caesium Chloride (CsCl with $a = 0.411$Å) and the mineral Perovskite ($CaTiO_3$ with $a = 2.94$Å). [3]



Ceasium Chloride



Perovskite

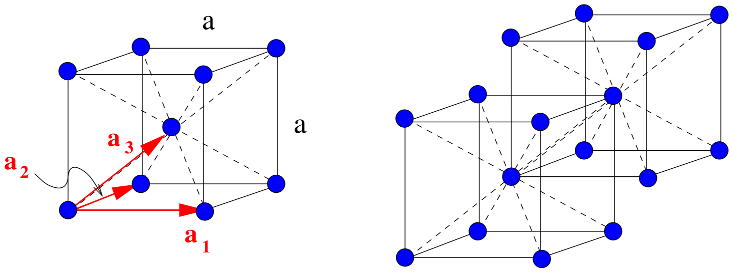
There is one full Caesium atom in each primitive cell of a CsCl crystal, there are eight blue dots at the vertices of the cube, but only one-eighth of each dot is inside the primitive cell. Similarly there are six red dots on the faces of the cube for $CaTiO_3$ but only half of each dot is inside the cube, so there are three Oxygen atoms in each primitive cell.

Note that CsCl and $CaTiO_3$ have different crystal structures, but the same lattice structures — their bases are different.

[3] The Perovskite structure is an important ingredient in geology: it is believed that the lower part of the Earth's mantle, between 700 and 2,500 km down, could be more than 90% brigmanite — Magnesium Silicate ($MgSiO_3$) with the Perovskite structure — which is probably the most common mineral on Earth.

2. Body centred cubic

Putting an extra lattice point at the centre of every primitive cell of a simple cubic lattice gives a distinct lattice structure called **body centred cubic**. A body centred cubic lattice can be viewed as two interwoven simple cubic lattices, as shown on the right below.



The picture on the left above is not a primitive cell, it contains two lattice points, but is still a useful way of visualising a body centred cubic lattice — it is called a **conventional cell**. A set of primitive lattice vectors is shown above,

$$\boldsymbol{a}_1 = a\hat{\mathbf{x}}, \qquad \boldsymbol{a}_2 = a\hat{\mathbf{y}}, \qquad \boldsymbol{a}_3 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}).$$

The volume of a primitive cell is

$$V_c = |\boldsymbol{a}_1.(\boldsymbol{a}_2 \times \boldsymbol{a}_3)| = \frac{a^3}{2}.$$

An alternative set, which is more symmetric, is

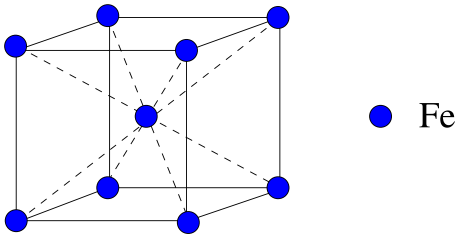
$$\boldsymbol{a}'_1 = \frac{a}{2}(-\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}) \qquad \boldsymbol{a}'_2 = \frac{a}{2}(\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}}) \qquad \boldsymbol{a}'_3 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}}),$$

with has the same volume,

$$V_c = |\boldsymbol{a}'_1.(\boldsymbol{a}'_2 \times \boldsymbol{a}'_3)| = \frac{a^3}{2},$$

as it must do if it is to be a primitive cell.

Examples of materials that crystallise in body centred form are iron, Fe, potassium, K, and Sodium, Na.

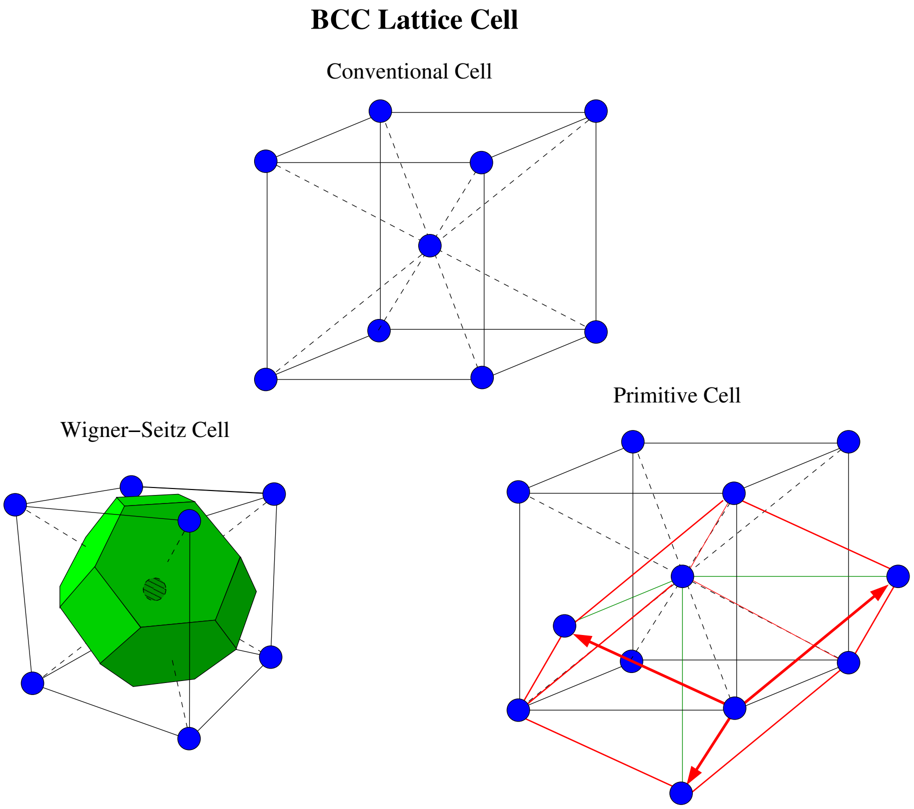


Iron

Water ice forms hexagonal crystals at atmospheric pressure and freezing temperature (0°C) but at high pressure (2.3 GPa, about 23,000 times atmospheric pressure at sea level) water ice forms at room temperature and takes a body centered cubic structure, known as hot ice (technically ice VII, water has a very rich phase structure in solid form and there are at least 19 different forms of ice[4]).

Note that CsCl is *not* a body centred lattice: the Cl atom at the centre of the cell is different to the Ce atoms at the vertices, so the central point is not equivalent to the vertices — it is not a lattice point. Do not confuse the lattice structure of CsCl with that of Iron — they are different.

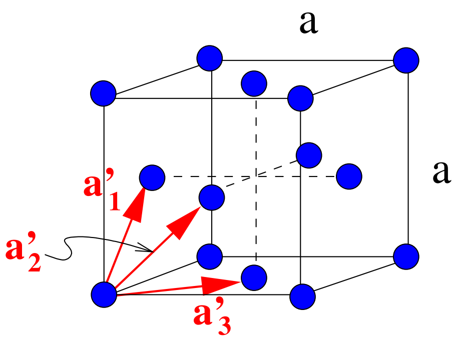
The Wigner-Seitz cell for a body centred cubic lattice is a truncated octahedron:



3. Face centred cubic

Putting an extra lattice point at the centre of the faces of a primitive cell of a simple cubic lattice gives another distinct lattice structure called **face centred cubic**. A face centred cubic lattice can be viewed as four interwoven simple cubic lattices.

[4] With apologies to Kurt Vonnegut ice IX takes a tetragonal phase which is only stable below 140 K and pressures between 200 MPa and 400 MPa.



The picture above is not a primitive cell because it contains four lattice points, it is a **conventional cell** of the face centred lattice. A set of primitive lattice vectors, as shown above, is

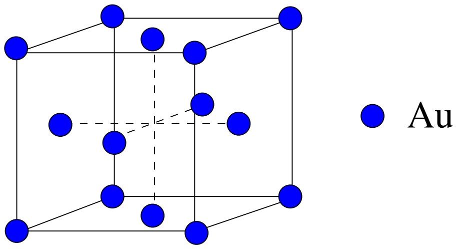
$$\boldsymbol{a}_1' = \frac{a}{2}(\hat{\mathbf{y}} + \hat{\mathbf{z}}) \qquad \boldsymbol{a}_2' = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{z}}) \qquad \boldsymbol{a}_3' = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}}),$$

The volume of a primitive cell is

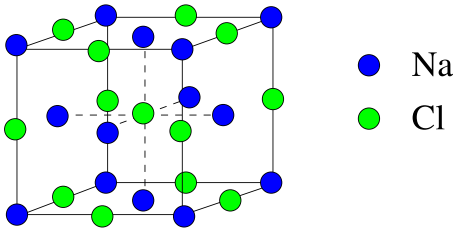
$$V_c = |\boldsymbol{a}_1.(\boldsymbol{a}_2 \times \boldsymbol{a}_3)| = \frac{a^3}{4},$$

where a is the size of a conventional cell.

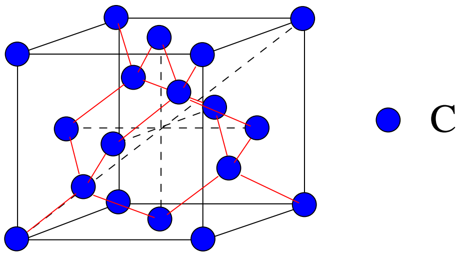
Examples of metals that crystallise in face centred form are aluminium, gold and lead, with bases consisting of a single atom at every lattice site.



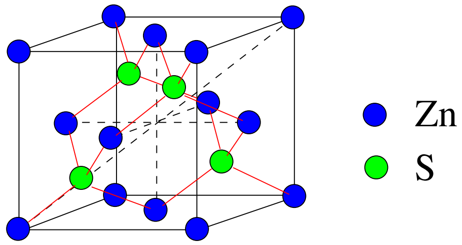
Salt, NaCl, is face centred, with $a = 3.56$Å, it is *not* simple cubic!



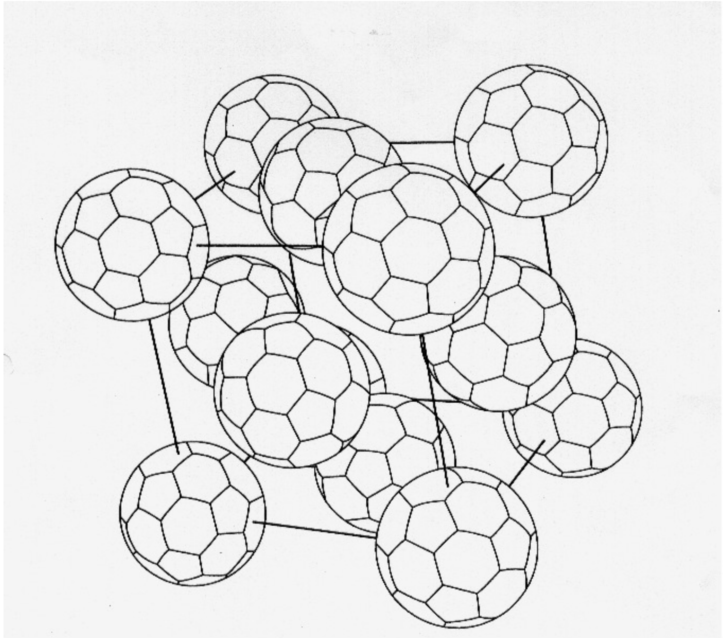
Diamond has a face centred structure with a basis consisting of two carbon atoms, one at the origin (front-bottom-left corner) and one at $\frac{a}{4}(\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}})$, and a identical pair at all lattice sites of course. This structure allows each carbon to be linked to its four nearest neighbours, each a distance $\frac{\sqrt{3}a}{4}$ away, by covalent bonds. Si, Ge and Sn have the same structure as diamond.



Zinc sulphide, ZnS, has a similar structure, except the base pair is ZnS rather than two identical carbon atoms,



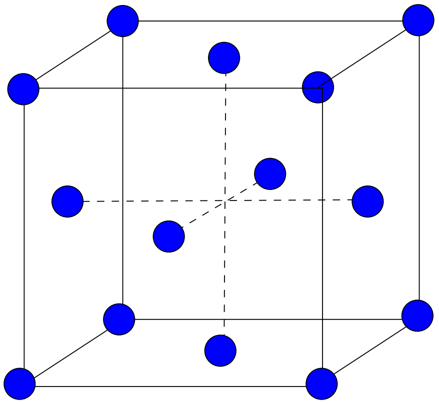
Carbon 60 (buckyballs) has also been found to crystallise in face centred cubic form — in this case the basis consists of sixty carbon atoms!



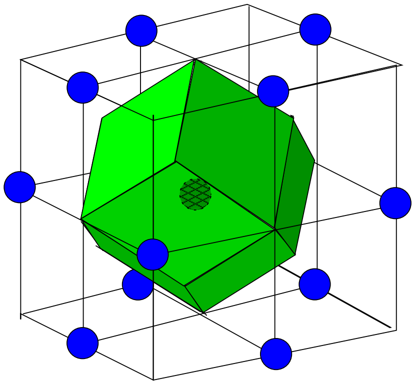
The Wigner-Seitz cell for a face centred cubic lattice is a truncated rhombic dodecahedron:

FCC Lattice Cell

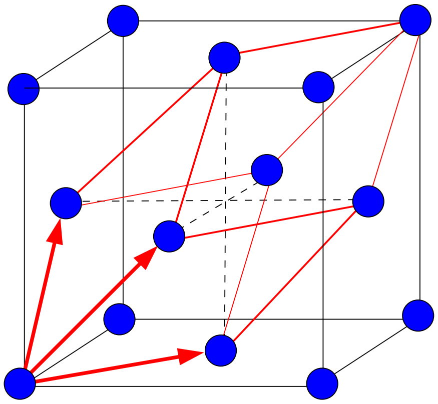
Conventional Cell



Wigner–Seitz Cell



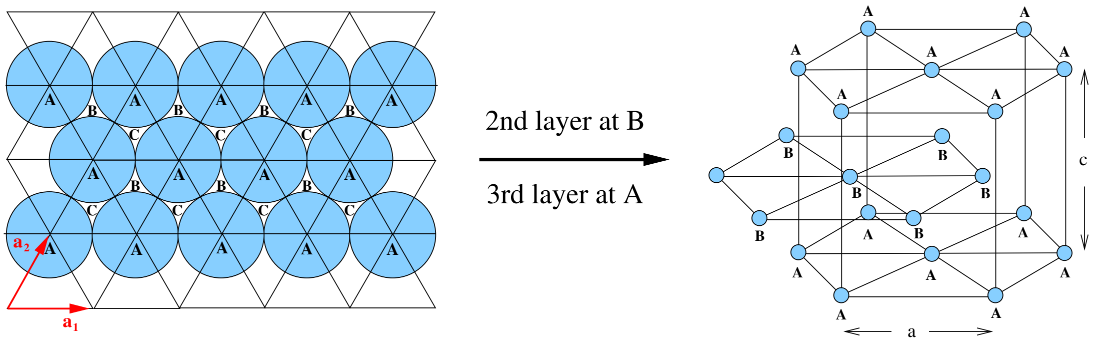
Primitive Cell



4. Hexagonal close packed structure

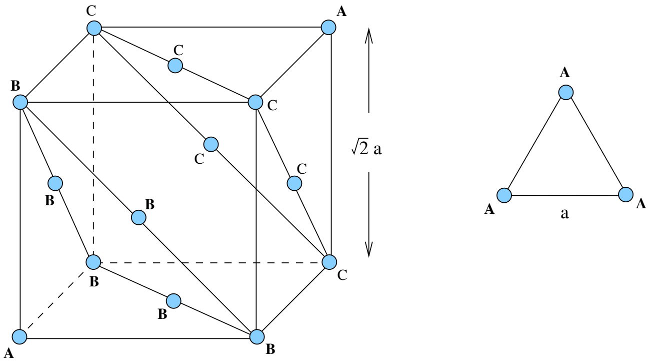
Strictly speaking this is not a Bravais lattice, but it is nevertheless a useful structure to consider as it not infrequently occurs in Nature, *eg.* Mg, Ti, Zn. The hexagonal close packed structure consists of two interwoven 3-dimensional hexagonal lattices and, like diamond, it is really a Bravais lattice (3-d hexagonal) with a basis consisting of two identical atoms. It is constructed by stacking 2-dimesnional hexagonal lattices on top of each other in the sequence ABAB... as shown in the upper figure below:

Hexagonal Close Packed Structure



2–d hexagonal lattice
array of spheres, radii a/2

ABABA....



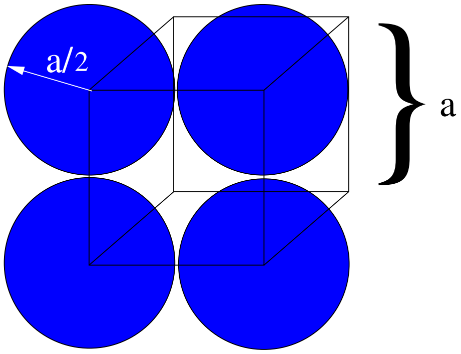
ABCABC...

For optimal close packing with identical spheres $c = \sqrt{\frac{8}{3}}a$. Magnesium for example crystallises in a hexagonal close packed structure with $a = 3.21$Å and $c = 5.21$Å, giving $\frac{c}{a} = 1.62$.

Different sequences of stacking hexagonal lattices give different structures. For example, as shown in the lower picture above, ABCABC... is equivalent to face centred cubic. Other sequences are possible, *e.g* ABACABAC... for some rare earth metals.

Filling fractions

Solids have higher densities than liquids or gases, because their atoms are closely packed. For example we can calculate the fraction of space filled by a spherical monatomic basis in a simple cubic crystal. For a cell size a the basis atoms just touch if their radius is $\frac{a}{2}$.



Each primitive cell has a volume $V_c = a^3$ and contains one complete sphere with volume $\frac{4\pi}{3}\left(\frac{a}{2}\right)^3 = \frac{\pi}{6}a^3$, so the fraction of space that is filled by solid spheres of radius $\frac{a}{2}$, the **packing fraction** is

$$\frac{V_{Sphere}}{V_c} = \frac{\pi}{6} = 0.524...$$

For some other structures the packing fractions are:

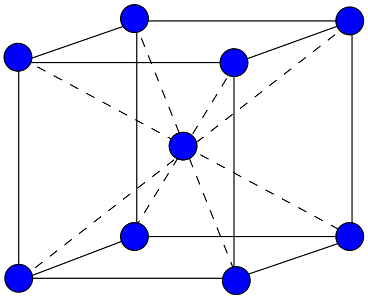
$$\mathbf{FCC}: \frac{\sqrt{2}\pi}{6} = 0.740...$$

$$\mathbf{BCC}: \frac{\sqrt{3}\pi}{8} = 0.68...$$

$$\mathbf{Diamond}: \frac{\sqrt{3}\pi}{16} = 0.34...$$

(the first two are for a monatomic spherical basis).

We finish this section with a couple of observations. First, note that the decomposition `Crystal = Lattice + Basis` is not necessarily unique. For example a body centred cubic lattice with a single monatomic basis (*e.g.* iron) is identical to a simple cubic lattice with a basis consisting of two identical atoms, one at the origin and one at the centre, $\frac{a}{2}(\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}})$,



In the same vein a simple cubic lattice with a monatomic basis is the same as a face centred cubic lattice with a diatomic basis consisting of two identical atoms, one at one corner of a conventional cell and one in the centre.

Secondly we observe that, once the basis is included, the symmetry of the crystal might be smaller than that of the lattice. The list of possible crystal point groups and space groups is larger than those of lattices:

Lattices: 7 point groups; 14 space groups
Crystals: 32 point groups; 230 space groups

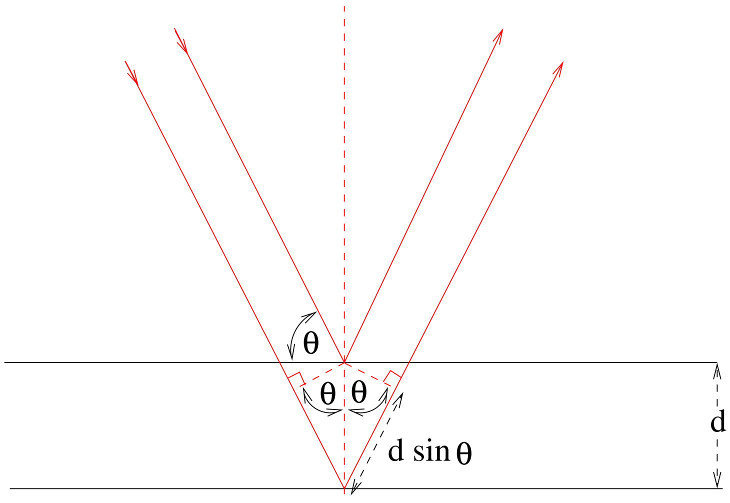
We shall not list all possible crystal space groups here. In four dimensions there are 52 Bravais lattices (different lattice space groups).

3. Reciprocal Lattices

Bragg Law

Experimentally crystal structure can be determined by **diffraction** experiments. Typical atomic separations in a crystal are of the order of 1Å$= 10^{-10}$ m so we need wavelengths of this order to resolve the structure. For electromagnetic radiation this corresponds to X-rays, though we can also use electrons or neutrons whose de Broglie wavelength is $\lambda \approx 1$Å.

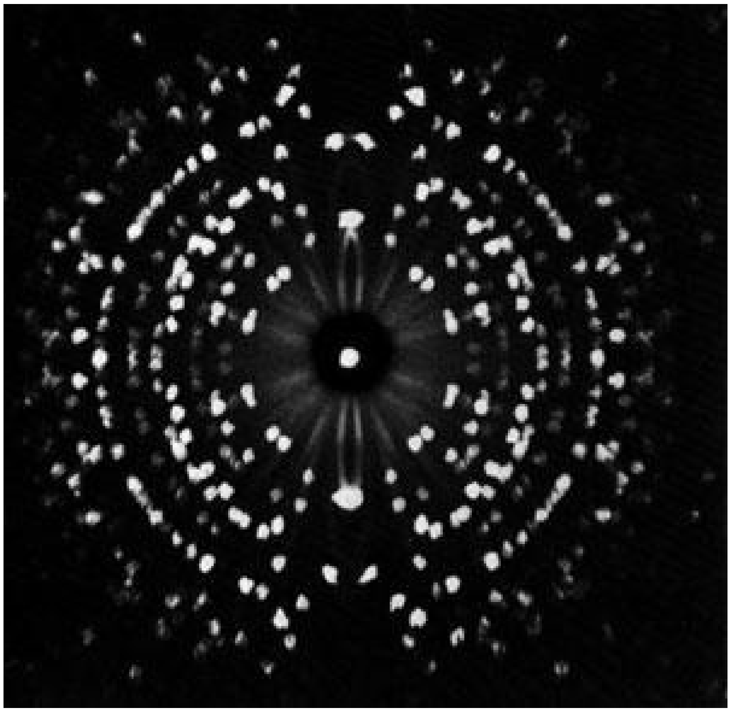
For concreteness let's consider X-rays reflecting off 2-dimensional planes in a crystal. Generically the X-rays experience partial reflection — part of the wave is transmitted and the remainder reflected. The reflected wave can experience interference between lattice planes, either constructive or destructive depending on the angle of incidence. In the figure below there is constructive interference when the path difference between the two waves shown is an integral multiple N of the wavelength,



There is constructive interference when

$$\boxed{2d\sin\theta = N\lambda.} \tag{1}$$

This is known as **Bragg's Law**. Since N is an integer only some specific angles, given by $\sin\theta = \frac{N\lambda}{2d}$, will give strong reflection — angles of incidence that do not satisfy this criterion for any integer N will tend to be transmitted rather than reflected. There will be peaks in intensity, **Bragg peaks**, for special directions such that angle θ satisfies (1) — other directions will receive no scattered X-rays. Bragg peaks manifest themselves as bright spots as seen in this X-ray diffraction pattern for a crystal of Alum (hydrated potassium aluminium sulfate, $\mathrm{K\,Al\,(SO_4)_2.12H_2O}$).



This simple derivation of the Bragg law assumes that X-rays scatter off smooth 2-dimensional planes, like partially transparent mirrors, but in reality they scatter off the electrons in atoms which are localised near points in the plane. To make further progress we need a more realistic mathematical model of the diffraction process. First we define a **lattice plane**.

Lattice planes and Miller indices

A lattice plane is a two-dimensional plane passing through any three non-colinear points of a three-dimensional lattice. Due to periodicity of the original lattice a lattice plane always contains an infinite number of points. A lattice plane is in fact always one of the five two-dimensional Bravais lattices.

For example consider an orthorhombic lattice with primitive lattice vectors $\boldsymbol{a}_1 = a\,\hat{\mathbf{x}}$, $\boldsymbol{a}_2 = b\,\hat{\mathbf{y}}$ and $\boldsymbol{a}_3 = c\,\hat{\mathbf{z}}$. A general lattice point can be represented by the lattice vector

$$\mathbf{L} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3 = n_1 a\,\hat{\mathbf{x}} + n_2 b\,\hat{\mathbf{y}} + n_3 c\,\hat{\mathbf{z}},$$

with n_1, n_2 and n_3 three integers. So $\mathbf{L}$ has Cartesian co-ordinates $x = n_1 a$, $y = n_2 b$ and $z = n_3 c$.

A linear relation between x, y and z defines a plane, *e.g.*

$$\frac{h}{a}x + \frac{k}{b}y + \frac{l}{c}z = p, \tag{2}$$

with h, k, l and p fixed constants. If we allow p to vary, equation (2) defines a family of parallel planes. So, if (x, y, z) is a lattice point, the constraint

$$hn_1 + kn_2 + ln_3 = p \tag{3}$$

defines a family of parallel planes, one for each value of p (the plane with $p = 0$ contains the origin). To describe this family of parallel planes it is sufficient to consider $p = 0$, since we can always choose the origin to lie in any given lattice plane. So we need only consider

$$hn_1 + kn_2 + ln_3 = 0. \tag{4}$$

For an infinite number of solutions to this equation, (n_1, n_2, n_3), which are not co-linear, h, k and l must be rational numbers, and we can always multiply (4) by the least common multiple of their denominators to make them integers — so we can choose h, k and l to be integers without any loss of generality. The smallest three integers (h, k, l) that define a family of parallel lattice planes are called **Miller indices**.

Note:

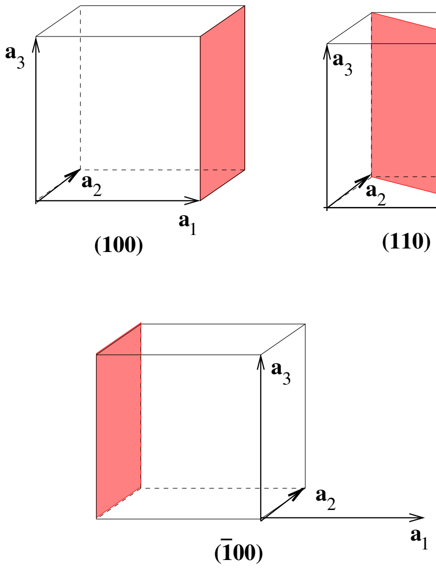
i) If a lattice plane is parallel to one of the primitive lattice vectors then the corresponding co-efficient in (3) is infinity and the Miller index is 0.

ii) When there is no possibility of confusion, commas are omitted from the triple (h, k, l) and (hkl) denotes either a single lattice plane or the set of equally spaced parallel planes, one for each value of p.

iii) By convention the Miller indices associated with a negative co-efficient in (3) is indicated with a bar above it, e.g. $(hk\bar{l})$.

iv) Another convention is that square brackets, $[hkl]$, denotes the direction normal to the plane (hkl). For simple cubic lattices $[hkl]$ is in the same direction as some lattice vector $\mathbf{L}$, but this is not the case for all of the Bravais lattices.

Examples of Miller indices for lattice planes in a simple cubic lattice



Reciprocal Lattice

The above simple derivation of Bragg's law ignores the periodic structure of the lattice planes and we have to be more sophisticated in order to understand fully the kind of X-ray diffraction pattern shown above. X-rays scatter elastically off electrons in the atoms that make up the crystal. Denote the density of electrons at a point $\mathbf{r}$ by $\rho(\mathbf{r})$ (with dimensions of $1/length^3$). Since the crystal is periodic $\rho(\mathbf{r})$ should be a periodic function, $\rho(\mathbf{r}+\mathbf{L}) = \rho(\mathbf{r})$ for any lattice vector $\mathbf{L}$. Since $\rho(\mathbf{r})$ is periodic we can write it as a three-dimensional Fourier series.

As a warm-up exercise, first consider the simple case of a one dimensional monatomic lattice, *i.e.* a line of periodically spaced atoms, each a distance a from its nearest neighbours on either side, so the one dimensional electron density is a periodic function of its argument x,

$$\rho(x) = \rho(x+a).$$

Periodic functions can be expanded as a Fourier series

$$\rho(x) = \rho_0 + \sum_{m=1}^{\infty} A_m \cos\left(\frac{2\pi m x}{a}\right) + \sum_{m=1}^{\infty} B_m \sin\left(\frac{2\pi m x}{a}\right).$$

$$\rho_0 = \frac{1}{a}\int_0^a \rho(x)dx$$

is just the average density over a single period and the co-efficients A_m and B_m can be calculated from $\rho(x)$ in the standard way

$$A_m = \frac{2}{a}\int_0^a \cos\left(\frac{2\pi m x}{a}\right)\rho(x)dx$$

$$B_m = \frac{2}{a}\int_0^a \sin\left(\frac{2\pi m x}{a}\right)\rho(x)dx.$$

It will be convenient to re-express the Fourier series as a sum of complex exponentials,

$$\rho(x) = \sum_{m=-\infty}^{\infty} \rho_m e^{\frac{2\pi i m x}{a}},$$

where the Fourier co-efficients $A_m = \rho_m + \rho_{-m}$ and $B_m = i(\rho_m - \rho_{-m})$ for $m \geq 1$ are real numbers. The Fourier co-efficients in exponential form, ρ_m and ρ_{-m}, are complex in general but must satisfy $\rho_m^* = \rho_{-m}$ since $\rho(x)$ is real. In fact $\rho_m = \frac{1}{2}A_m + \frac{1}{2i}B_m$ and $\rho_{-m} = \frac{1}{2}A_m - \frac{1}{2i}B_m$ for $m \geq 1$. The co-efficients ρ_m are obtained from

$$\rho_m = \frac{1}{a}\int_0^a \rho(x)e^{-\frac{2\pi i m x}{a}}dx$$

for all integral m.

We seek a similar decomposition for all of the three dimensional Bravais lattices. Consider first a simple cubic lattice, with lattice spacing a. This is very like three copies of the one dimensional lattice and we can write

$$\rho(\mathbf{r}) = \sum_{m_1=-\infty}^{\infty}\sum_{m_2=-\infty}^{\infty}\sum_{m_3=-\infty}^{\infty} \rho_{m_1,m_2,m_3} e^{\frac{2\pi i m_1 x}{a}} e^{\frac{2\pi i m_2 y}{a}} e^{\frac{2\pi i m_3 z}{a}} \tag{5}$$

The only subtlety is that this *cannot* be written as

$$\left(\sum_{m_1=-\infty}^{\infty} \rho_{m_1} e^{\frac{2\pi i m_1 x}{a}}\right)\left(\sum_{m_2=-\infty}^{\infty} \rho_{m_2} e^{\frac{2\pi i m_2 y}{a}}\right)\left(\sum_{m_3=-\infty}^{\infty} \rho_{m_3} e^{\frac{2\pi i m_3 z}{a}}\right)$$

because there is no reason to assume that ρ_{m_1,m_2,m_3} can be factorised into $\rho_{m_1}\rho_{m_2}\rho_{m_3}$, and in general it cannot. Equation (5) can be written more compactly as

$$\rho(\mathbf{r}) = \sum_{\{m_1,m_2,m_3\}} \rho_{m_1,m_2,m_3} e^{\frac{2\pi i \mathbf{m}.\mathbf{r}}{a}} = \sum_{\mathbf{G}} \rho_{\mathbf{G}} e^{i\mathbf{G}.\mathbf{r}}$$

where $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$,

$$\mathbf{G} = \frac{2\pi}{a}\big(m_1\hat{\mathbf{x}} + m_2\hat{\mathbf{y}} + m_3\hat{\mathbf{z}}\big),$$

and the sum means the sum over all integer triples (m_1, m_2, m_3).

We can write a similar decomposition for $\rho(\mathbf{r})$ for any three dimensional Bravais lattice

$$\rho(\mathbf{r}) = \sum_{\mathbf{G}} \rho_{\mathbf{G}} e^{i\mathbf{G}.\mathbf{r}}, \tag{6}$$

where $\rho_{\mathbf{G}}$ are independent of $\mathbf{r}$ and the sum is over all vectors $\mathbf{G}$ for which

$$\rho(\mathbf{r}) = \rho(\mathbf{r}+\mathbf{L}) \quad \Rightarrow \quad \sum_{\mathbf{G}} \rho_{\mathbf{G}} e^{i\mathbf{G}.\mathbf{r}} = \sum_{\mathbf{G}} \rho_{\mathbf{G}} e^{i\mathbf{G}.(\mathbf{r}+\mathbf{L})} \tag{7}$$

for any lattice vector $\mathbf{L}$.

As for one-dimensional Fourier transforms the Fourier co-efficients $\rho_{\mathbf{G}}$ are derivable from the original electron density function $\rho(\mathbf{r})$

$$\rho_{\mathbf{G}} = \frac{1}{V_c} \int_{\substack{Primitive\\ Cell}} \rho(\mathbf{r}) e^{-i\mathbf{G}.\mathbf{r}} dV.$$

The set of all allowed $\mathbf{G}$'s satisfying (7) can be found as follows: define three vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ and $\boldsymbol{b}_3$ in terms of primitive lattice vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$

$$\boldsymbol{b}_1 = 2\pi \frac{\boldsymbol{a}_2 \times \boldsymbol{a}_3}{\boldsymbol{a}_1.(\boldsymbol{a}_2 \times \boldsymbol{a}_3)}, \qquad \boldsymbol{b}_2 = 2\pi \frac{\boldsymbol{a}_3 \times \boldsymbol{a}_1}{\boldsymbol{a}_1.(\boldsymbol{a}_2 \times \boldsymbol{a}_3)}, \qquad \text{and} \qquad \boldsymbol{b}_3 = 2\pi \frac{\boldsymbol{a}_1 \times \boldsymbol{a}_2}{\boldsymbol{a}_1.(\boldsymbol{a}_2 \times \boldsymbol{a}_3)}. \tag{8}$$

With this definition it is automatic that

$$\boldsymbol{b}_i.\boldsymbol{a}_j = 2\pi\delta_{ij}$$

where δ_{ij} is the Kronecker δ, equal to 1 if $i = j$ and zero otherwise. Then, for any three integers m_1, m_2 and m_3,

$$\mathbf{G} = m_1\boldsymbol{b}_1 + m_2\boldsymbol{b}_2 + m_3\boldsymbol{b}_3 \tag{9}$$

satisfies

$$e^{i\mathbf{G}.\mathbf{L}} = e^{2\pi i(n_1m_1+n_2m_2+n_3m_3)} = 1 \tag{10}$$

for any lattice vector $\mathbf{L} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3$, so (7) is automatic.

The set of all vectors $\mathbf{G}$ satisfying (9) itself constitutes a lattice, called the **reciprocal lattice**, with primitive lattice vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{b}_3$.

For a 2-dimensional lattice, just set $\mathbf{a}_3 = \hat{\mathbf{z}}$ and use

$$\boldsymbol{b}_1 = \frac{2\pi(\boldsymbol{a}_2 \times \hat{\mathbf{z}})}{|\boldsymbol{a}_1 \times \boldsymbol{a}_2|}, \qquad \boldsymbol{b}_2 = -\frac{2\pi(\boldsymbol{a}_1 \times \hat{\mathbf{z}})}{|\boldsymbol{a}_1 \times \boldsymbol{a}_2|}.$$

Examples:

i) **Simple Cubic:** primitive lattice vectors,

$$\boldsymbol{a}_1 = a\hat{\mathbf{x}}, \qquad \boldsymbol{a}_2 = a\hat{\mathbf{y}}, \qquad \boldsymbol{a}_3 = a\hat{\mathbf{z}};$$

the reciprocal lattice has primitive lattice vectors

$$\boldsymbol{b}_1 = \frac{2\pi}{a}\hat{\mathbf{x}}, \qquad \boldsymbol{b}_2 = \frac{2\pi}{a}\hat{\mathbf{y}}, \qquad \boldsymbol{b}_3 = \frac{2\pi}{a}\hat{\mathbf{z}}.$$

It is a simple cubic lattice with lattice spacing $\frac{2\pi}{a}$.

ii) **FCC:** conventional cell size a, primitive cell volume $V_c = \frac{a^3}{4}$,

$$\begin{aligned}
\boldsymbol{a}_1 &= \frac{a}{2}(\hat{\mathbf{y}}+\hat{\mathbf{z}}), & \boldsymbol{b}_1 &= \frac{2\pi}{(a^3/4)}\left(\frac{a}{2}\right)^2\Big((\hat{\mathbf{x}}\times\hat{\mathbf{y}})+(\hat{\mathbf{z}}\times\hat{\mathbf{x}})+(\hat{\mathbf{z}}\times\hat{\mathbf{y}})\Big) = \frac{2\pi}{a}(-\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}});\\
\boldsymbol{a}_2 &= \frac{a}{2}(\hat{\mathbf{z}}+\hat{\mathbf{x}}), & \boldsymbol{b}_2 &= \frac{2\pi}{a}(\hat{\mathbf{x}}-\hat{\mathbf{y}}+\hat{\mathbf{z}});\\
\boldsymbol{a}_3 &= \frac{a}{2}(\hat{\mathbf{x}}+\hat{\mathbf{y}}), & \boldsymbol{b}_3 &= \frac{2\pi}{a}(\hat{\mathbf{x}}+\hat{\mathbf{y}}-\hat{\mathbf{z}}).
\end{aligned}$$

The reciprocal lattice is body centred cubic, with conventional cell lattice spacing $\frac{4\pi}{a}$.

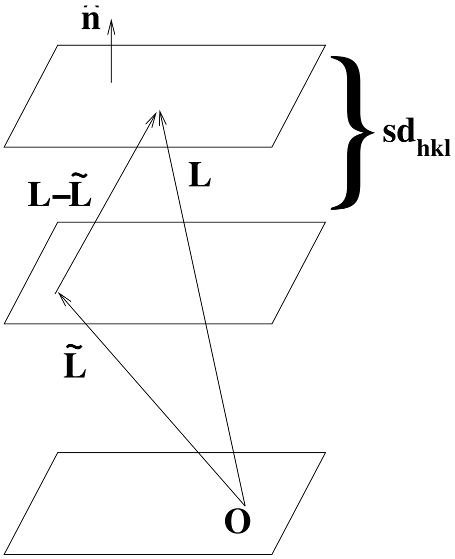
iii) **BCC:** with conventional cell size a the reciprocal lattice is face centred cubic with conventional cell size $\frac{4\pi}{a}$ (the proof is left as an exercise).

When necessary the original lattice will be referred to as the **direct lattice**, to distinguish it from the reciprocal lattice.

Suppose we have a family of lattice planes, (hkl), with minimal separation d_{hkl}. If $\mathbf{L}$ is a lattice vector in one plane and $\widetilde{\mathbf{L}}$ a lattice vector in another plane, a distance $s\, d_{khl}$ away from the first (with s any positive integer), then

$$(\mathbf{L}-\widetilde{\mathbf{L}}).\hat{\mathbf{n}} = sd_{hkl}$$

where $\hat{\mathbf{n}}$ is a unit normal to the planes.



This implies that

$$e^{\frac{2\pi i}{d_{hkl}}\hat{\mathbf{n}}.(\mathbf{L}-\widetilde{\mathbf{L}})} = e^{2\pi i s} = 1$$

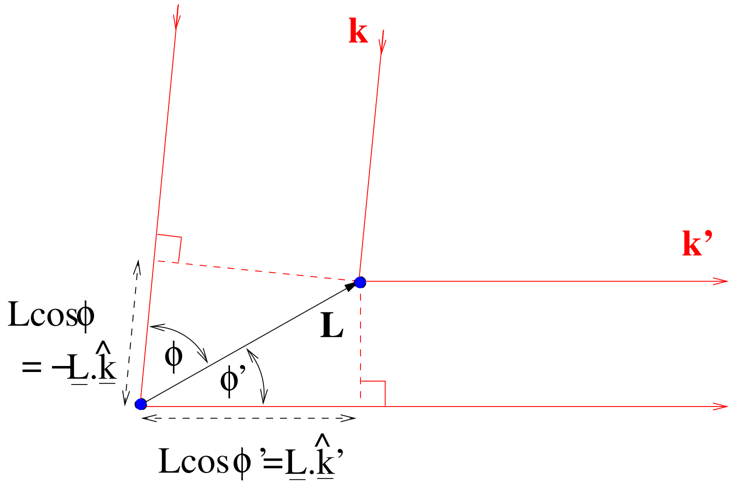
for all $\mathbf{L} - \widetilde{\mathbf{L}}$ (by varying s, $\mathbf{L}$ and $\widetilde{\mathbf{L}}$ this will include all direct lattice vectors). From the definition (10) this in turn implies that $\mathbf{G} = \frac{2\pi}{d_{hkl}}\hat{\mathbf{n}}$ is a reciprocal lattice vector. It is in fact the shortest reciprocal lattice vector that is normal to the (hkl) planes, hence

$$\mathbf{G}_{hkl} = \frac{2\pi}{d_{hkl}}\hat{\mathbf{n}}$$

has length $\frac{2\pi}{d_{hkl}}$, where d_{hkl} is the distance between neighboring planes among the (hkl) set of planes.

Von Laue condition

We can now derive a more powerful version of the Bragg condition, called the Von Laue condition, which takes into account the fact that lattice planes are collections of lattice points. Consider a beam of X-rays scattering elastically off identical atoms sitting at two lattice points separated by a lattice vector $\mathbf{L}$. Elastic scattering means that the energy, and hence wavelength λ, of the X-rays does not change, only their direction changes. If the incoming beam has wavevector $\mathbf{k} = |\mathbf{k}|\hat{\mathbf{k}}$ in the $\hat{\mathbf{k}}$ direction and the outgoing beam has $\mathbf{k}' = |\mathbf{k}|'\hat{\mathbf{k}}'$ in the $\hat{\mathbf{k}}'$ direction then $|\mathbf{k}| = |\mathbf{k}'| = \frac{2\pi}{\lambda}$ ($\hat{\mathbf{k}}$ and $\hat{\mathbf{k}}'$ are unit vectors in the directions $\mathbf{k}$ and $\mathbf{k}'$).



From the diagram above the path difference between two X-rays scattering off the two atoms is $\mathbf{L}.(\hat{\mathbf{k}}' - \hat{\mathbf{k}})$. Constructive interference requires

$$\mathbf{L}.(\hat{\mathbf{k}}' - \hat{\mathbf{k}}) = N\lambda$$

where N is an integer. Hence

$$\mathbf{L}.(\mathbf{k}' - \mathbf{k}) = 2\pi N,$$

since $k = k' = \frac{2\pi}{\lambda}$. There will be a huge enhancement in the intensity of the scattered wave if this is true for all lattice vectors $\mathbf{L}$, that is if

$$e^{i\mathbf{L}.(\mathbf{k}-\mathbf{k}')} = 1 \tag{11}$$

for all $\mathbf{L}$, which is equivalent to the statement that

$$\mathbf{G} = \mathbf{k} - \mathbf{k}'$$

is a reciprocal lattice vector, (10). From this follows

$$-\mathbf{k}' = \mathbf{G} - \mathbf{k} \qquad \Rightarrow \qquad |\mathbf{k}'|^2 = \mathbf{G}^2 - 2\mathbf{G}.\mathbf{k} + |\mathbf{k}|^2,$$

giving

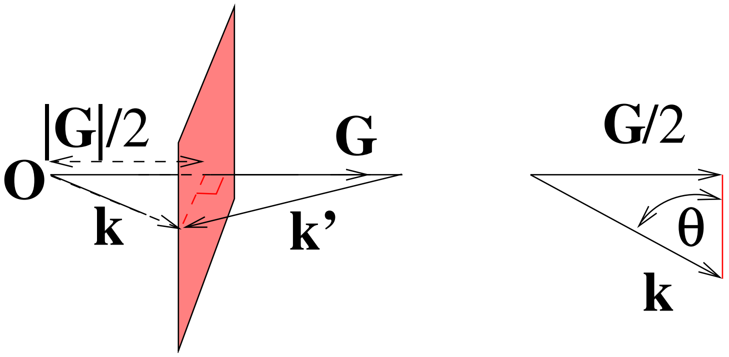
$$\boxed{\mathbf{G}^2 = 2\mathbf{G}.\mathbf{k}} \tag{12}$$

since $|\mathbf{k}'|^2 = |\mathbf{k}|^2$. This is the **von Laue condition**, a scattered X-ray will show a peak in intensity if the incoming wavevector $\mathbf{k}$ satisfies this condition for some reciprocal lattice vector $\mathbf{G}$.

This is related to the Bragg condition (1) as follows. Since $\mathbf{G}$ is a reciprocal lattice vector and it is an integral multiple, $\mathbf{G} = N\mathbf{G}_{hkl}$, of some shortest reciprocal lattice vector, $\mathbf{G}_{hkl}$, for three integers h, k and l. If (hkl) have no common divisor[5] then $\mathbf{G}_{hkl} = \frac{2\pi}{d_{hkl}}\hat{\mathbf{n}}$ has magnitude $|\mathbf{G}_{hkl}| = \frac{2\pi}{d_{hkl}}$ where d_{hkl} is the distance between neighbouring (hkl) lattice planes. The von Laue condition is

$$\frac{|\mathbf{G}|}{2} = \hat{\mathbf{G}}.\mathbf{k} = |\mathbf{k}|\sin\theta$$

where the angle θ is defined in the figure below,



Hence

$$\frac{|\mathbf{G}|}{2} = \frac{\pi N}{d_{hkl}} = |\mathbf{k}|\sin\theta = \frac{2\pi}{\lambda}\sin\theta \qquad \Rightarrow \qquad 2d_{hkl}\sin\theta = N\lambda,$$

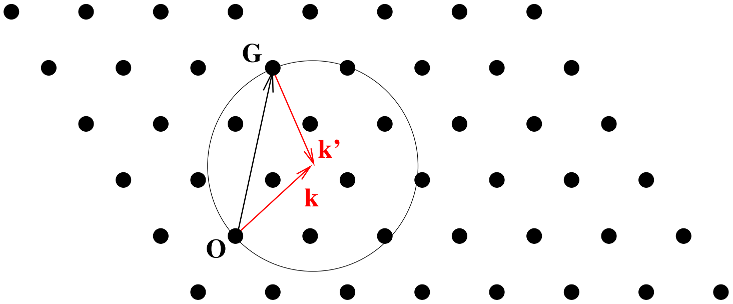
which is the Bragg condition (1) with $d = d_{hkl}$.

[5] The k in (hkl) here is an integer describing reciprocal lattice planes, *not* the wave number of the incoming X-ray!

From the figure above it can be seen that the maximum intensity in the scattered ray is achieved when the tip of the wavevector $\mathbf{k}$ lies in a plane which is the perpendicular bisector of a reciprocal lattice vector $\mathbf{G} = N\mathbf{G}_{hkl}$ for some (hkl) — this called the **Bragg plane** for the incoming wave. Most $\mathbf{k}$ will not lie in a Bragg plane and so will not give peak intensity for the scattered wave.

Ewald construction

A neat way of visualising the von Laue condition is the **Ewald construction**. Choose an origin $\mathbf{O}$ at a point in the reciprocal lattice and place the tail of $\mathbf{k}$ at $\mathbf{O}$. Draw a circle of radius $|\mathbf{k}|$ centred on the tip of $\mathbf{k}$, so it passes through the tail of $\mathbf{k}$. $\mathbf{k}$ will generate a Bragg peak if and only if another reciprocal lattice point $\mathbf{G}$ (other than $\mathbf{O}$) lies on the circle.

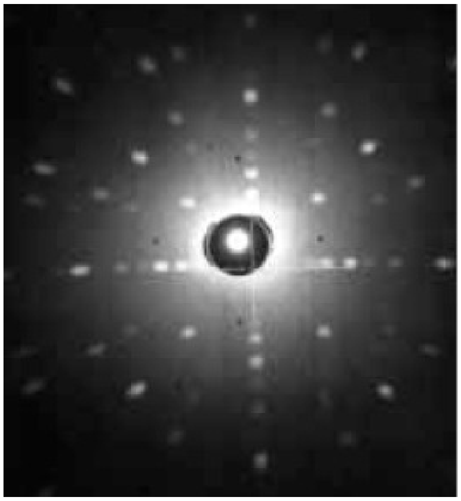


Three common methods of observing diffraction peaks are:

1) **Laue method:** fix the direction of $\mathbf{k}$ relative to the crystal and allow $|\mathbf{k}|$ to vary (*i.e.* vary the wave-length), effectively thickening the circle in the Ewald construction above so that it encompasses some $\mathbf{G}$
2) **rotating crystal method:** fix $\mathbf{k}$ and rotate the crystal, equivalent to rotating the lattice points in the Ewald construction about the origin.
3) **powder method:** use a powder consisting of many small crystals, in random orientations, with $\mathbf{k}$ fixed. There will always be some small crystals with the lattice in the correct orientation to give a peak.

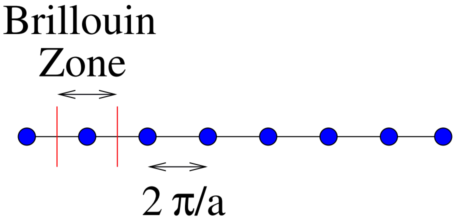
The Ewald construction makes it clear that if $|\mathbf{k}|$ is less than the reciprocal lattice spacing there will be no Bragg peaks, *i.e.* the wavelength is too long. If $|\mathbf{k}|$ is very large compared to the reciprocal lattice spacing, *i.e.* very short wave-lengths, there will be very many such $\mathbf{G}$'s and very many allowed directions $\mathbf{k}'$, when this happens the Bragg peaks wash out and the pattern is lost. A clear pattern is only seen if $|\mathbf{k}|$ is larger than the reciprocal lattice spacing, but not too large, corresponding to wavelengths of the order of the direct lattice spacing which, for most crystals, is of the order of an Å. For electromagnetic radiation this corresponds to X-rays, but electrons or neutrons with velocities momenta corresponding to de Broglie wavelengths of a few Å can also be used.

Below is the X-ray diffraction pattern for diamond, taken using the von Laue method. Note the 4-fold symmetry which reflects the underlying cubic structure of diamond:



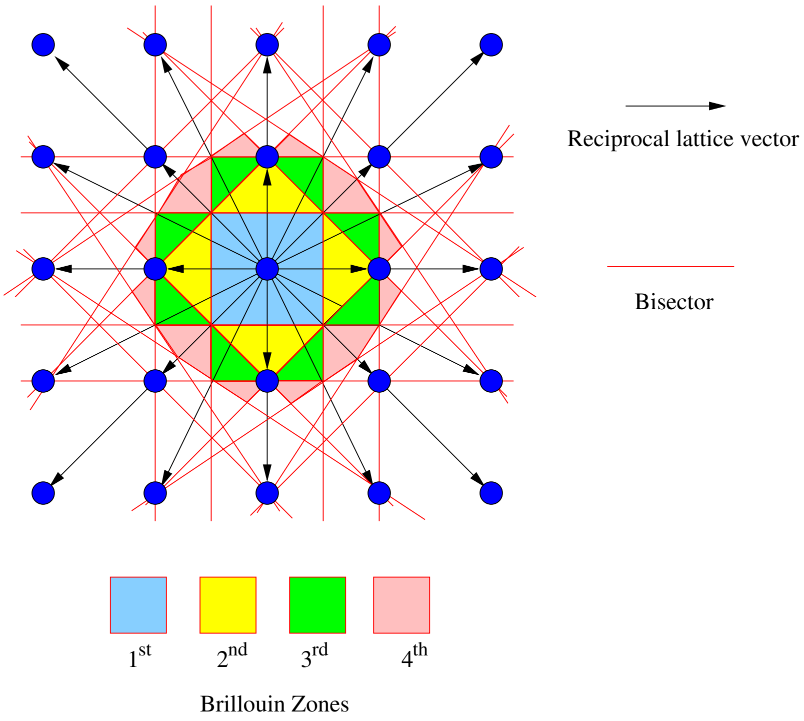
Brillouin zones

A Wigner-Seitz cell of the reciprocal lattice is called a **Brillouin zone**. Brillouin zones are another very useful way of understanding how X-ray diffraction patterns can arise — they will also play a central rôle in understanding crystal vibrations and the movement of electrons through crystals to be studied later. For a one-dimensional lattice for example, with lattice spacing a, the reciprocal lattice has lattice spacing $\frac{2\pi}{a}$ and the region between $-\frac{\pi}{a}$ and $\frac{\pi}{a}$ is a Brillouin zone.



In two or three dimensions the von Laue condition requires that the tip of $\mathbf{k}$, the wavevector of the incoming X-ray, lie on a plane which is the perpendicular bisector of a reciprocal lattice vector $\mathbf{G}$. Consider first a 2-dimensional square lattice, with primitive lattice vectors $\boldsymbol{a}_1 = a\hat{\mathbf{x}}$ and $\boldsymbol{a}_2 = a\hat{\mathbf{y}}$. The reciprocal lattice is also square, with primitive lattice vectors $\boldsymbol{b}_1 = \frac{2\pi}{a}\hat{\mathbf{x}}$ and $\boldsymbol{b}_2 = \frac{2\pi}{a}\hat{\mathbf{y}}$ and reciprocal lattice vectors have the form $\mathbf{G} = m_1\boldsymbol{b}_1 + m_2\boldsymbol{b}_2$, with m_1 and m_2 integers. In the figure below the blue square is bounded by four red lines, each is a perpendicular bisector of a reciprocal lattice vector. The four lattice vectors that are used to construct the blue square are $\pm\boldsymbol{b}_1$ and $\pm\boldsymbol{b}_2$ ($m_1 = \pm 1$, $m_2 = \pm 1$). The blue square is a Wigner-Seitz cell for the reciprocal lattice and is called the **first Brillouin zone**. An incoming X-ray whose wavevector $\mathbf{k}$ has its tail on the central reciprocal lattice point and its head anywhere on the boundary of the blue square will generate a Bragg peak.

Brillouin Zones for Square Lattice



The yellow triangles are bounded on the outside by perpendicular bisectors of the four reciprocal lattice vectors

$$\mathbf{G} = \boldsymbol{b}_1 + \boldsymbol{b}_2, \qquad \mathbf{G} = \boldsymbol{b}_1 - \boldsymbol{b}_2, \qquad \mathbf{G} = -\boldsymbol{b}_1 + \boldsymbol{b}_2, \qquad \mathbf{G} = -\boldsymbol{b}_1 - \boldsymbol{b}_2,$$

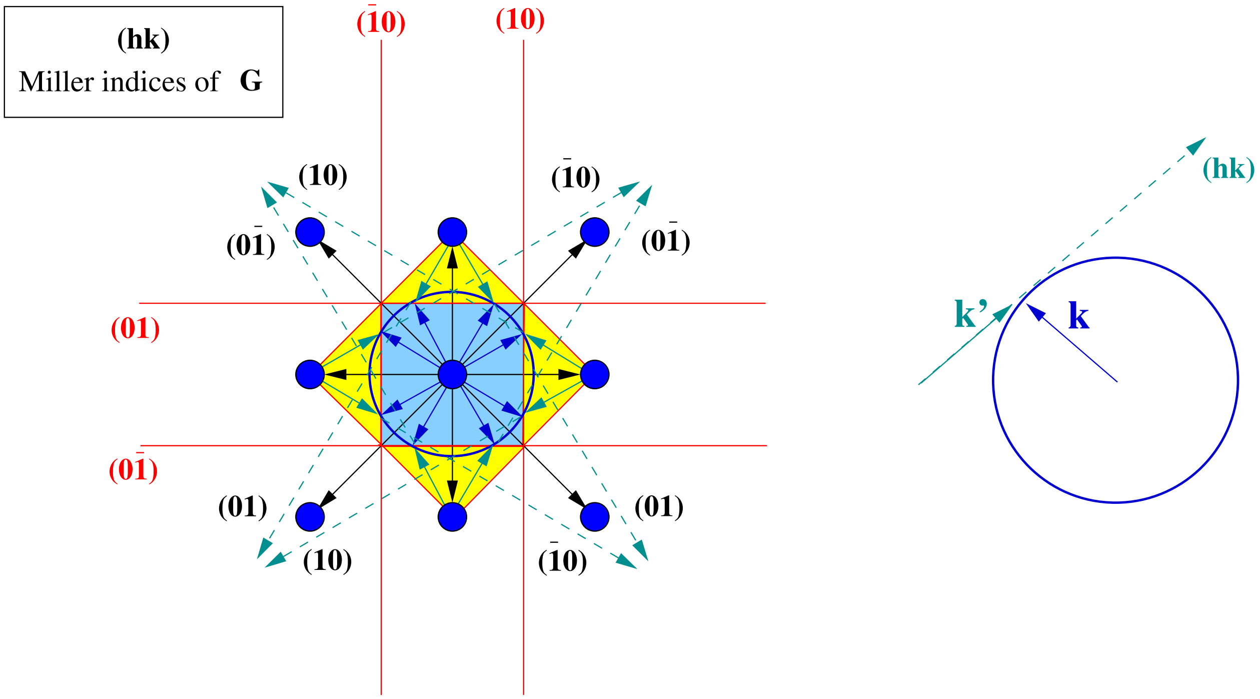
and on the inside by the first Brillouin zone, the blue square. They can be pieced together to make a yellow square which is also a Wigner-Seitz cell of the reciprocal lattice, identical in size and shape to the first Brillouin zone. This cell is called the **second Brillouin zone**. An incoming X-ray whose wavevector **k** has its tail on the central reciprocal lattice point and its head anywhere on the boundary of the yellow triangles will generate a Bragg peak.

The green triangles can be pieced together to make a square identical to the blue one — this is the **third Brillouin zone** (can you work out which reciprocal vectors are bisected by the red boundaries?). The pink shapes constitute the fourth Brillouin zone, and so on.

Blue arrows in the figure below give examples of **k**-directions that generate Bragg peaks from the boundary of the first Brillouin zone. The tip if the wavevector is rotated to give the blue circle, only the specific directions where this circle intersects the boundary of a Brillouin zone (red lines) corresponds to an incident direction that gives a Bragg peak. The reflected waves $\mathbf{k}'$ are shown in a lighter blue and, for clarity, they have been extended by dotted blue arrows and labelled by the Miller indices of the reciprocal vector that is bisected by the relevant red line.

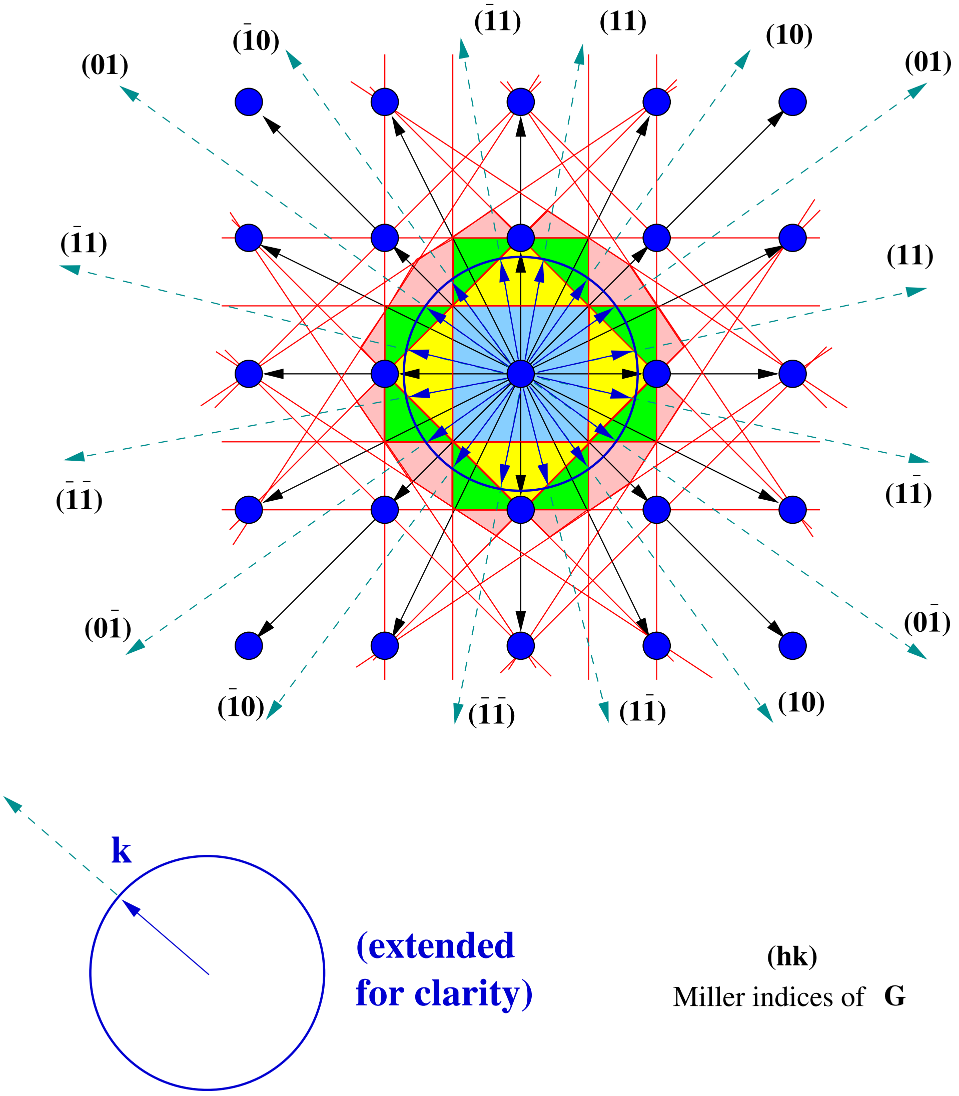
Bragg peak directions for square lattice

for fixed $|\mathbf{k}|$ extending into 2nd Brillouin zone

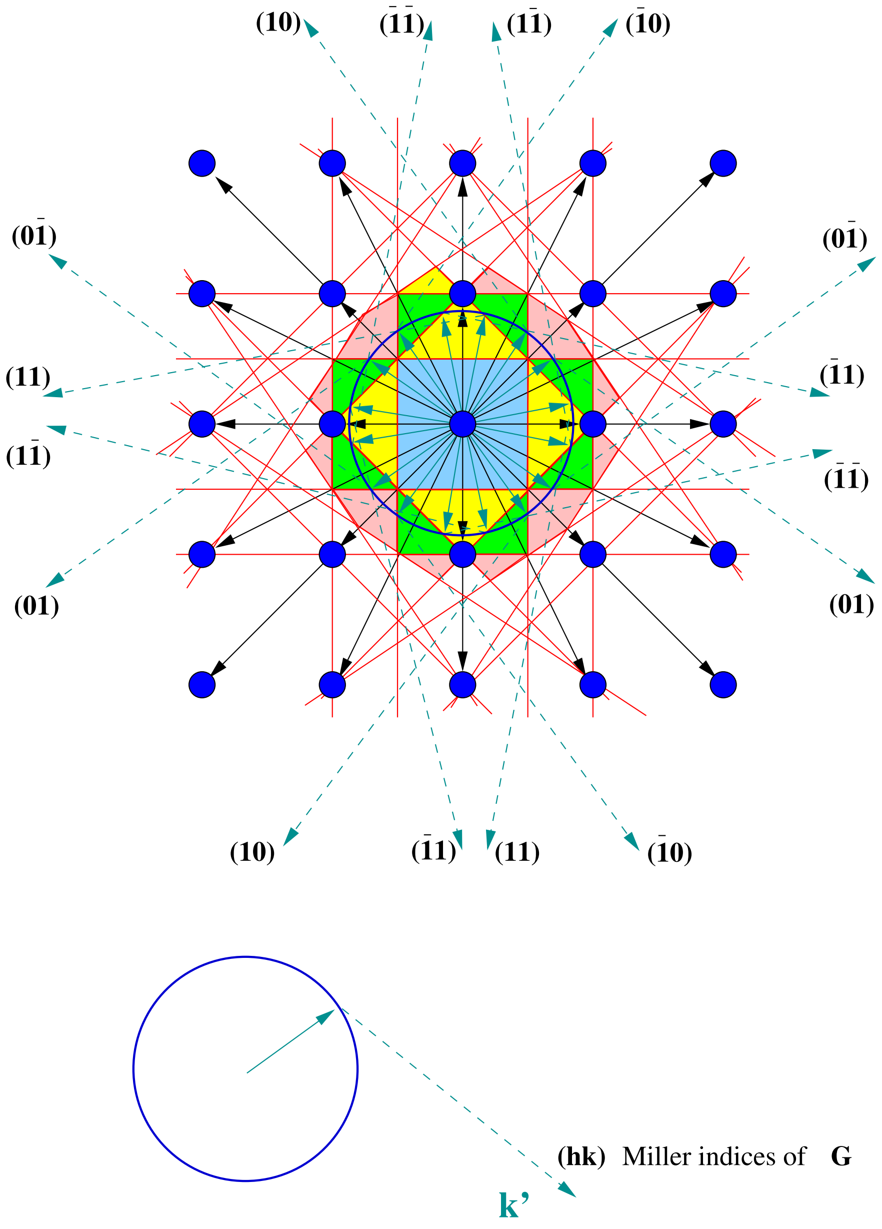


Shorter wavelengths (longer **k**) can scatter off more Brillouin zones: the following figure shows incident directions that give Bragg peaks by scattering off second and even third Brillouin zone boundaries. The second figure below shows the direction of the outgoing (scattered) wave for the same length of **k**.

Bragg peak k−directions for square lattice



Bragg peak k'−directions for square lattice



In summary, a Bragg peak is present if and only if the tip of **k** lies on the boundary of a Brillouin zone in the above construction.

Structure factors

So far we have assumed that lattice sites, and only lattice sites, act as point scatterers. Representing the scattered wave by a complex number (the physical wave is the real part) each lattice site $\mathbf{L}$ contributes $e^{i(\mathbf{k}-\mathbf{k}').\mathbf{L}}$ to the scattered wave, so the total scattered amplitude is proportional to[6] $\sum_{\mathbf{L}} e^{i(\mathbf{k}-\mathbf{k}').\mathbf{L}}$. If $\mathbf{k}-\mathbf{k}' = \mathbf{G}$ is a reciprocal lattice vector then $e^{i(\mathbf{k}-\mathbf{k}').\mathbf{L}} = 1$ for every lattice site and every term in the sum adds coherently. If $\mathbf{k}-\mathbf{k}'$ is not a reciprocal lattice vector every term in the sum has a different phase and they combine destructively to give a total of zero.

For a crystal with anything other that a monatomic basis the true story is a little more complicated. Electromagnetic waves scatter predominantly off electrons (electrons react to an incoming wave much more readily than positive ion cores, as they are much lighter and more responsive). Denote the electron density $\rho(\mathbf{r})$ then, in general, the scattered amplitude is proportional to

$$F(\mathbf{k}-\mathbf{k}') := \int dV \rho(\mathbf{r}) e^{i(\mathbf{k}-\mathbf{k}').\mathbf{r}},$$

where the integral is over the volume of the crystal.

For a monatomic crystal the electron density resides only at lattice sites and we can write

$$\rho(\mathbf{r}) = n_0 \delta(\mathbf{r}-\mathbf{L})$$

where n_0 is the number of electrons in the atom free to respond to the incoming wave and $-e$ is the charge on an electron, but ρ will be more complicated than this for a more general crystal type. With the von Laue condition, $\mathbf{k}-\mathbf{k}' = \mathbf{G}$, we have

$$F = \int_{Crystal} dV \rho(\mathbf{r}) e^{i\mathbf{G}.\mathbf{r}} = \mathcal{N}_c \int_{Cell} dV \rho(\mathbf{r}) e^{i\mathbf{G}.\mathbf{r}},$$

where $\mathcal{N}_c$ is the number of cells in the crystal. The integral over a single cell,

$$S_{\mathbf{G}} = \int_{Cell} dV \rho(\mathbf{r}) e^{i\mathbf{G}.\mathbf{r}},$$

is called the **structure factor** — a dimensionless number, in general complex.

If the basis consists of s atoms at points $\mathbf{r}_j$ in the unit cell, where $j = 1, \dots, s$, and $\rho_j(\mathbf{r})$ is the electron density of the j-th atom then

$$S_{\mathbf{G}} = \sum_{j=1}^{s} \int_{Cell} dV \rho_j(\mathbf{r}) e^{i\mathbf{G}.\mathbf{r}} = \sum_{j=1}^{s} e^{i\mathbf{G}.\mathbf{r}_j} \int_{Cell} dV \rho_j(\mathbf{r}) e^{i\mathbf{G}.(\mathbf{r}-\mathbf{r}_j)} = \sum_{j=1}^{s} f_j e^{i\mathbf{G}.\mathbf{r}_j},$$

where

$$f_j := \int_{Cell} dV \rho_j(\mathbf{r}) e^{i\mathbf{G}.(\mathbf{r}-\mathbf{r}_j)}$$

[6] Remember (11), $e^{i(\mathbf{k}-\mathbf{k}').\mathbf{L}}$=1 for constructive interference — if it is not unity for all $\mathbf{L}$ different lattice points will give different complex phases and the sum will be zero.

is called the **atomic structure factor**. To a good approximation f_j is independent of $\mathbf{r}_j$ and $\mathbf{G}$, since we expect $\rho_j(\mathbf{r})$ to be strongly localised about $\mathbf{r} = \mathbf{r}_j$, $\rho_j(\mathbf{r}) \approx n_j \delta(\mathbf{r} - \mathbf{r}_j) \Rightarrow f_j \approx n_j$ where n_j is the number of electrons in atoms j that are free to respond to the incoming X-ray.

Example 1: Caesium Chloride has a simple cubic structure with a basis consisting of two atoms ($s = 2$), which we take to be a Caesium atom at $\mathbf{r}_1 = 0$ and a Chlorine atom at $\mathbf{r}_2 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$, using a conventional cell basis $\boldsymbol{a}_1 = a\hat{\mathbf{x}}$, $\boldsymbol{a}_2 = a\hat{\mathbf{y}}$, $\boldsymbol{a}_3 = a\hat{\mathbf{z}}$. Sodium and Chlorine have different electronic structures and we expect them respond differently to X-rays, so $f_1 \neq f_2$. The reciprocal lattice is also cubic, with

$$\mathbf{G} = \frac{2\pi}{a}(h\hat{\mathbf{x}} + k\hat{\mathbf{y}} + l\hat{\mathbf{z}}).$$

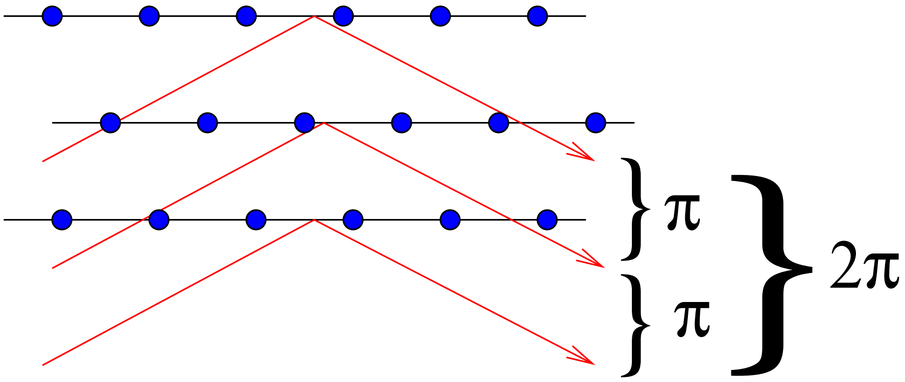
This gives

$$S_{hkl} = f_1 + e^{i\frac{a}{2}(\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}}).\mathbf{G}} f_2 = f_1 + e^{i\pi(h+k+l)} f_2 = \begin{cases} f_1 - f_2 & \text{for } h+k+l \text{ odd;} \\ f_1 + f_2 & \text{for } h+k+l \text{ even.} \end{cases}$$

Indeed experimentally reflections from (200) and (110) planes are stronger than from (100) and (300) planes.

Example 2: Sodium has a BCC structure with a monatomic basis, but we can also think of this a simple cubic structure with a diatomic basis, $s = 2$, consisting of identical atoms of sodium at $\mathbf{r}_1 = 0$ and at $\mathbf{r}_2 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$. This is similar to CsCl, but now $f_1 = f_2$ and we expect all Bragg peaks corresponding to $h + k + l$ odd to be completely absent, and indeed this is the case.

The absence of $h+k+l$ odd planes for BCC crystals can be understood intuitively in a simple two-dimensional example with adjacent lines of atoms off-set from one another,



When the phase of the wave reflected from adjacent layers differ by π they interfere destructively, but then the next to adjacent layers must necessarily differ in phase by 2π and interfere constructively.

Diffraction experiments on crystals require wavelengths of a few Å corresponding to X-rays for electromagnetic radiation, but we can also use electrons or neutrons with de

Broglie wavelength of similar size. For X-rays the scatters are electrons in the crystal, but for neutrons it is the atomic nuclei that cause scattering while for electrons it is the combined electrostatic potential of the crystal electrons plus the positively charged atomic nuclei that cause scattering. We therefore get different information about the crystal from X-rays, neutron and electron scattering.

4. Crystal Binding

The way in which atoms are bound together to form crystals depends in detail on inter-atomic forces between the atoms making up the crystal. We shall discuss two cases in some depth: inert elements and ionic crystals, but you should bear in mind that there are other cases, such as covalent bonding, that will not be covered in this course.

Inert elements: (*i.e.* noble gases: Ne, Ar, Kr, Xe). These gases tend to form face centred cubic crystals when they solidify, with a monatomic basis. (The physics of solid Helium is very different and will not be covered here.)

To understand their structure we model the force between two atoms separated by $\mathbf{r}$ using the **Lennard-Jones** potential,

$$U(\mathbf{r}) = \frac{B}{r^{12}} - \frac{A}{r^6} = 4\varepsilon \left\{ \left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right\},$$

where A and B are two constants which can be traded for an energy ε and a length σ. The second term above represents an attraction between atoms due to dipole-dipole interactions while the first term is a repulsion due to quantum effects — when the atoms get so close to one another that their outermost electronic orbitals start to overlap the Pauli exclusion principle wants to prevent the electron wave-functions from overlapping too much.

The energy ε and the length σ are characteristics of each element and they can be determined from experiments performed on the gaseous phase, determining the equation of state by measuring virial co-efficients and viscosity,

	Melting Point ($^\circ K$)	$\varepsilon(10^{-23}J)$	σ(Å)
Ne	24	50	2.74
Ar	84	167	3.40
Kr	117	225	3.65
Xe	161	320	3.98

The total binding energy of the crystal is obtained by summing the interactions over all pairs of atoms, remembering to divide by 2 to avoid over-counting. For a crystal with $\mathcal{N}$ atoms,

$$U_{Tot} = \left(\frac{\mathcal{N}}{2}\right) 4\varepsilon \sum_{\mathbf{L}\neq 0} \left\{ \left(\frac{\sigma}{|\mathbf{L}|}\right)^{12} - \left(\frac{\sigma}{|\mathbf{L}|}\right)^{6} \right\}.$$

A stable configuration requires that the crystal is at a minimum of the potential energy. If the lattice spacing is varied then the lattice vectors $\mathbf{L}$ will change length. Let $\widetilde{\mathbf{L}}$ be lattice vectors for a lattice with primitive cells having unit volume. Then a lattice with primitive cells having volume R^3 will have lattice vectors $\mathbf{L} = R\widetilde{\mathbf{L}}$. For a monatomic crystal based on a simple cubic lattice R is the same thing as the inter-atomic spacing, but for other Bravais lattices it is not necessarily exactly the same as the inter-atomic spacing though it will be proportional to it. In any case the potential energy of the whole crystal is

$$U_{Tot} = 2\mathcal{N}\varepsilon \left\{ A_{12}\left(\frac{\sigma}{R}\right)^{12} - A_6\left(\frac{\sigma}{R}\right)^6 \right\}, \tag{13}$$

where

$$A_n := \sum_{\mathbf{L}\neq 0} \frac{1}{|\widetilde{\mathbf{L}}|^n}.$$

Varying R is the same as varying the nearest neighbour separation.

For example in a one-dimensional crystal there is an atom at each lattice site, labelled by an integer k, $\widetilde{\mathbf{L}} = k\hat{\mathbf{x}}$ with $\hat{\mathbf{x}}.\hat{\mathbf{x}} = 1$, and $\mathbf{L} = kR\hat{\mathbf{x}}$ so $|\mathbf{L}| = kR$ and

$$A_n = \sum_{k\neq 0} \frac{1}{k^n} = 2\sum_{k=1}^{\infty} \frac{1}{k^n}.$$

The sum $\zeta(n) = \sum_{k\neq 0} \frac{1}{k^n}$ is known as the Riemann ζ-function, and it can be calculated analytically when n is even, for example $\zeta(12) = \frac{691\pi^{12}}{638512857}$.

For three dimensional crystals the sums will depend on the lattice type and must be carried out numerically. For FCC lattices the results are

$$A_6 = 14.45392\cdots, \qquad A_{12} = 12.13188\cdots$$

(any lattice point in a FCC lattice has 12 nearest neighbours and successive terms in the sum fall off very rapidly, particularly for A_{12} for which by far the greatest contribution to the sum comes from just the nearest neighbors). The equilibrium separation R_0 is obtained by setting

$$\frac{dU_{Tot}}{dR} = 0 \qquad \Rightarrow \qquad -12A_{12}\frac{\sigma^{12}}{R_0^{13}} + 6A_6\frac{\sigma^6}{R_0^7}$$

giving

$$R_0^6 = 2\left(\frac{A_{12}}{A_6}\right)\sigma^6 \qquad \Rightarrow \qquad R_0 = 1.090\sigma.$$

The experimentally measured values of R_0 in real crystals are

	Ne	Ar	Kr	Xe
$\frac{R_0}{\sigma}$	1.14	1.11	1.10	1.09

The increasing discrepancies in Kr, Ar and Ne are due to quantum effects as the outer electron shells are more and more tightly bound in the smaller atoms.

Using $\frac{R_0}{\sigma} = 1.09$ in (13) gives the binding energy per atom in equilibrium

$$\frac{1}{\mathcal{N}} U_{Tot}(R_0) = -8.6\varepsilon.$$

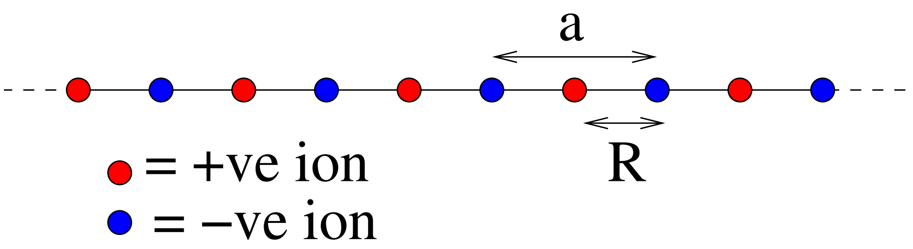
Note that values of ε given above, and hence the binding energy per atom, are proportional to the melting point of the crystals.

Ionic crystals: (*e.g.* NaCl, CsCl, ZnS). Crystals made up of positive and negative ions, such as salt, in a regular array are called *ionic* crystals. The binding force for ionic crystals is due to the Coulomb interaction of the positive and negative charges on the ions. Assuming the atoms are singly ionised the binding energy is obtained from the Coulomb energy between particles of charge $\pm e$ a distance r apart, $\frac{e^2}{4\pi\epsilon_0 r}$. This is a much longer range force than that arising from the Lennard-Jones potential for inert elements. If the separation between nearest neighbour ion pairs of opposite charge is $\mathbf{R}$ and the total number of ion pairs (molecules) is $\mathcal{N}$ then the total electrostatic energy in the crystal is

$$U_{Col} = \frac{e^2\mathcal{N}}{4\pi\epsilon_0}\left\{-\frac{1}{|\mathbf{R}|} + \sum_{\mathbf{L}\neq 0}\left(\frac{1}{|\mathbf{L}|} - \frac{1}{|\mathbf{L}+\mathbf{R}|}\right)\right\}. \tag{14}$$

The sum over $\frac{1}{L}$ comes from like sign ions at each lattice point and is positive because like sign ions repel each another.

For example in a one-dimensional crystal, consisting of a regular line of molecules a distance a apart, the nearest neighbour ionic separation is $R = \frac{a}{2}$,



and

$$\begin{aligned} U_{Col} =& \frac{e^2\mathcal{N}}{4\pi\epsilon_0}\left(\cdots - \frac{1}{3R} + \frac{1}{2R} - \frac{1}{R} - \frac{1}{R} + \frac{1}{2R} - \frac{1}{3R} + \cdots\right) \\ =& \frac{e^2\mathcal{N}}{2\pi\epsilon_0}\left(-\frac{1}{R} + \frac{1}{2R} - \frac{1}{3R} + \cdots\right) = -\frac{e^2\mathcal{N}}{2\pi\epsilon_0 R}\left(1 - \frac{1}{2} + \frac{1}{3} - \cdots\right). \end{aligned}$$

Note that we use $\mathcal{N}$ here, rather than $\frac{\mathcal{N}}{2}$ as for the inert elements, because we are summing over $2\mathcal{N}$ ions and, dividing by one-half to avoid over-counting just reduces this to $\mathcal{N}$.

We need the sum

$$1 - \frac{1}{2} + \frac{1}{3} - \cdots = \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}.$$

This is a convergent series,

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} = \ln 2, \tag{15}$$

as is seen by Taylor expanding[7]

$$\begin{aligned} \ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots \\ \underset{x=1}{\Rightarrow} \qquad \ln 2 &= 1 - \frac{1}{2} + \frac{1}{3} - \cdots . \end{aligned} \tag{16}$$

Thus

$$U_{Col} = -\frac{e^2 \mathcal{N} \alpha}{4\pi\epsilon_0 R}$$

with $\alpha = 2\ln 2 = 1.386294\ldots$.

For a three dimensional crystal the sum over lattice points in (14) must be carried out numerically and the dimensionless number

$$\alpha = -1 + R \sum_{\mathbf{L}\neq 0} \left(\frac{1}{L} - \frac{1}{|\mathbf{L}+\mathbf{R}|} \right) \tag{17}$$

is called the **Madelung constant**. Again it depends on the sequence in which the crystal is put together and it is best to compute it by first assembling small neutral blocks and then putting them together to form the crystal.

The Madelung constant depends on the lattice structure:

Structure	α	Example
SC	1.762675	CsCl
BCC	1.747565	
FCC	1.6381	NaCl

The total energy includes repulsion of the atoms when they get too close to one another, due to the exclusion principle and electron wave-function overlap — this is the same effect as for inert elements. It can be modelled as a $\frac{1}{R^m}$ repulsive potential (for noble gases $m = 12$) but, unlike the inert element case, it is not possible to obtain the form from experiments on the gaseous phase. With this assumption the total potential is

$$U_{Tot} = \mathcal{N} \left(\frac{C}{R^m} - \frac{e^2 \alpha}{4\pi\epsilon_0 R} \right), \tag{18}$$

[7] While (15) is correct, the infinite series is convergent, it's value depends on the order in which it is summed, it is said to be *conditionally convergent*. Physically this means, for an infinite crystal, the Coulomb energy stored in the crystal would depend on how the crystal is assembled — real crystals however are never truly infinite and the sums will always really be finite with unambiguous values.

where C is a positive constant. The equilibrium separation, R_0, is obtained by demanding

$$\left.\frac{\partial U_{Tot}}{\partial R}\right|_{R_0} = 0 \qquad \Rightarrow \qquad -\frac{mC}{R_0^{m+1}} + \frac{e^2\alpha}{4\pi\epsilon_0 R_0^2} = 0,$$

giving

$$R_0^{m-1} = \frac{4\pi\epsilon_0 mC}{e^2\alpha}.$$

Putting this value of R_0 into (18) gives the binding energy per ion par

$$\frac{U_{Tot}(R_0)}{\mathcal{N}} = -\frac{e^2\alpha}{4\pi\epsilon_0}\left(\frac{m-1}{m}\right)\frac{1}{R_0}.$$

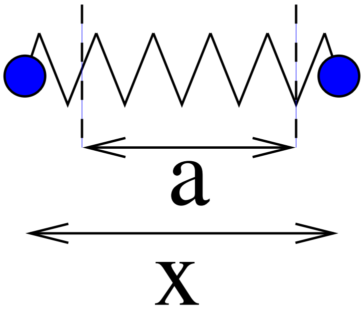
The value of m does not affect the result much, as long as m is large.

5. Crystal Vibrations – Phonons

A real crystal is not a perfect lattice, the atoms and molecules making up the crystal will vibrate about their equilibrium positions. These vibrations will propagate through the crystal at definite speeds, as sound waves. There will also be vibrations due to thermal motion — a warm crystal is continuously humming!

One-dimensional crystal (monatomic basis)

To illustrate the concepts, consider again a one-dimensional monatomic crystal consisting of identical atoms a distance a apart. For small amplitude vibrations we can model the atomic vibrations by thinking of each pair of atoms being linked with a spring with identical spring constant $C > 0$ for each pair, with the spring relaxed when the atoms are a distance a apart. The restoring force on the n-th atom due to the $(n+1)$-th atom on its right is $F = C(x - a)$ (the force is to the right if $x > a$).

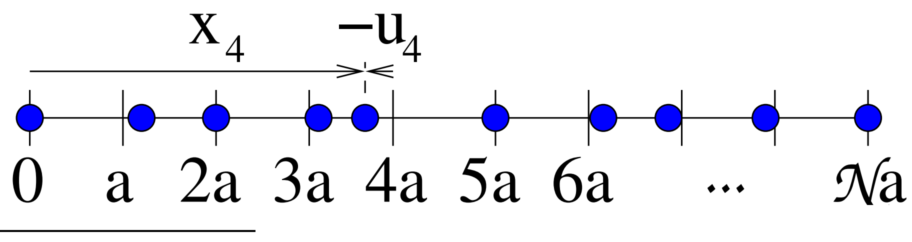


In a chain of such atoms, which are vibrating around their equilibrium positions, denote the position of the n-th atom by x_n. The equilibrium position of the n-th atom is na but when the crystal vibrates $x_n \neq na$ in general. To construct a specific mathematical model we need to specify boundary conditions: we choose[8] $\mathcal{N} + 1$ atoms and fix $x_0 = x_{\mathcal{N}} = 0$. If the atoms are vibrating x_n is a function of time $x_n(t)$. Denote the displacement of the n-th atom from its equilibrium position by $u_n(t)$,

$$u_n(t) = x_n(t) - na,$$

then the total force on the n-th atom is the sum of the forces due to the atoms on either side,

$$F_n = C(x_{n+1} - x_n - a) - C(x_n - x_{n-1} - a) = C(u_{n+1} - 2u_n + u_{n-1}).$$



[8] Alternatively we could use periodic boundary conditions on $\mathcal{N}$ atoms and set $x_0 = x_{\mathcal{N}}$ without specifying its value. For very large $\mathcal{N}$ which boundary conditions we choose makes little difference.

If the mass of each atom is M, then Newton's second law implies

$$M\ddot{u}_n = C(u_{n+1} - 2u_n + u_{n-1}). \tag{19}$$

This gives a set of $\mathcal{N}$ coupled linear ODE's for the $u_n(t)$, which we can solve. The solutions are oscillating. Using a complex notation write

$$u_n(t) = \varepsilon_0 e^{-i(\omega t - Kna)} \tag{20}$$

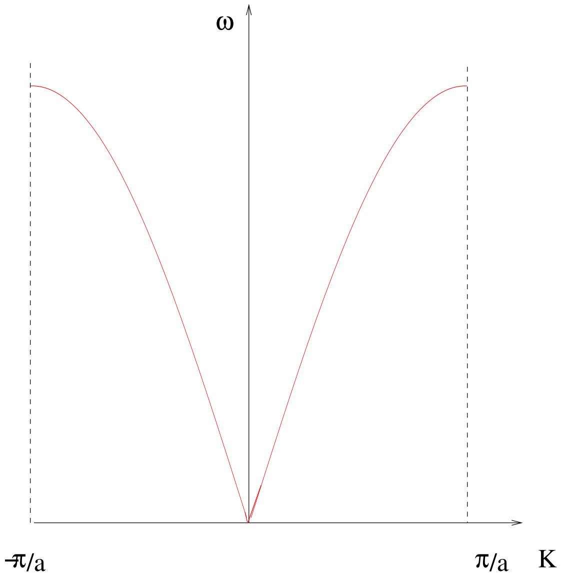
with ω, K and ε_0 constants (the actual displacements are the real part of these complex u_n). ω is an angular frequency, K is a wave-number ($K > 0$ represents waves moving to the right and $K < 0$ waves moving to the left) and ε_0 the amplitude of the displacement. Using this form in (19) gives

$$-\omega^2 M = C(e^{iKa} + e^{-iKa} - 2) = 2C(\cos Ka - 1) = -4C\sin^2\left(\frac{KA}{2}\right).$$

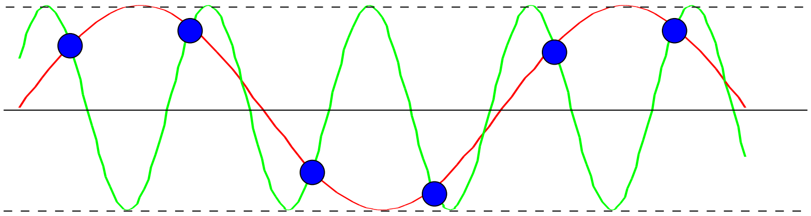
Taking the positive square root we get a relation between ω and K

$$\boxed{\omega = 2\sqrt{\frac{C}{M}}\sin\left|\frac{Ka}{2}\right|.} \tag{21}$$

Since the wavenumber $|K| = \frac{2\pi}{\lambda}$ is related to the wavelength λ equation (21) relates the frequency to the wavelength $\omega(K)$ — it is an example of a *dispersion relation*.



Since in (20) $\frac{u_{n+1}(t)}{u_n(t)} = e^{iKa} = e^{i\left(K+\frac{2p\pi}{a}\right))a}$, for any integer p, we need only consider K in the range $-\frac{\pi}{a} < K \leq \frac{\pi}{a}$, or equivalently $\lambda = \frac{2\pi}{|K|} \geq 2a$, wavelengths with $\lambda < 2a$ are meaningless! This can be visualised using the figure below.



For ease of visualisation the displacements u_n at one instant of time are represented vertically here and the horizontal displacement represents the equilibrium position of the atoms, na. The green curve has a wavelength one-third of the red curve, but the red curve is perfectly adequate for representing the displacements, there is nothing to be gained by considering the shorter wavelength.

The range of wavevectors $|K| \le \frac{\pi}{a}$ is precisely the First Brillouin zone of the one-dimensional crystal. For $\mathcal{N}$ large, but still finite, we can decompose a general vibration of the crystal into a linear superposition of normal modes. With periodic boundary conditions, $u_0 = u_{\mathcal{N}} \Rightarrow e^{iK\mathcal{N}a} = 1$ and so we must have $K\mathcal{N} = \frac{2\pi p}{a}$ with p an integer. So $K = \frac{2p}{\mathcal{N}}\left(\frac{\pi}{a}\right)$ and $-\frac{\pi}{a} \le K \le \frac{\pi}{a} \quad \Rightarrow \quad p = \pm 1, \pm 2, \cdots, \frac{\mathcal{N}}{2}$. There is a finite number, $\mathcal{N}$, of modes ($p = 0$ corresponds to a rigid translation of the whole crystal and is uninteresting). In other words the allowed values of K,

$$K = \pm\frac{2\pi}{\mathcal{N}a}, \pm\frac{4\pi}{\mathcal{N}a}, \pm\frac{6\pi}{\mathcal{N}a}, \cdots, \pm\frac{\pi}{a},$$

are *discrete* for $\mathcal{N}$ finite — we get a continuum of K-values only in the $\mathcal{N} \to \infty$ limit.

Note that:

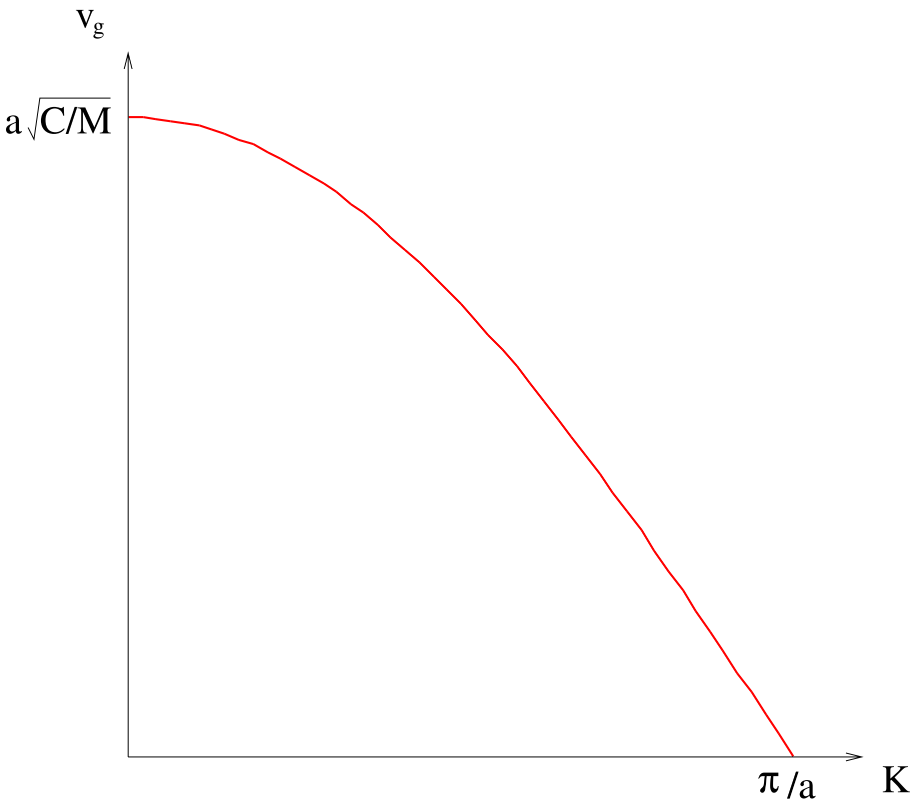
- For K small and positive, $0 < K << \frac{\pi}{a}$, (21) gives $\omega \approx \sqrt{\frac{C}{M}}Ka$ leading to a linear relation between frequency and wavelength with velocity

$$v_p = \frac{\omega}{K} = \sqrt{\frac{C}{M}}\,a.$$

  The larger the spring constant, C, *i.e.* the stiffer the crystal, the greater the speed of propagation of sound waves.

- More generally, away from small K, the velocity depends on the wavelength. A wavepacket made up of a combination of different wavelengths will tend to disperse because long wavelengths (small K) move faster than shorter wavelengths (with K near $\pm\frac{\pi}{a}$). Waves move with *group velocity*

$$v_g = \frac{d\omega}{dK} = \sqrt{\frac{C}{M}}\,a\cos\left(\frac{Ka}{2}\right).$$



- For small K, $v_g \approx v_p$ and the group velocity is the same as v_p,[9] the dispersion relation is linear.
- For $K = \pm\frac{\pi}{a}$ the group velocity $v_g = 0$: we have **standing waves**. The displacements of neighbouring atoms are exactly out of phase

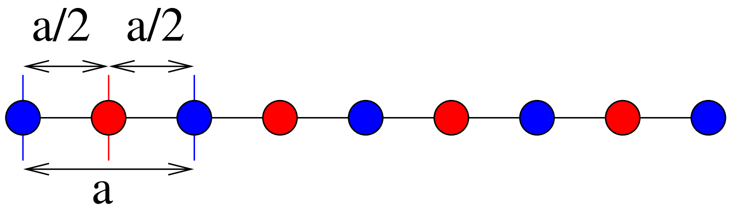
$$\frac{u_{n+1}(t)}{u_n(t)} = e^{i\pi} = -1.$$

Sound waves with these wavelengths are reflected off the Brillouin zone boundary.

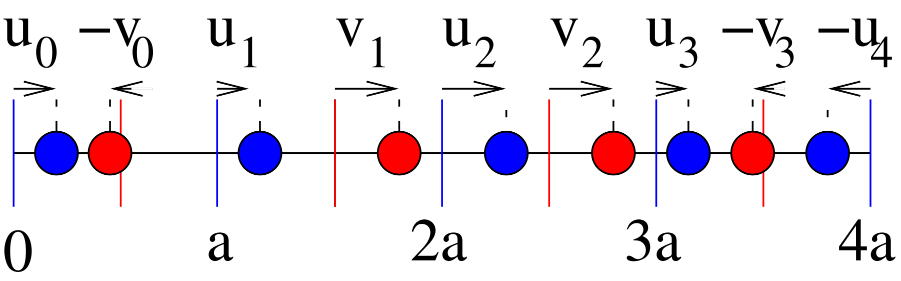
One-dimensional crystal (diatomic basis)

For a basis consisting of two atoms (*e.g.* positive and negative ions in an ionic crystal) with different masses M_1 and M_2 there are further interesting phenomena. Again take the lattice spacing to be a and suppose that the equilibrium separation between M_1 and M_2 atoms is $\frac{a}{2}$ (in the picture below M_1 atoms are blue and M_2 atoms are red).

[9] $v_p = \frac{\omega}{K}$, for any K, is called the *phase* velocity. An observer moving with speed v_p would see a constant phase in the atomic displacements — this is not necessarily a physical velocity. In most situations energy, and other physical quantities, are transported with the group velocity.



Denote the displacements of the n-th M_1 atom from equilibrium by u_n and that of the n-th M_2 atom by v_n. For small displacements we can model the forces as springs between nearest neighbour atoms and, for simplicity, we shall assume that the spring constants are all the same, C. In the picture below the vertical lines represent the equilibrium positions,



Then Newton's equations are

$$\begin{aligned} M_1 \ddot{u}_n &= C(v_n - u_n) - C(u_n - v_{n-1}) = C(v_n + v_{n-1} - 2u_n) \\ M_2 \ddot{v}_n &= C(u_{n+1} - v_n) - C(v_n - u_n) = C(u_{n+1} + u_n - 2v_n). \end{aligned}$$

Looking for a (complex) solution of the form

$$\begin{aligned} u_n(t) &= \varepsilon_1 e^{i(Kna - \omega t)} \\ v_n(t) &= \varepsilon_2 e^{i(Kna - \omega t)} \end{aligned} \qquad \Rightarrow \qquad \begin{aligned} -M_1\omega^2\varepsilon_1 &= C\big((1+e^{-iKa})\varepsilon_2 - 2\varepsilon_1\big) \\ -M_2\omega^2\varepsilon_2 &= C\big((e^{iKa}+1)\varepsilon_1 - 2\varepsilon_2\big) \end{aligned}$$

(again the physical displacements are the real parts of the complex $u_n(t)$ and $v_n(t)$.) This can be written in matrix form

$$\begin{pmatrix} M_1\omega^2 - 2C & C(1+e^{-iKa}) \\ C(1+e^{iKa}) & M_2\omega^2 - 2C \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} = 0.$$

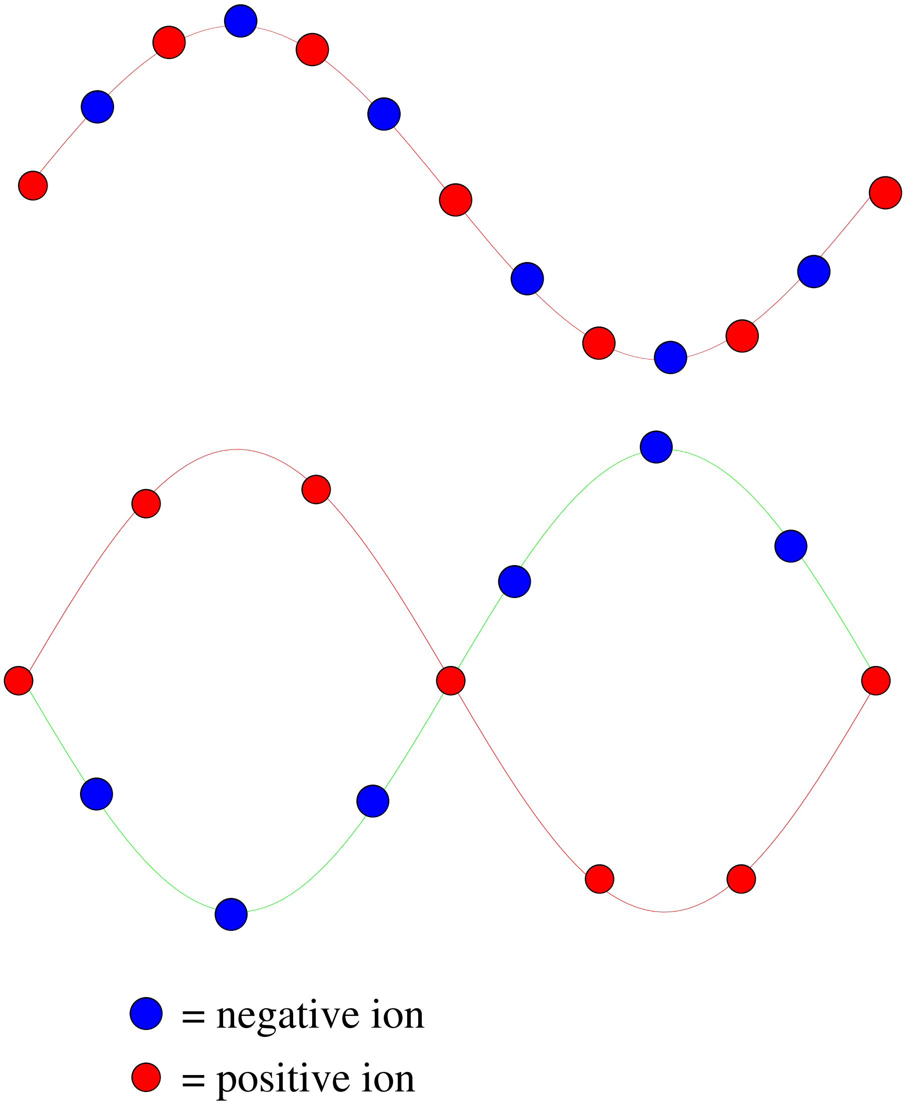
If the matrix is invertible the only solution is $\varepsilon_1 = \varepsilon_2 = 0$, a solution with ε_1 and ε_2 not both zero only exists if the matrix is not invertible *i.e.* the determinant is zero. This requires

$$M_1 M_2 \omega^4 - 2C(M_1+M_2)\omega^2 + C^2(2 - e^{iKa} - e^{-iKa}) = 0,$$

or

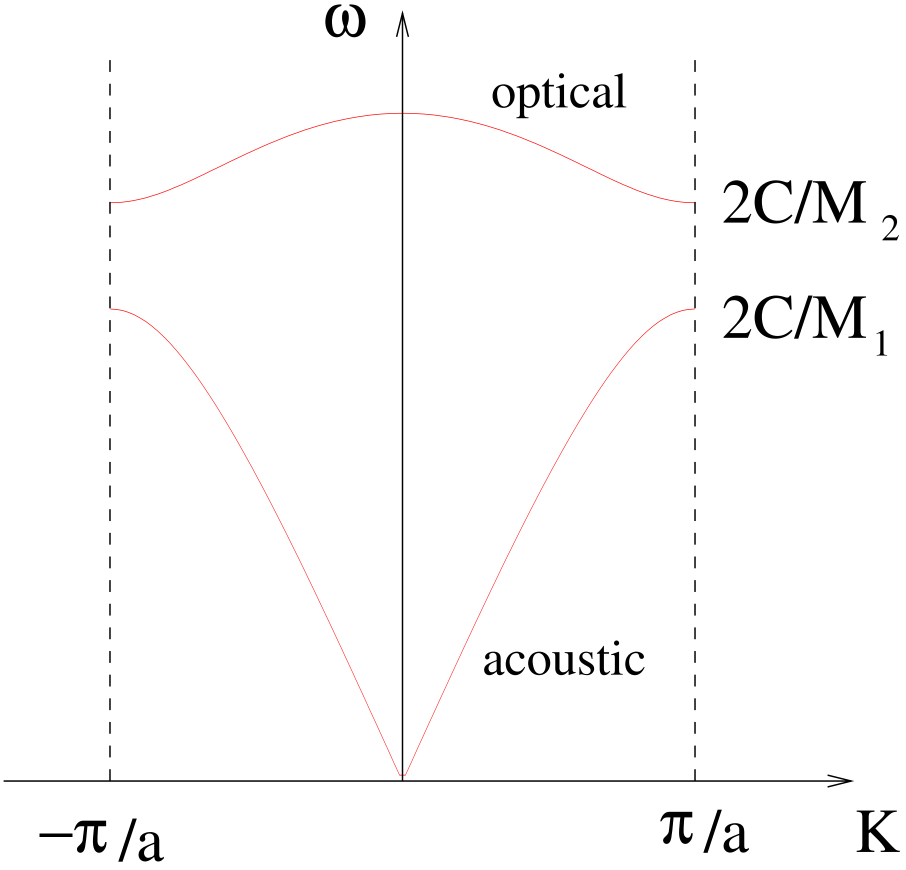
$$\omega^2 = \frac{C(M_1+M_2) \pm C\sqrt{(M_1+M_2)^2 - 4M_1M_2\sin^2\left(\frac{Ka}{2}\right)}}{M_1M_2}.$$

We see that there are now *two* different frequencies for each value of $-\frac{\pi}{a} \leq K \leq \frac{\pi}{a}$, corresponding to two different vibrational modes for each K. The lower sign (lower frequency) requires $\varepsilon_1 = \varepsilon_2$, so M_1 and M_2 are oscillating in phase, while the upper sign (higher frequency) requires $\varepsilon_1 = -\varepsilon_2$, so M_1 and M_2 are oscillating exactly out of phase — while M_1 is displaced to the left the adjacent M_2 is displaced to the right. These two possibilities are shown below, where the M_1 atoms are red and the M_2 atoms are blue (again, for clarity, the displacements u_n and v_n are represented vertically and the equilibrium positions, na and $\left(n + \frac{1}{2}\right) a$, horizontally)



Experimentally the different modes can be preferentially excited in an ionic crystal if M_1 are positive ions and M_2 are negative ions. Then a passing electromagnetic wave will push the positive and negative ions in different directions, because they are pushed in opposite directions by an electric field. However an acoustic vibration (hit the crystal with a hammer!) does not distinguish between positive and negative ions, they are both pushed in the same direction by a passing acoustic wave. For a given K the lower frequency mode is called the *acoustic* mode, because it can be excited by a passing sound wave through the crystal, while the upper frequency is called the *optical* mode, because it can be excited by a passing electromagnetic wave (light) through the crystal.

The dispersion relation, shown below, has two *branches*, an acoustic branch and an optical branch.

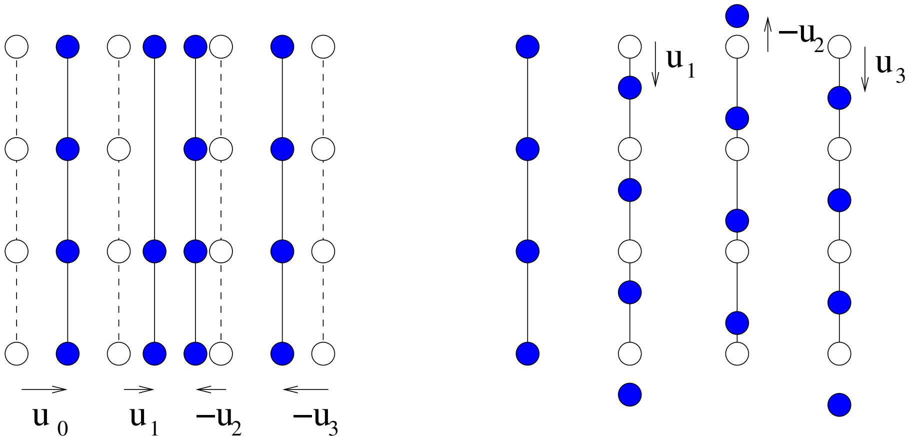


For K small, $0 < K << \frac{\pi}{a}$, $\sin^2\left(\frac{Ka}{2}\right) \approx \frac{K^2a^2}{4}$ and

$$\omega^2 = \begin{cases} \frac{2(M_2+M_2)C}{M_2M_2} - \frac{C(Ka)^2}{2(M_1+M_2)} + \cdots & \text{Optical branch } (\frac{\varepsilon_1}{\varepsilon_2} = -1); \\ \frac{C(Ka)^2}{2(M_1+M_2)} + \cdots & \text{Acoustic branch } (\frac{\varepsilon_1}{\varepsilon_2} = 1). \end{cases}$$

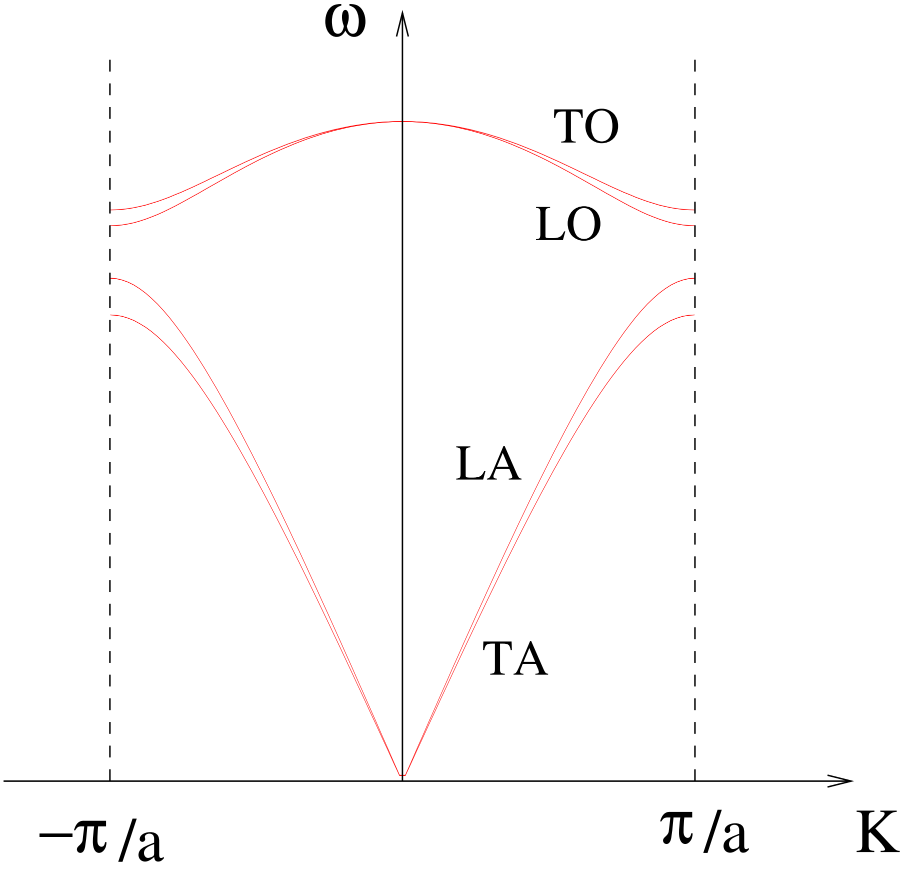
For the optical branch ω^2 is a maximum at $K = 0$, so $v_g = 0$ there, and the dispersion relation looks like an inverted parabola for small K, while the acoustic branch has a linear dispersion relation, $\omega \approx \sqrt{\frac{C}{2(M_1+M_2)}}\,aK$ and $v_g = v_p = \sqrt{\frac{C}{2(M_1+M_2)}}\,a$.

In two dimensions there are even more possibilities. For a monatomic basis, when there is only one mode in one dimension, there are two different modes in two dimensions, the atoms can be displaced in the *same* direction as the wavevector **K** as shown on the left in the picture below (a *longitudinal* mode) or at right-angles to the wavevector as shown on the right in the picture below (*transverse* mode).

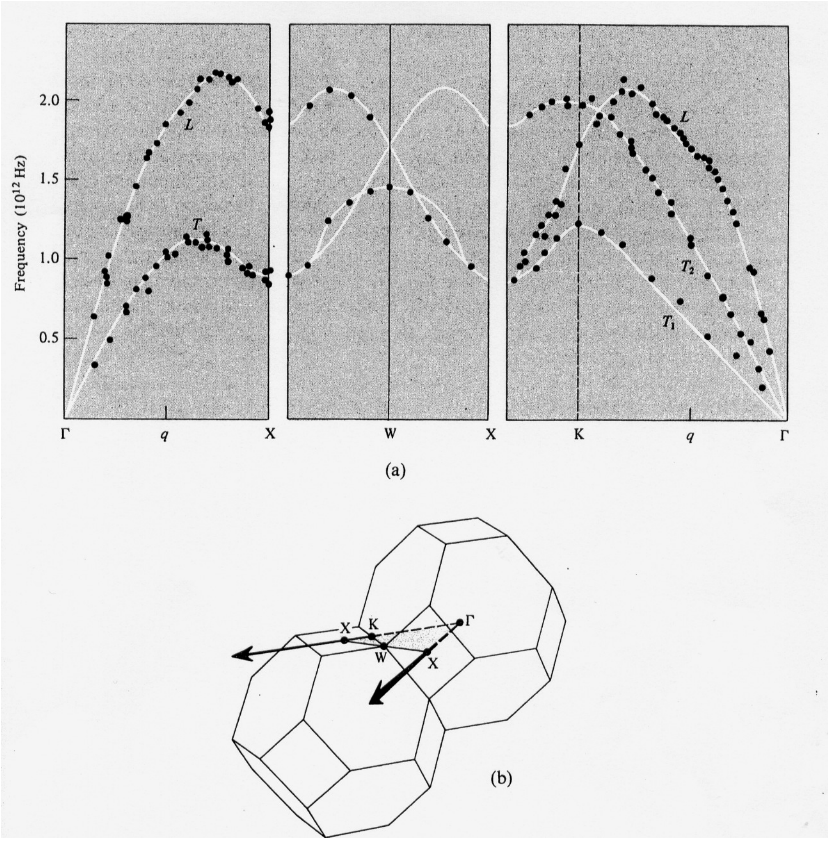


The amplitude ε_0 in one-dimension becomes a vector, $\underline{\varepsilon}_0$, in two-dimensions, with $\mathbf{K}$ parallel to $\underline{\varepsilon}_0$ in the longitudinal case and $\mathbf{K}.\underline{\varepsilon}_0 = 0$ in the transverse case. If the crystal is anisotropic and the spring constants are different in different directions, the dispersion relation will be different for the longitudinal and transverse modes.

For a diatomic 2-dimensional crystal there can be up to four modes: longitudinal optical (LO), transverse optical (TO), longitudinal acoustic (LA) and transverse acoustic (TA), each with a different dispersion relation



In three dimensions there can be two different transverse optical and transverse acoustic modes for each frequency, giving six different modes: one LO, two TO, one LA and two TA. The dispersion relation can become very complicated as it can be different for different directions $[hkl]$. For example the dispersion relations measured experimentally in lead (FCC), in various crystal directions, are shown below



Data are shown for wavevectors in three different directions as indicated in the Wigner-Seitz cell of the reciprocal lattice in (b) (lead has a face centred cubic structure so the reciprocal lattice is body centred cubic and the Wigner-Seitz cell is a truncated octahedron). Γ marks the centre of the Wigner-Seitz and $\mathbf{K}$ traces out a triangle with sides $\Gamma - K - X$, $X - W - X$ and $X - \Gamma$. Lead is not an ionic crystal and only acoustic modes appear in the upper panel. On the line $\Gamma - X$ there is only one transverse acoustic branch and this bifurcates into two on $X - W - X$, which combine again into a single branch at the second X but bifurcates again before reaching K. The direction $\Gamma - X$ is [100] and $\Gamma - K$ is [110].

Quantisation

To understand fully the nature of crystal vibrations it is necessary to take quantum mechanical effects into account. In quantum mechanics a classical wave can sometimes best be described by particles in the quantum theory. A quantum of crystal vibration is called a **phonon** — a particle of sound.

The vibrations of the crystal atoms or molecules about their equilibrium positions can be modelled using a harmonic oscillator. In quantum mechanics the energy levels of a harmonic oscillator are labelled by a non-negative integer $n = 0, 1, 2, 3, \ldots$ and are equally spaced

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega,$$

where ω is the characteristic frequency of the oscillator. In thermal equilibrium, in contact with a heat bath at temperature T, the probability of a given oscillator being in energy eigenstate n is given by the Boltzmann distribution

$$P_n = \frac{e^{-\frac{E_n}{k_B T}}}{\sum_{n=0}^{\infty} e^{-\frac{E_n}{k_B T}}} = \frac{e^{-\left(n+\frac{1}{2}\right)\frac{\hbar\omega}{k_B T}}}{\sum_{n=0}^{\infty} e^{-\left(n+\frac{1}{2}\right)\frac{\hbar\omega}{k_B T}}} = \frac{y^n}{\sum_{n=0}^{\infty} y^n},$$

where k_B is Boltzmann's constant and $y = e^{-\frac{\hbar\omega}{k_B T}}$ lies in the range $0 \leq y < 1$. The denominator in this expression for P_n is determined by the requirement that the probabilities sum to one, $\sum_{n=0}^{\infty} P_n = 1$. For y in this range

$$\sum_{n=0}^{\infty} y^n = \frac{1}{1-y} \qquad \Rightarrow \qquad P_n = y^n(1-y).$$

The expectation value of n, *i.e.* its most likely value, denoted by $< n >$, is the weighted sum

$$< n >= \sum_{n=0}^{\infty} n P_n = (1-y)\sum_{n=0}^{\infty} n y^n,$$

which can be evaluated using

$$\sum_{n=0}^{\infty} n y^n = y\frac{d}{dy}\left(\sum_{n=0}^{\infty} y^n\right) = y\frac{d}{dy}\left(\frac{1}{1-y}\right) = \frac{y}{(1-y)^2},$$

giving

$$< n >= \frac{y}{1-y} = \frac{1}{e^{\frac{\hbar\omega}{k_B T}} - 1}.$$

This is the **Planck distribution**.

Label the possible crystal vibrational modes by their wavenumber K and a discrete variable s (denoting the different modes: TO, TA, *etc*), then the thermal energy in vibrational modes of the crystal, when it is at a temperature T, is the expectation value of the energy

$$\begin{aligned} U =< E > &= \left\langle \sum_{K,s}\left(n_{K,s} + \frac{1}{2}\right)\hbar\omega_{K,s}\right\rangle = \sum_{K,s}\left(< n_{K,s} > + \frac{1}{2}\right)\hbar\omega_{K,s} \\ &= \sum_K \sum_s \frac{\hbar\omega_{K,s}}{\left(e^{\frac{\hbar\omega_{K,s}}{k_B T}} - 1\right)} + \sum_K \sum_s \frac{\hbar\omega_{K,s}}{2}. \end{aligned}$$

The last term on the right hand side here is a constant, independent of T, and can be ignored in the calculation of thermal properties of crystals below.

For simplicity first consider a monatomic one-dimensional crystal, where we can ignore s (there is only one mode for each K) and $K = \frac{2p}{\mathcal{N}}\left(\frac{\pi}{a}\right)$ with $p = \pm 1, \pm 2, \ldots$. The $\sum_K$ is equivalent to $\sum_p$ but for large $\mathcal{N}$ we can replace the sum with an integral, $\sum_K \to \int D(\omega)d\omega$, where $D(\omega)$ denotes the number of quantum states in the frequency range ω to $\omega + d\omega$. $D(\omega)$ is called the **density of states**, it is calculated below. Thus we get

$$U = \int_0^\infty \frac{D(\omega)\hbar\omega}{\left(e^{\frac{\hbar\omega}{k_BT}} - 1\right)} d\omega. \qquad (22)$$

More generally, for a polyatomic basis and/or in higher dimensions when there is more than one mode for each K, the internal energy of the crystal is

$$\boxed{U = \sum_s \int_0^\infty \frac{D(\omega_s)\hbar\omega_s}{\left(e^{\frac{\hbar\omega_s}{k_BT}} - 1\right)} d\omega_s.}$$

Density of states

To calculate $\mathcal{D}(\omega)$, again initially in one dimension to simplify the demonstration, consider a one-dimensional crystal with lattice spacing a and periodic boundary conditions. The allowed wavevectors are $K = \frac{2p}{\mathcal{N}}\frac{\pi}{a}$ with $p \pm 1, \pm 2, \ldots$, so the spacing between successive wavevectors is $\frac{2}{\mathcal{N}}\frac{\pi}{a}$ and the number of modes in a range δK is $\frac{\mathcal{N}a}{2\pi}\delta K$. The number of modes δN in a frequency range $\delta\omega$ is therefore

$$\delta N = \frac{dN}{d\omega}\delta\omega = 2\left(\frac{\mathcal{N}a}{2\pi}\right)\frac{dK}{d\omega}\delta\omega = D(\omega)\delta\omega$$

(the extra factor of 2 here is inserted to allow for the fact that there are two modes for each ω, one moving to the left and one to the right). Since $\mathcal{N}a = L$, the length of the crystal, this gives

$$\left(\frac{L}{\pi}\right)\frac{dK}{d\omega}\delta\omega = D(\omega)\delta\omega \qquad \Rightarrow D(\omega) = \left(\frac{L}{\pi}\right)\frac{dK}{d\omega},$$

and we can calculate the density of states $D(\omega)$ if we know the dispersion relation $\omega(K)$.

For example the dispersion relation (19) for a one-dimensional crystal, with $\omega_0 = 2\sqrt{\frac{C}{M}}$, reads

$$\omega(K) = \omega_0 \sin\left|\frac{Ka}{2}\right| \qquad \Rightarrow \qquad \frac{d\omega}{dK} = \frac{a}{2}\omega_0\cos\left|\frac{Ka}{2}\right| \qquad (K \geq 0)$$

$$\Rightarrow \qquad D(\omega) = \frac{2L}{a\pi}\frac{1}{\omega_0}\frac{1}{\cos\left(\frac{|K|a}{2}\right)} = \frac{2\mathcal{N}}{\pi}\frac{1}{\sqrt{\omega_0^2 - \omega^2}}$$

Note that at the Brillouin zone boundary, $K \to \frac{\pi}{a}$, $\omega \to \omega_0$ and $D(\omega) \to \infty$. A divergence in the density of states at certain characteristic frequencies is not uncommon and is called a **van Hove singularity**.

In three dimensions we can use the same ideas to get the density of states. Consider a crystal with simple cubic symmetry with $\mathcal{N}$ primitive cells and lattice spacing a. If the linear dimensions are L_1, L_2 and L_3 then the volume is $V = L_1 L_2 L_3 = \mathcal{N} a^3$. For simplicity we take $L_1 = L_2 = L_3 := L = \mathcal{N}^{\frac{1}{3}} a$ and assume $\mathcal{N}^{\frac{1}{3}}$ is an integer, for large $\mathcal{N}$ this is not a significant restriction, at least as far as intrinsic properties of the crystal are concerned. Imposing periodic boundary conditions implies

$$e^{i(K_x x + K_y y + K_z z)} = e^{i\left(K_x(x+L) + K_y(y+L) + K_z(z+L)\right)}$$

$$\Rightarrow \qquad K_x, K_y, K_z = 0, \pm\frac{2\pi}{L}, \pm\frac{4\pi}{L}, \ldots, \pm\frac{\mathcal{N}^{\frac{1}{3}}\pi}{L}.$$

There is therefore one value of K per volume $\left(\frac{2\pi}{L}\right)^3 = \frac{8\pi^3}{V}$ in K-space. The number of quantum modes in a volume $d^3K = dK_x dK_y dK_z$ of K-space is therefore $d^3N = \frac{V}{8\pi^3} d^3K$. For large $\mathcal{N}$ we can approximate the discrete distribution of modes in K-space by a continuum and imagine integrating over a sphere of radius K and area $4\pi K^2$ in K-space, so the radius K is the only variable left,

$$\frac{V}{8\pi^3} d^3K = \frac{V}{8\pi^3} dK_x dK_y dK_z \quad \underset{\int d\Omega}{\longrightarrow} \quad \frac{V}{2\pi^2} K^2 dK.$$

The number of modes inside such a sphere, with volume $\frac{4\pi}{3}K^3$ (*i.e.* with wavenumber less than K), is

$$N = \frac{V}{8\pi^3}\frac{4\pi}{3}K^3 = \frac{V}{6\pi^2}K^3.$$

This now gives the three-dimensional density of states as

$$dN = D(\omega) d\omega = \frac{dN}{d\omega} d\omega = \frac{dN}{dK}\frac{dK}{d\omega} d\omega = \frac{V}{2\pi^2} K^2 \frac{dK}{d\omega} d\omega \qquad \Rightarrow$$

$$\boxed{D(\omega) = \frac{V}{2\pi^2} K^2 \frac{dK}{d\omega}.} \tag{23}$$

which can be evaluated once the dispersion relation, $\omega(K)$, is known.

Debye model

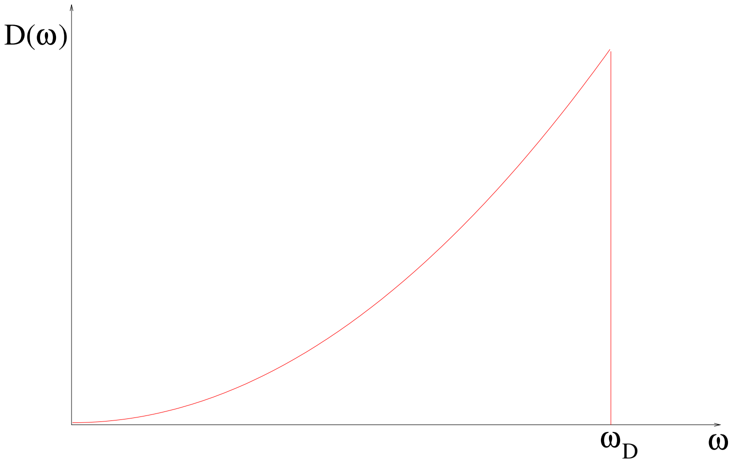
The Debye model makes the simplifying assumption that the dispersion relation is linear, $\omega = vK$, where $v = \frac{d\omega}{dK}$, the speed of sound, is independent of ω. From this we get the density of states

$$D(\omega) = \frac{V}{2\pi^2}\frac{\omega^2}{v^3}.$$

If there are $\mathcal{N}$ primitive cells in the crystal then there is a maximum frequency ω_D, a cut-off frequency, determined by

$$\mathcal{N} = \int_0^{\omega_D} D(\omega) d\omega = \frac{V}{2\pi^2}\frac{1}{v^3}\int_0^{\omega_D} \omega^2 d\omega = \frac{V}{6\pi^2}\frac{\omega_D^3}{v^3} \qquad \Rightarrow \qquad \omega_D = \left(6\pi^2 \frac{\mathcal{N}}{V}\right)^{\frac{1}{3}} v.$$

The maximum angular frequency ω_D is called the **Debye frequency**. With this cut-off the density of states for the Debye model looks like this:



The contribution to the thermodynamic internal energy is

$$U = \int_0^{\omega_D} \frac{D(\omega)\hbar\omega}{e^{\frac{\hbar\omega}{k_B T}} - 1} d\omega = \frac{V\hbar}{2\pi^2 v^3}\int_0^{\omega_D} \frac{\omega^3 d\omega}{e^{\frac{\hbar\omega}{k_B T}} - 1} \tag{24}$$

for each polarisation. For simplicity we shall just take v to be the same for each of the three acoustic modes, then the total internal energy is three times (24). Changing the integration variable from ω to $x = \frac{\hbar\omega}{k_B T}$ gives

$$U = \frac{3V(k_B T)^4}{2\pi^2 v^3 \hbar^3}\int_0^{x_D} \frac{x^3 dx}{e^x - 1},$$

where $x_D = \frac{\hbar\omega_D}{k_B T}$.

It is conventional to define a temperature, Θ_D called the **Debye temperature**, by $k_B \Theta_D = \hbar\omega_D$,

$$\Theta_D = \frac{\left(6\pi^2 \frac{\mathcal{N}}{V}\right)^{\frac{1}{3}} \hbar v}{k_B},$$

with $\frac{\mathcal{N}}{V} := n_c$ the number of primitive cells per unit volume. Then $x_D = \frac{\Theta_D}{T}$ and

$$U = 9\mathcal{N} k_B T \left(\frac{T}{\Theta_D}\right)^3 \int_0^{\frac{\Theta_D}{T}} \frac{x^3 dx}{e^x - 1}.$$

$U(T,V)$ depends on the volume through $\Theta_D \propto V^{-\frac{1}{3}}$.

Other thermodynamic quantities can be obtained from $U(T,V)$. The heat capacity of the crystal at constant volume, for example, is

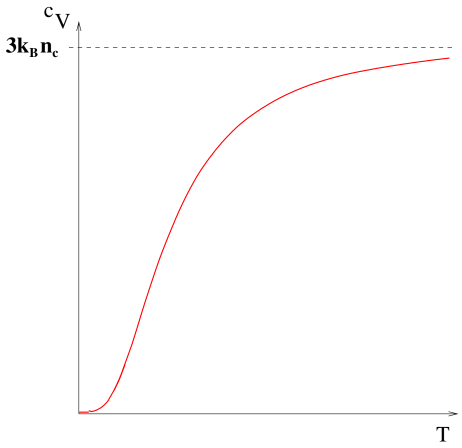
$$C_V = \left(\frac{\partial U}{\partial T}\right)_V .$$

This is most easily calculated from (24), multiplied by 3 to account for the three acoustic modes. The only T dependence in (24) is in $e^{\frac{\hbar\omega}{k_B T}}$, so

$$C_V = \frac{3V\hbar}{2\pi^2 v^3}\frac{\hbar}{k_B T^2}\int_0^{\omega_D}\frac{\omega^4 e^{\frac{\hbar\omega}{k_B T}}\,d\omega}{\left(e^{\frac{\hbar\omega}{k_B T}}-1\right)^2} = \frac{3V}{2\pi^2 v^3}\frac{k_B^4 T^3}{\hbar^3}\int_0^{x_D}\frac{x^4 e^x\,dx}{(e^x-1)^2}$$

$$= 9\mathcal{N}k_B\left(\frac{T}{\Theta_D}\right)^3\int_0^{x_D}\frac{x^4 e^x\,dx}{(e^x-1)^2}.$$

The specific heat, $c_V = \frac{C_V}{V}$, is plotted below:



We can evaluate the integral in (24) in certain limits:

- Low temperatures: $k_B T << \hbar\omega_D$, $x_D \to \infty$,

$$\int_0^\infty \frac{x^3}{e^x-1} = \sum_{n=1}^\infty \int_0^\infty x^3 e^{-nx}dx = \sum_{n=1}^\infty \frac{1}{n^4}\int_0^\infty u^3 e^{-u}dx \qquad \text{where } u = nx$$

$$= \Gamma(4)\sum_{n=1}^\infty \frac{1}{n^4} = 3!\sum_{n=1}^\infty \frac{1}{n^4} = \frac{\pi^4}{15},$$

leading to thermal energy

$$U \approx \frac{3\pi^4}{5}\mathcal{N}k_B T\left(\frac{T}{\Theta_D}\right)^3$$

and specific heat

$$c_V = \frac{C_V}{V} \approx \frac{12\pi^4}{5}\left(\frac{\mathcal{N}}{V}\right) k_B \left(\frac{T}{\Theta_D}\right)^3 = \frac{2\pi^2}{5} k_B \left(\frac{k_B T}{\hbar v}\right)^3 . \tag{25}$$

There formula are only correct for T small, in particular

$$\lim_{T\to 0} \frac{C_v}{VT^3} = \frac{2\pi^2}{5}\frac{k_B^4}{(\hbar v)^3}$$

is constant. This is an important result from the Debye approximation, the specific heat due to crystal vibrations goes like $\sim T^3$ at low T. For a metallic crystal there is another contribution to the specific heat, due to electrons free to roam around the crystal, which we shall evaluate later. It may be necessary to go temperatures as low as $T < \frac{\Theta_D}{50}$ to see this T^3 behaviour.

- In the opposite limit, of high temperatures, $x_D << 1$, we can expand $\frac{1}{e^x - 1} = \frac{1}{x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots} = \frac{1}{x}\left(1 - \frac{x}{2} + \frac{x^2}{12} - \cdots\right)$ and

$$U \approx 9\mathcal{N} k_B T \left(\frac{T}{\Theta_D}\right)^3 \frac{x_D^3}{3} = 3\mathcal{N} k_B T$$

is linear in T, hence the specific heat is constant

$$c_V = \frac{C_V}{V} \approx \frac{3\mathcal{N} k_B}{V}.$$

This is the classical result — constant specific is indeed observed at large T and is known as the **Dulong-Petit** result. The Dulong-Petit value for the specific heat of a crystal can be understood from the equipartition theorem: each degree of freedom in the crystal has the same energy $\frac{1}{2}k_B T$, each atom has 3 co-ordinates labelling its position and 3 momenta giving 6 degrees of freedom, hence the internal energy is $U = 6\mathcal{N}\frac{k_B T}{2} = 3\mathcal{N} k_B T$.[10] This classical result assumes that all degrees of freedom are excited but, if T is not very large $T << \Theta_D$, not all degrees of freedom can be excited and the specific heat is reduced

Values of Θ_D for some elements are: 158°K (Na); 400°K (Mg); 470°K (Fe); 2230°K (C).

Einstein model

The linear dispersion relation, $\omega = vK$, in the Debye model is a reasonable approximation for acoustic modes at small K, it is not a good model for optical modes in a crystal with a polyatomic basis. Einstein suggested a simplified density of states

$$D(\omega) = \mathcal{N}\delta(\omega - \omega_E)$$

[10] The internal energy of a monatomic gas is $\frac{3}{2}\mathcal{N} k_B T$, not $3\mathcal{N} k_B T$, because the degrees of freedom associated with the positions of the atoms in an ideal gas do not contribute to the energy and so do not contribute to the internal energy. In a crystal the position does contribute as it takes energy to move an atom away from its equilibrium position.

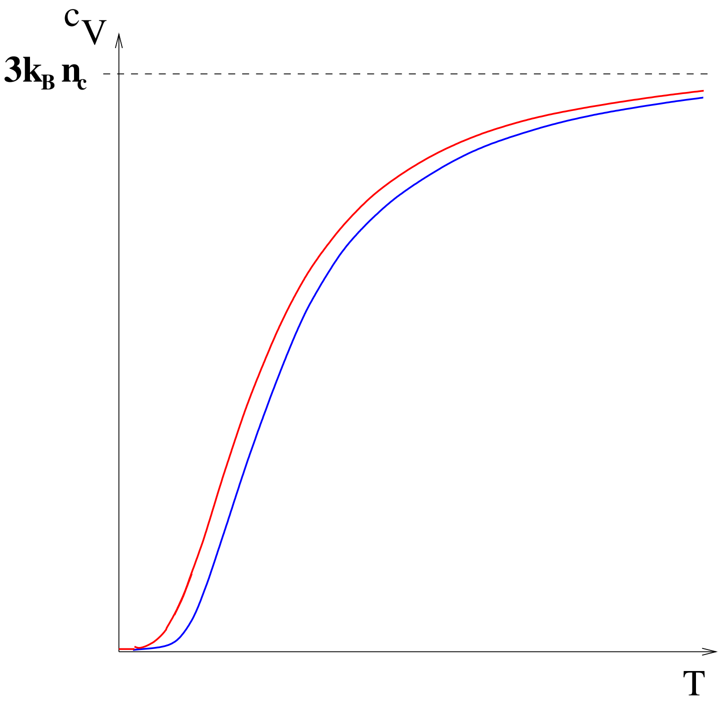
in this case, where ω_E is a fixed frequency and $\delta(\omega - \omega_E)$ is a Dirac δ-function, vanishing unless $\omega = \omega_E$. The integral over x in (24) is trivial in this case: if there are p optical modes, all with the same ω_E,

$$U = \frac{p\mathcal{N}\hbar\omega_E}{e^{\frac{\hbar\omega_E}{k_BT}} - 1}$$

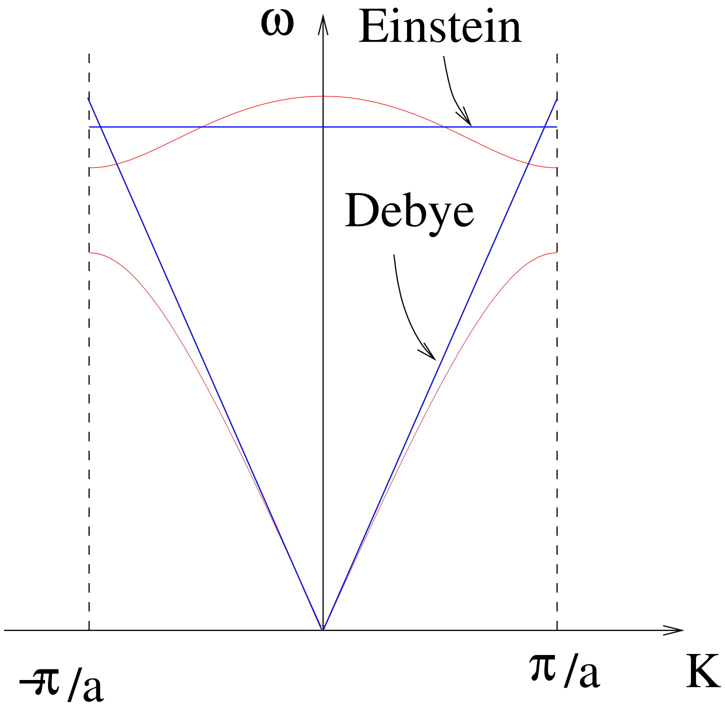
and the specific heat is

$$c_V = \left(\frac{\mathcal{N}}{V}\right)\frac{(\hbar\omega_E)^2}{k_BT^2}\frac{p\,e^{\frac{\hbar\omega_E}{k_BT}}}{(e^{\frac{\hbar\omega_E}{k_BT}} - 1)^2} \quad \to \quad \begin{cases} p\left(\frac{\mathcal{N}}{V}\right)k_B, & T \to \infty \\ p\left(\frac{\mathcal{N}}{V}\right)\left(\frac{\hbar\omega_E}{k_BT}\right)^2 k_B e^{-\frac{\hbar\omega_E}{k_BT}}, & T \to 0. \end{cases}$$

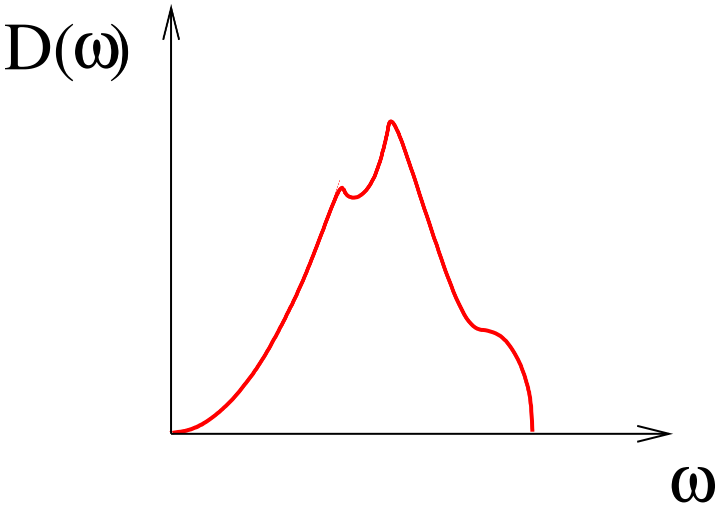
The Einstein result is the same for large T as the Debye result, the specific heat approaches a constant at large T, but at low T the specific heat for optical modes in the Einstein model is much less than that of the acoustic modes in the Debye model. The two are compared below (with $p = 3$): the red curve is the Debye model and blue Einstein model,



It is stressed that these calculations only take into account the vibrational modes of the crystal, any contribution from free electrons is ignored. The low T results are only valid for crystals that are electrical insulators, metallic crystals have an extra contribution to the specific heat coming from free electrons in the crystal. We shall see later that electrons contribute a linear term to the acoustic mode specific heat in a metallic crystal, giving $c_V \approx AT + BT^3$ at low T, with A and B constants. At very low temperatures the linear term dominates the cubic term and the metallic specific heat is linear in T.

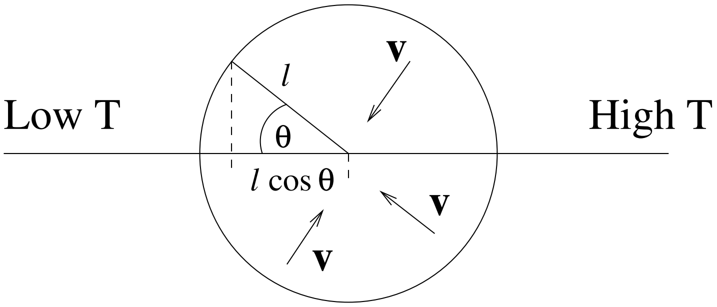


Both the Debye and the Einstein models are crude approximations to the dispersion relation in real crystals, they are plotted in blue above and compared to the one-dimensional diatomic results for acoustic and optical modes calculated earlier. Real crystals are more complicated: a real experimental dispersion relation for phonons, determined by neutron scattering, for acoustic modes in aluminium, is shown below,



Thermal conductivity

Heat energy in a crystal is due to vibrating atoms and so we expect phonons to conduct heat. For simplicity consider a crystal with monatomic basis. Denote the equilibrium energy density in phonons (lattice vibrations) by $w(\mathbf{r})$ (so the internal energy is $\int_{crystal} w(\mathbf{r})dV$) and the phonon velocity by $\mathbf{v}$ (in the presence of a temperature gradient $w(\mathbf{r})$ will vary from place to place). Now introduce a temperature gradient $T(x)$ in the x-direction and let the phonon mean free path (the average distance between phonon collisions) be l. Then the average time between phonon collisions is $\tau = \frac{l}{v}$. Any phonon arriving at a general point $\mathbf{r}_0$ of the crystal has, on average, come from a sphere of radius l centred on $\mathbf{r}_0$, this sphere represents the locus of points from which the phonons arriving at $\mathbf{r}_0$ last scattered and $w(\mathbf{r})$ will be different at different points on this sphere so, in the presence of a temperature gradient, phonons arriving from different directions will carry different energy — those coming from directions in which the temperature is hotter will have greater energy than those coming from directions in which the temperature is cooler. If $T(x)$ is constant in the y and z-directions then $w(\mathbf{r})$ will be too and $w(x)$ depends only on x.



There will be a net flux of energy, a thermal current, in the direction of decreasing T as heat energy diffuses from regions of higher T to lower T. The x component of $\mathbf{v}$ is $v_x = v\cos\theta$ and, denoting an infinitesimal area element of the sphere by $dA = l^2 \sin\theta d\theta d\phi$ the thermal current is

$$\begin{aligned}
J =& \frac{1}{4\pi l^2}\int_{sphere} v_x w(x) dA \\
=& \frac{2\pi}{4\pi l^2}\int_0^\pi (v\cos\theta) w(x_0 - l\cos\theta) l^2 \sin\theta d\theta \\
\approx& \frac{v}{2}\int_0^\pi \left\{ w(x_0) - l\cos\theta\left(\frac{dw}{dx}\right)_{x_0}\right\}\cos\theta\sin\theta d\theta \\
=& \frac{v}{2}\int_{-1}^1 \left\{ w(x_0)\alpha - \alpha l\left(\frac{dw}{dx}\right)_{x_0}\right\}\alpha d\alpha \qquad\qquad (\alpha = \cos\theta) \\
=& -\frac{vl}{3}\left(\frac{dw}{dx}\right)_{x_0}.
\end{aligned}$$

Now $\frac{dw}{dx}$ is related to the thermal gradient, $\frac{dT}{dx}$, by the chain rule

$$\frac{dw}{dx} = \frac{dw}{dT}\frac{dT}{dx}.$$

Since there is a thermal gradient the system is not in thermal equilibrium but we still expect the thermal energy per unit volume $w(T, V)$ to depend on V as well as T, $\frac{dw}{dT}$ here is more correctly written $\frac{\partial w}{\partial T}\Big|_V$ which is the specific heat at constant volume, c_V, so

$$J = -\frac{c_V v l}{3}\frac{dT}{dx} = -\frac{c_V v^2 \tau}{3}\frac{dT}{dx},$$

where $\tau = \frac{v}{l}$ is average time between phonon collisions. The **thermal conductivity**, κ, is defined as the ratio of the thermal current to the thermal gradient,

$$J = -\kappa \frac{dT}{dx},$$

and we get the important result that the thermal conductivity

$$\boxed{\kappa = \frac{c_V v^2 \tau}{3}} \tag{26}$$

is proportional to the specific heat of the crystal.

Two limiting cases:

- At high T, $c_V = 3k_B n_c$ is a constant. It is reasonable to expect that the collision rate will be proportional to the phonon density,

  $$\tau^{-1} \propto < n >= \frac{1}{e^{\frac{\hbar\omega}{k_B T}} - 1} \underset{T\to\infty}{\approx} \frac{k_B T}{\hbar\omega} \propto T,$$

  Since the phonon velocity is independent of the temperature (it is determined by the dispersion relation), we expect

  $$\kappa \propto 1/T$$

  at high T. Experimentally $\kappa \propto \frac{1}{T^\nu}$ with ν between 1 and 2.

- For low T, $< n > \approx e^{-\frac{\hbar\omega}{k_B T}} \Rightarrow \tau \propto e^{\frac{\hbar\omega}{k_B T}} \to \infty$ as $T \to 0$. Hence $\kappa \to \infty$, except that the photon mean free path l is necessarily limited by the crystal size or, more realistically, the distribution of lattice imperfections or chemical impurities in the crystal, so τ tends to some finite value τ_0 as $T \to 0$ and $\kappa \to \frac{c_v v^2 \tau_0}{3}$. In the Debye approximation $c_V \propto T^3$ at low T, so

  $$\kappa \propto T^3.$$

Crystal momentum and Umklapp processes

We can always map any wavevector into the first Brillouin zone by adding a reciprocal lattice vector. If $\mathbf{K}$ is not in the first Brillouin zone there always exists a reciprocal lattice vector $\mathbf{G}$ such that $\mathbf{K} + \mathbf{G}$ is. This is a three-dimensional generalisation of our earlier observation that, for a one-dimensional crystal with lattice spacing a, we need only consider wave-numbers $|K| \leq \frac{\pi}{a}$. If two phonons with wave-vectors $\mathbf{K}_1$ and $\mathbf{K}_2$, both in the first Brillouin zone, collide and merge to give a single phonon with wave-vector $\mathbf{K}_3'$ then conservation of momentum says that

$$\hbar\mathbf{K}_1 + \hbar\mathbf{K}_2 = \hbar\mathbf{K}_3', \tag{27}$$

but $\mathbf{K}_3'$ may not be in the first Brillouin zone. However we can always find a reciprocal lattice vector $\mathbf{G}$ so that $\mathbf{K}_3 = \mathbf{K}_3' + \mathbf{G}$ is in the first Brillouin zone,

$$\hbar\mathbf{K}_1 + \hbar\mathbf{K}_2 = \hbar\mathbf{K}_3 + \hbar\mathbf{G}. \tag{28}$$

If $\mathbf{G} = 0$ then we obviously have

$$\hbar\mathbf{K}_1 + \hbar\mathbf{K}_2 = \hbar\mathbf{K}_3$$

identically, this called a **normal** process (N-process). Even if $\mathbf{G} \neq 0$ it still plays no role in the physics and equation (28) is completely equivalent to

$$\hbar\mathbf{K}_1 + \hbar\mathbf{K}_2 = \hbar\mathbf{K}_3. \tag{29}$$

As explained at the bottom of page 40 for a one-dimensional crystal wave-vectors outside the first Brillouin are not important for phonon physics and the same is true in three dimensions. (27) and (28) are indistinguishable physically. A $\mathbf{G} \neq 0$ process is called an **umklapp** process (U-process).[11] An umklapp process involves Bragg reflection of the final state phonon from a Brillouin zone boundary. The momentum $\hbar K$ is called the **crystal momentum** and it is not conserved absolutely in an umklapp process, it is only conserved up to a reciprocal lattice vector. Conservation laws in physics are a consequence of symmetries of the underlying dynamics and in free space conservation of momentum is a consequence of translation invariance. A crystal does not have translational invariance under arbitrary small displacements, it only has translational invariance under discrete translations by a direct lattice vector. This is a smaller symmetry than invariance under all possible translations of any magnitude and the resulting conservation law, conservation of crystal momentum, is less powerful than in free space — we only have conservation of momentum up to a reciprocal lattice vector.

At a temperature T we only expect phonons with $\hbar\omega \lesssim k_B T$ to be present and, if T is not too high, this means $\omega << \omega_D$ that $\mathbf{K}_1$ so $\mathbf{K}_2$ will be small and deep within the first Brillouin zone so that $\mathbf{K}_3$ is also well within the first Brillouin zone. Umklapp processes will then be very rare and conservation of crystal momentum is exact momentum conservation. If this is the case then there is no dissipation in phonon collisions, momentum and energy are conserved and we expect the thermal conductivity $\kappa \to \infty$ at low T (this argument assumes a perfect crystal and ignores impurities and imperfections in the crystal). As the temperature increases umklapp processes become more common and momentum leaks out of the phonons and through umklapp processes giving rise to dissipation and energy loss. Of course the total physical momentum is still conserved, $\hbar\mathbf{G}$ is absorbed by the crystal as it is buffeted about by the phonons.

[11] “Umklapp” means “flip over” in German.

6. Metals

In a metallic crystal, such as magnesium or iron, we need to take account of the fact that some electrons are free to move around the crystal and are not necessarily bound to specific atoms as they are in an insulator. There is a background sea of mobile electrons. To a first approximation we can treat these electrons as an ideal gas, though we must be careful to take into account the quantum nature of the electrons.

Consider a cubic crystal of size L, so the volume is $V = L^3$. Impose periodic boundary conditions on the electron wave-functions,

$$\Psi(x+L, y, z) = \Psi(x, y, z),$$

and similarly for the y and z directions. Assume any wave-function can be expanded in a basis of plane-waves,

$$\psi_{\mathbf{k}}(\mathbf{r}, t) = e^{i(\mathbf{k}.\mathbf{r}-\omega_{\mathbf{k}}t)}.$$

$\psi_{\mathbf{k}}$ is periodic, with period L, if $k_x = 0, \pm\frac{2\pi}{L}, \pm\frac{4\pi}{L}, \ldots$, with similar conditions on k_y and k_z. Unlike the case of phonons there is no upper limit on k_i being imposed here.

In the absence of any interactions the Shrödinger equation for $\psi_{\mathbf{k}}$ is

$$i\hbar\frac{\partial\psi_{\mathbf{k}}}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi_{\mathbf{k}}$$

which gives

$$\hbar\omega_{\mathbf{k}} = \frac{\hbar^2}{2m}\mathbf{k}.\mathbf{k} \qquad \Rightarrow \qquad \varepsilon_{\mathbf{k}} = \hbar\omega_{\mathbf{k}} = \frac{\hbar^2}{2m}k^2.$$

The momentum is

$$\hat{\mathbf{p}}\psi_{\mathbf{k}} = -i\hbar\nabla\psi_{\mathbf{k}} = \hbar\mathbf{k}\psi_{\mathbf{k}} \qquad \Rightarrow \qquad \mathbf{p} = \hbar\mathbf{k},$$

as usual. The quantum nature of the electrons, together with the periodic boundary conditions implies that there is one quantum state, one wave-vector, for every volume $\left(\frac{2\pi}{L}\right)^3$ in k-space. If there are N mobile electrons[12] in the volume V each must occupy a separate quantum state, because of the exclusion principle. One might expect the electrons quantum states to fill a sphere in k-space of volume $N\left(\frac{2\pi}{L}\right)^3 = N\left(\frac{8\pi^3}{V}\right)$, except there is an extra factor of 2 due to the fact that electrons have spin-1/2 and therefore have two spin states, spin up and spin down, for each value of $\mathbf{k}$, so they actually fill a sphere with half this volume, $N\left(\frac{4\pi^3}{V}\right)$. For finite N the distribution of quantum states in k-space is discrete but in the limit of large N it can be approximated with a smooth continuous distribution within a sphere of volume

$$\frac{4\pi}{3}k_F^3 = N\left(\frac{4\pi^3}{V}\right)$$

[12] N might not be the same as the number of primitive cells in the crystal, there could be more than one mobile electron per primitive cell, *e.g.* Mg. For a monovalent metal with a monatomic basis, *e.g.* Na, K, N=$\mathcal{N}$.

and radius

$$k_F = \left(3\pi^2 \frac{N}{V}\right)^{\frac{1}{3}}.$$

This is called the **Fermi sphere** and k_F is the **Fermi wave-number**, $p_F := \hbar k_F$ is the **Fermi momentum**. The energy of a state with wave-number k_F is called the **Fermi energy**: for non-interacting electrons the Fermi energy is

$$\varepsilon_F = \frac{p_F^2}{2m} = \frac{\hbar^2 k_F^2}{2m} = \frac{\hbar^2}{2m}\left(\frac{3\pi^2 N}{V}\right)^{\frac{2}{3}} = \frac{1}{2} m v_F^2, \tag{30}$$

where

$$v_F = \frac{p_F}{m} = \frac{\hbar k_F}{m} = \frac{\hbar}{m}\left(\frac{3\pi^2 N}{V}\right)^{\frac{1}{3}}$$

is the **Fermi velocity**. In a simple cubic crystal with a monatomic, monovalent basis with lattice spacing 5Å for example, $\frac{N}{V} = \left(\frac{1}{5\times 10^{-10}}\right)^3 m^{-3} = 8 \times 10^{27} m^{-3}$ giving $v_F = 7 \times 10^5 m/s \approx 2 \times 10^{-3} c$, a remarkably high velocity. Compare this with the thermal velocity of particles of mass m in a perfect gas: $v_T = \sqrt{\frac{2k_B T}{m}} \approx 3000 ms^{-1}$ at $T = 300K$. The Fermi energy for most metals is about an order of magnitude higher than chemical energies.

When interactions between electrons and the lattice are included the lattice structure will distort the Fermi surface away from a sphere to a shape with less symmetry than a sphere but which reflects the symmetry of the underlying lattice.

At zero temperature all quantum states with $0 \leq k \leq k_F$ are filled and all quantum states with $k > k_F$ are empty. At finite temperature thermal fluctuations can kick an electron with $k < k_F$ into a quantum state with $k > k_F$ provided $\epsilon(k) \lesssim \epsilon(k_F) + k_B T$, leaving a state with $k < k_F$ empty. Such an empty state is called a **hole**.

From (30) the Fermi energy depends on N, conversely N depends on the Fermi energy. We can determine the density of states for free (non-interacting) electrons at any energy from

$$\varepsilon = \frac{\hbar^2}{2m}\left(\frac{3\pi^2 N_\varepsilon}{V}\right)^{\frac{2}{3}} \qquad \Rightarrow \qquad N_\varepsilon = \left(\frac{2m\varepsilon}{\hbar^2}\right)^{\frac{3}{2}} \frac{V}{3\pi^2}.$$

N_ε is the total number of quantum states, with energy less then ε, available to an electron. The density of states is then

$$\mathcal{D}(\varepsilon) = \frac{dN_\varepsilon}{d\varepsilon} = \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}} \varepsilon^{\frac{1}{2}} = \frac{3}{2}\frac{N_\varepsilon}{\varepsilon}. \tag{31}$$

Fermi-Dirac distribution function

In order to understand the physics of electrons in metals it is important to know the probability distribution for the number of electrons expected to have energy ε at any given temperature. For a gas of photons, Planck proposed that the photons corresponding to light with frequency $\nu = \frac{\omega}{2\pi}$ could only carry energy which is an integral multiple of $\varepsilon = \hbar\omega$, $E = n\varepsilon$, with $n = 1, 2, 3, \ldots$ and this notion generalises to other particles, such as electrons. One subtlety is that for particles that are not simply free or non-interacting, such as electrons in a solid or particles in a non-ideal gas, the energy can be shifted by a constant, called the **chemical potential** μ, so $\varepsilon \to \varepsilon - \mu$. Essentially μ is the amount of energy needed to add one more particle to a system of N particles, if $\mu < 0$ then particles are attracted to the system and $|\mu|$ is a binding energy, if $\mu > 0$ then particles are pushed away from the system (chemical potentials are covered in more depth in the statistical mechanics module).

The Boltzmann distribution for system of particles with allowed energy levels $E_n = n(\epsilon - \mu)$ in thermal equilibrium with a heat bath at temperature T gives the probability of energy level ε being occupied,

$$P(E_n) = \frac{e^{-\frac{n(\varepsilon-\mu)}{k_BT}}}{Z},$$

with $Z = \sum_{n=0}^{\infty} e^{-\frac{n(\varepsilon-\mu)}{k_BT}}$ chosen[13] that the total probability $\sum_{n=0}^{\infty} P(\varepsilon) = 1$.

Let $y = e^{-\frac{\varepsilon-\mu}{k_BT}}$ then for bosons, such as photons, $Z = \sum_{n=0}^{\infty} y^n = \frac{1}{1-y}$ giving the distribution function for bosons with angular frequency ω,

$$f_B(\varepsilon) := P(\varepsilon) = \frac{y}{1-y} = \frac{1}{e^{\frac{\varepsilon-\mu}{k_BT}} - 1}, \tag{32}$$

provided $\varepsilon > \mu$. $f_B(\varepsilon)$ represents the probability of finding a boson with energy ϵ when the ambient temperature is T — it is called the **Bose-Einstein** distribution.

For a system of fermionic particles, such as electrons, the basic principle is the same, except that it must be remembered that each individual term for a given n in the sum for Z in the Bose-Einstein distribution corresponds physically to n bosons occupying the same quantum state. As many bosons as one wishes can go into the same energy state, but the Pauli exclusion states that *at most* one Fermion can occupy each quantum state, so for Fermions $n = 0$ or 1 only, $n \geq 2$ is not allowed. This means $Z = \sum_{n=0}^{1} y^n = 1 + y$, leading to the **Fermi-Dirac** distribution function for fermions, $f_F(\varepsilon) := P(\varepsilon) = \frac{y}{y+1}$ or

$$\boxed{f_F(\varepsilon) = \frac{1}{e^{\frac{\varepsilon-\mu}{k_BT}} + 1}.} \tag{33}$$

The difference in sign in the denominators of (32) and (33) is the source a great difference between the behaviour of Fermions and Bosons at low temperatures.

[13] Z is called the **partition function** for the system.

If $\mu < 0$ and $\frac{|\mu|}{k_BT} >> \frac{\varepsilon}{k_BT}$ the 1 in the denominator is irrelevant and both the Bose-Einstein and the Fermi-Dirac distributions look the same. This is the case in a classical gas where both reduce to the Boltzmann distribution

$$f(\varepsilon) \approx e^{\frac{\mu}{k_BT}} e^{-\frac{\varepsilon}{k_BT}} = \frac{e^{-\frac{\varepsilon}{k_BT}}}{Z},$$

with $Z = e^{-\frac{\mu}{k_BT}} = \sum e^{-\frac{\varepsilon}{k_BT}}$. We can calculate μ assuming that the gas particles are free, so the allowed energies, $\varepsilon = \frac{1}{2}mv^2$, are just their kinetic energies at speed v and, in the classical limit, replace the sum with an integral. Then using (31), with the factor 2 arising from spin states of an electron removed so $V \to \frac{V}{2}$,

$$\begin{aligned}
Z &= \int_\varepsilon e^{-\frac{\varepsilon}{k_BT}} \frac{dN_\varepsilon}{d\varepsilon} d\varepsilon = \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}} \int_0^\infty \varepsilon^{\frac{1}{2}} e^{-\frac{\varepsilon}{k_BT}}\, d\varepsilon \\
&= \frac{V}{4\pi^2}\left(\frac{2mk_BT}{\hbar^2}\right)^{\frac{3}{2}} \int_0^\infty u^{\frac{1}{2}} e^{-u}\, du \qquad \text{with} \qquad u = \frac{\varepsilon}{k_BT} \\
&= \frac{V}{4\pi^2}\left(\frac{2mk_BT}{\hbar^2}\right)^{\frac{3}{2}} \frac{\sqrt{\pi}}{2} = V\left(\frac{mk_BT}{2\pi\hbar^2}\right)^{\frac{3}{2}} = e^{-\frac{\mu}{k_BT}}.
\end{aligned}$$

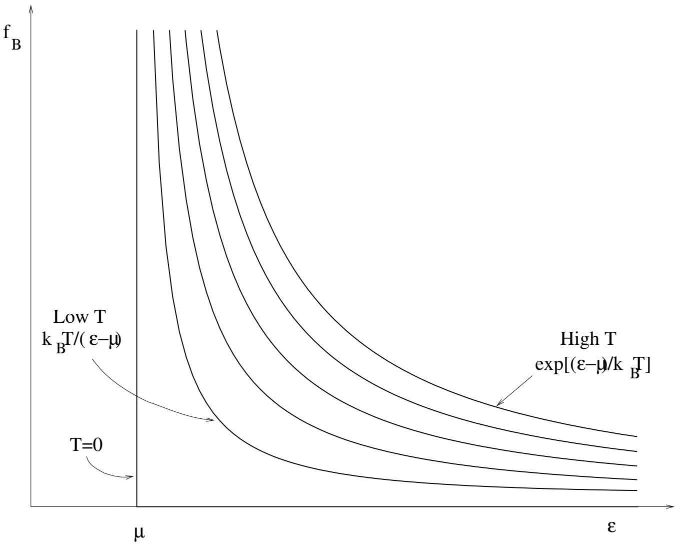
hence

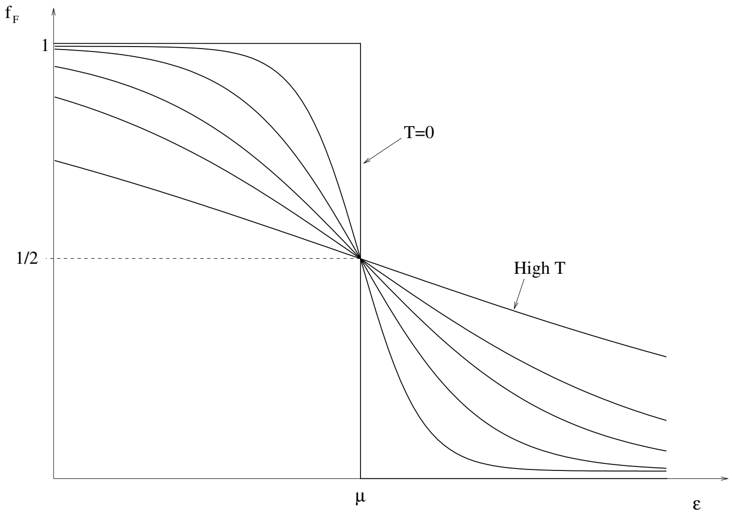
$$\mu = -\frac{3}{2}k_BT\ln(k_BT) + const$$

is negative.

The Boltzmann distribution is the large T, low density, limit of the distribution — the classical limit.

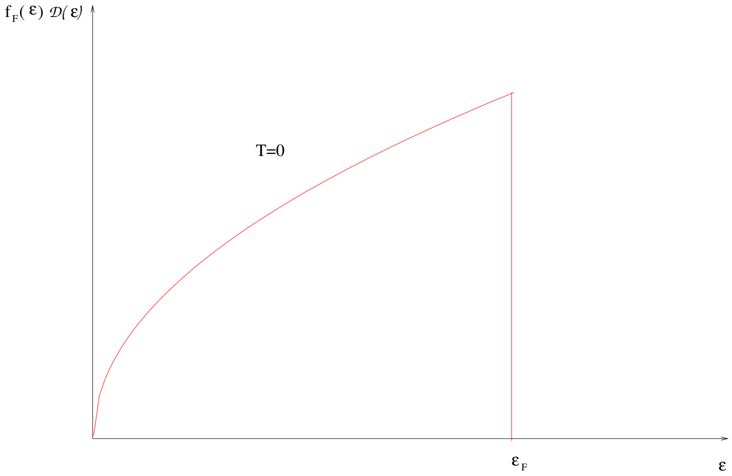
When quantum effects are important the 1 in the denominator of equations (32) and (33) cannot be ignored — roughly speaking this happens when the particle density becomes large enough for their wave-packets to overlap significantly, that is when the separation between the particles becomes of the order of, or less than, their de Broglie wave-length. The Bose-Einstein and Fermi-Dirac distributions, for fixed μ and different T, are plotted below:





At $T = 0$ the Bose-Einstein distribution has no states with $\varepsilon > \mu$ occupied, the only occupied states are those with $\varepsilon = \mu$. All particles occupy the *same* quantum state, a situation known as Bose-Einstein condensation. This phenomenon occurs in superfluids and superconductors.

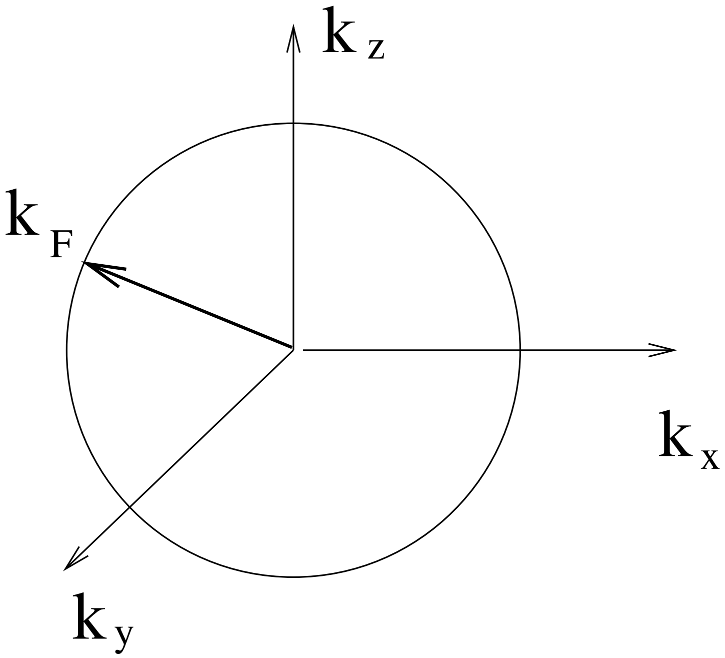
At $T = 0$ the Fermi-Dirac distribution has every quantum state with $\varepsilon < \mu$ occupied, $f_F(\varepsilon) = 1$ for $\varepsilon < \mu$, and every state with $\varepsilon > \mu$ unoccupied, $f_F(\varepsilon) = 0$ for $\varepsilon > \mu$. $\varepsilon = \mu = \varepsilon_F$ is the Fermi surface. [14]



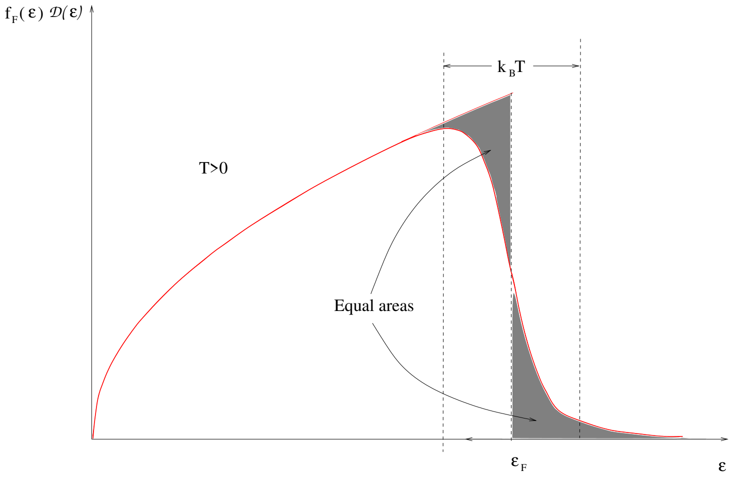
At $T = 0$ the product $f_F(\varepsilon)\mathcal{D}(\varepsilon)$ cuts off at $\varepsilon = \varepsilon_F$ and drops to zero.

The distribution of quantum sates in k-space at $T = 0$ is the interior of a solid ball of radius k_F:

[14] Actually the plot of $f_F(\varepsilon)$ above is produced assuming μ is independent of T. As we have seen this is not a good assumption for a classical ideal gas (the classical case is best visualised by the high T curves in f_B), but is often a good assumption for low T, in particular when $k_BT{<<}\mu$. As we shall see for a real Fermi gas μ has a slight T dependence near ε_F and the curves for $f_F(\varepsilon)$ at different T do not all cross at exactly the same point.



For $T > 0$ some electrons are thermally excited above ε, leaving unoccupied states below ε. The excited and unoccupied states lie in a band of width k_BT around ε_F.



The above graph of $f_F(\varepsilon)\mathcal{D}(\varepsilon)$ as a function of ε shows the number of electrons with energy ε: the total area under the curve is N, the number of electrons in the crystal. The chemical potential can be determined as a function of T and N by the condition

$$N = \int_0^\infty f_F(\varepsilon)\mathcal{D}(\varepsilon)d\varepsilon = \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}} \int_0^\infty \frac{\varepsilon^{\frac{1}{2}}}{\left(e^{\frac{\varepsilon-\mu}{k_BT}}+1\right)}d\varepsilon. \tag{34}$$

Heat capacity of a free electron gas

Classically, an ideal gas of N particles has heat capacity[15]

$$C_V = \frac{3}{2} N k_B, \tag{35}$$

and specific heat

$$c_V = \frac{C_V}{V} = \frac{3}{2}\frac{N}{V} k_B = \frac{3}{2} n_e k_B, \tag{36}$$

where $n_e = \frac{N}{V}$ is the number of electrons per unit volume.[16] The observed c_V in metals is only about $\sim 1\%$ of (36), a phenomenon which perplexed nineteenth century physicists but which, from a more modern perspective, can be qualitatively be understood as being due to the Pauli exclusion principle. Only those electrons within a distance $\sim k_B T$ of the Fermi surface are free to contribute to the specific heat, electrons any deeper than $k_B T$ below the Fermi surface have their dynamics 'frozen' by the exclusion principle: there are no unoccupied energy states nearby and so these electrons have no degrees of freedom to contribute to the specific heat.

To get a quantitative expression for the specific heat due to free electrons in a metal we first need the internal energy

$$U(T) = \int_0^\infty \varepsilon f_F(\varepsilon) \mathcal{D}(\varepsilon) d\varepsilon \qquad (37).$$

The heat capacity is then $\frac{\partial U}{\partial T}\big|_V$ but one subtlety is that we need to calculate $U(T)$ at constant N, while the right hand side of (37) is a function of μ which in turn depends on T and N through (34). It is not possible to perform the integral (37) analytically so we resort to an approximation, but it is a very good approximation. Define the *Fermi temperature*, T_F by

$$k_B T_F = \varepsilon_F.$$

Typically $T_F \approx 50,000^\circ\ K$ is a very large temperature and we shall expand $U(T)$ as a function $\frac{T}{T_F}$ — this is known as a **Sommerfeld expansion**. The details are given in an appendix (Appendix A)

$$\frac{\mu}{\varepsilon_F} = 1 - \frac{\pi^2}{12}\left(\frac{k_B T}{\varepsilon_F}\right)^2 + O\left(\frac{k_B T}{\varepsilon_F}\right)^4. \tag{38}$$

This an extremely good approximation: remember $\varepsilon_F = k_B T_F$ typically corresponds to a temperature $T_F \approx 50,000^\circ\ K$ so, even at $T = 500^\circ\ K$, $\frac{k_B T}{\varepsilon_F}$ is only of order 10^{-2} and

[15] This can be understood in terms of the equipartition theorem: the energy of a free particle is distributed equally among the three degrees of freedom associated with the three components of momenta. The equipartition theorem of thermodynamics states that each degree of freedom receives an amount of energy $\frac{1}{2}k_B T$, so the energy of each particle is $\frac{3}{2}k_B T$. The total energy in a gas consisting of N particles is $\frac{3}{2}N k_B T$ and the heat capacity is $\frac{3}{2}N k_B$.

[16] The specific heat is a better representation of the intrinsic characteristics of the metal as it is independent of the size of the crystal, the heat capacity itself depends on how big the crystal is.

$\left(\frac{k_BT}{\varepsilon_F}\right)^2$ is of order 10^{-4}, so $\mu = \varepsilon_F$ to one part in 10,000 and the quadratic approximation to $\mu(T)$ in (38) is accurate to one part in one hundred million.

The integral in (37) can be evaluated using the same techniques as in appendix A. The calculation is left as an exercise and the answer is

$$U = \frac{3}{5} N\varepsilon_F \left(\frac{\mu}{\varepsilon_F}\right)^{\frac{5}{2}} + \frac{3\pi^2}{8} N\varepsilon_F \left(\frac{k_BT}{\varepsilon_F}\right)^2 \left(\frac{\mu}{\varepsilon_F}\right)^{\frac{1}{2}} + \dots .$$

Using the expansion (38) then gives

$$U = \frac{3}{5} N\varepsilon_F + \frac{\pi^2}{4} N\varepsilon_F \left(\frac{k_BT}{\varepsilon_F}\right)^2 + \dots .$$

The electronic specific heat is now given by

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V = \frac{N\pi^2 k_B^2 T}{2\,\varepsilon_F} + O\left(\frac{k_BT}{\varepsilon_F}\right)^3 .$$

This explains why electronic heat capacities are only about 1% of the classical value (35), they are reduced by the factor $\frac{\pi^2}{3}\frac{k_BT}{\varepsilon_F} \approx 10^{-2}$ by the exclusion principle. The specific heat is

$$c_V = \frac{N}{V}\frac{\pi^2 k_B^2 T}{2\,\varepsilon_F} + O\left(\frac{k_BT}{\varepsilon_F}\right)^3 = \frac{m}{\hbar^2}\frac{(n_e)^{\frac{1}{3}} k_B^2 T}{(3\pi^2)^{\frac{2}{3}}} + o\left(\frac{k_BT}{\varepsilon_F}\right)^3 , \qquad (39)$$

a formula that is accurate to one part in 10,000 for $T \approx 500^\circ\ K$.

The full specific heat of a metallic crystal includes the phonon contribution from the lattice, $\propto T^3$ at low temperatures and $3nk_B$ at high temperatures (the Dulong-Petit value).[17] At very low temperatures the electronic specific heat dominates in metals.

This calculation also has implications for the thermal conductivity κ, since $\kappa = \frac{c_v v_F^2 \tau}{3}$ and c_v is reduced by a factor of 1/100 from its naive classical value.

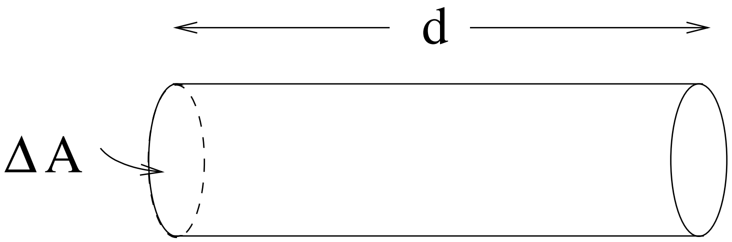
DC conductivity

Consider a segment of metal wire of length d and constant cross-sectional area ΔA. The electrical resistance, R, depends on the geometry, doubling the length doubles the resistance and thick wires have less resistance than thin wires: R is proportional to the length and inversely proportional to ΔA,

$$R = \frac{d}{\Delta A}\rho,$$

where ρ is called the resistivity of the metal (it has dimensions of *Ohms* × *length*). R is not intrinsic to the material but ρ is, for this reason ρ is more fundamental than R.

[17] Remember that $n = \frac{N}{V}$ is the number of primitive lattice cells per unit volume in the crystal, which is not the same as $n_e = \frac{N}{V}$, the number of electrons per unit volume, except for crystals with a monatomic basis of monovalent metals.



A voltage V applied along the wire will generate a current I proportional to ΔA,

$$I = j\Delta A$$

where j is called the current density, the current per unit area in the wire. This voltage will give rise to an electric field of magnitude $E = \frac{V}{d}$, so Ohm's law $V = IR$ can be written

$$Ed = (j\Delta A)\left(\frac{\rho d}{\Delta A}\right) \qquad \Leftrightarrow \qquad E = \rho j.$$

Physicists prefer to work with the conductivity, defined to be the inverse of the resistivity, $\sigma = 1/\rho$, and write $j = \sigma E$. This should really be written as a vector equation, currents have a direction associated with them so j is a vector, as is E. For an isotropic conductor

$$\boxed{\mathbf{j} = \sigma\mathbf{E}.}$$

It is a difficult, but important, task to calculate σ for any given material, metal or semi-conductor, from a knowledge of the microscopic structure at the atomic level. To get some understanding of the underlying physics, consider the force acting on a charge $-e$, such as an electron, due to an applied electric field $\mathbf{E}$: Newton's second law implies

$$\mathbf{F} = m\frac{d\mathbf{v}}{dt} = -e\mathbf{E}.$$

If $\mathbf{E}$ is constant this immediately integrates to

$$\mathbf{p}(t) - \mathbf{p}(0) = -e\mathbf{E}t,$$

where $\mathbf{p} = m\mathbf{v}$ is the electron's momentum. In a perfectly pure crystal the electron would accelerate indefinitely but in a real crystal this acceleration will be impeded by collisions with phonons (at room temperature) or imperfections in the crystal (more important at low temperatures, when phonons are scarce). Suppose the electron is stopped in its tracks by these collisions and denote the average time between collisions by τ, then the average electron velocity will be $\mathbf{v} = \mathbf{a}\tau = -\frac{e}{m}\mathbf{E}\tau$. If the electron number density is n then the current density is proportional to n,

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

from which we derive the DC conductivity in terms of τ,

$$\boxed{\sigma = \frac{n_e e^2 \tau}{m}.} \tag{40}$$

This is called the **Drude** formula. It is useful in that it gives an intuitive understanding of what affects the conductivity, many rapid collisions means τ is small and the conductivity is low, very few collisions means τ is a long time and the conductivity is high. τ is in fact a function of temperature, at room temperature $\tau \propto 1/T$. The conductivity is also proportional to the electron density, which is reasonable — more electron means more current. A quantity which reflects how easily the electrons can move, without reference to the number of electrons is the **mobility**, μ, defined by[18]

$$\mathbf{v} = \mu \mathbf{E} \qquad \Leftrightarrow \qquad \mu = -\frac{\sigma}{n_e e}.$$

The minus sign is because the charge on the electron -s $-e$.

We shall consider AC conductivity in §8.

Wiedemann-Franz law

We can now derive a relation between the electrical and thermal conductivities in a metal, which follows from the fact that it is mobile electrons that are responsible for both. Using our expression for thermal conductivity (26) and the electron specific heat (39), $c_v = \frac{\pi^2}{2}\frac{k_B^2}{\varepsilon_F} T n_e$, where the Fermi energy ε_F is related to the Fermi velocity v_F by $\varepsilon_F = \frac{1}{2} m v_F^2$, gives $\kappa = \frac{c_v v_F^2 \tau}{3} = \frac{\pi^2 n}{3m} k_B T \tau$, we have

$$\boxed{\frac{\kappa}{\sigma} = \frac{\pi^2 k_B^2}{3e^2} T.}$$

The ratio of thermal to electrical conductivities in a metal is proportional to the absolute temperature of the crystal, with a co-efficient given by known physical constants. Historically this relation, known as the Wiedemann-Franz law, was discovered before it was understood why specific heats in metals were so low, 1% of their expected classical value — the thermal conductivity is also low for the same reason, Fermi blocking.

Hall effect

The Hall effect occurs in very thin slabs of conducting material placed in a transverse magnetic field[19] when a current is passed through the slab. Discovered in 1879 it was a very important step in understanding the nature of electric currents.

[18] Not to be confused with the chemical potential.

[19] Strictly speaking **B** is *the magnetic flux density* and the magnetic field is denoted by **H**. In a vacuum they are simply proportional to each other, $\mathbf{B}{=}\mu_0\mathbf{H}$, where μ_0 is the magnetic permeability of the vacuum, $\mu_0{=}4\pi{\times}10^{-7}$ in SI units. At the moment we do not to worry about this distinction and we shall, somewhat incorrectly, refer to **B** as the magnetic field. The distinction will however be important in §8 when we come to discuss magnetic properties of materials.

Consider the effect on the analysis above of including a magnetic field. The force on the electron is given by the Lorentz force law,

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

For a constant magnetic field in the z-direction, $\mathbf{B} = B\hat{\mathbf{z}}$, and a constant electric field perpendicular to $\mathbf{B}$, $\mathbf{E} = E\hat{\mathbf{x}}$, this gives

$$\begin{aligned} m\dot{v}_x &= -eE - ev_yB \\ m\dot{v}_y &= ev_xB \\ m\dot{v}_z &= 0. \end{aligned}$$

Combining the first two of these equations implies

$$\ddot{v}_x = -\frac{e^2B^2}{m^2}v_x,$$

which is the harmonic oscillator equation for v_x with frequency $\omega_c = \frac{eB}{m}$, called the **cyclotron frequency**. This is a characteristic frequency for a charged particle moving in a magnetic field. Solving this equation gives

$$v_x(t) = v_0\cos(\omega_c t)$$

where v_0 is a constant and, using this in the v_y equation above, gives

$$\dot{v}_y(t) = v_0\omega_c\cos(\omega_c t) \qquad \Rightarrow \qquad v_y(t) = v_0\sin(\omega_c t) + a,$$

with a a constant of integration. Then putting this form of $v_y(t)$ into the equation for $\dot{v}_x$ above yields

$$m\dot{v}_x = -m\omega_c v_0 sin(\omega_c t) = -eE - eB\big(v_0\sin(\omega_c t) + a\big),$$

which fixes a to be the ratio $a = -\frac{E}{B}$. Finally $\dot{v}_z = 0$ implies that v_z is a constant and the motion of the electron is given by

$$\mathbf{v}(t) = v_0\big(\cos(\omega_c t)\hat{\mathbf{x}} + \sin(\omega_c t)\hat{\mathbf{y}}\big) - \frac{E}{B}\hat{\mathbf{y}} + v_z\hat{\mathbf{z}}$$

which integrates to

$$\mathbf{r}(t) = \frac{v_0}{\omega_c}\big(\sin(\omega_c t)\hat{\mathbf{x}} - \cos(\omega_c t)\hat{\mathbf{y}}\big) - \frac{E}{B}t\hat{\mathbf{y}} + v_z t\hat{\mathbf{z}},$$

where we have chosen the origin so that $\mathbf{r}(0) = \mathbf{0}$. If otherwise unhindered the electron performs a circular motion in the $x - y$ plane superimposed on a constant drift in the direction $v_z\hat{\mathbf{z}} - \frac{E}{B}\hat{\mathbf{y}}$.

We can now include the effect of collisions by adding a term $\mathbf{F}_c = -\frac{m}{\tau}\mathbf{v}$ to the Lorentz force,

$$m\ddot{\mathbf{v}} = -\frac{m}{\tau}\mathbf{v} - e\big(\mathbf{E} + (\mathbf{v} \times \mathbf{B})\big).$$

Again with $\mathbf{B} = B\hat{\mathbf{z}}$ in the z-direction, but with a more general $\mathbf{E}$, this can be written in components as

$$\begin{aligned} m\dot{v}_x &= -e(E_x + v_y B) - \frac{mv_x}{\tau} \\ m\dot{v}_y &= -e(E_y - v_x B) - \frac{mv_y}{\tau} \\ m\dot{v}_z &= -eE_z - \frac{mv_z}{\tau}. \end{aligned}$$

Rather than solving these equations in complete generality, we just look for a steady state solution with $\dot{\mathbf{v}} = 0$, which will be sufficient for a discussion of DC currents. Such a solution is

$$\begin{aligned} v_x &= -\frac{\tau e}{m}E_x - \omega_c \tau v_y \\ v_y &= -\frac{\tau e}{m}E_y + \omega_c \tau v_x \\ v_z &= -\frac{\tau e}{m}E_z. \end{aligned}$$

Now focus on a thin slab of conducting material in the $x - y$ plane. The electrons are confined to the plane, so we shall assume $v_z = 0$ and $E_z = 0$. Consider a rectangular slab with its edges aligned in the x and y directions. If a current is passed through the slab in the x-direction, then $v_y = 0$ as well and

$$E_x = -\frac{mv_x}{\tau e}, \qquad E_y = \omega_c \frac{m}{e} v_x = -\omega_c \tau E_x$$

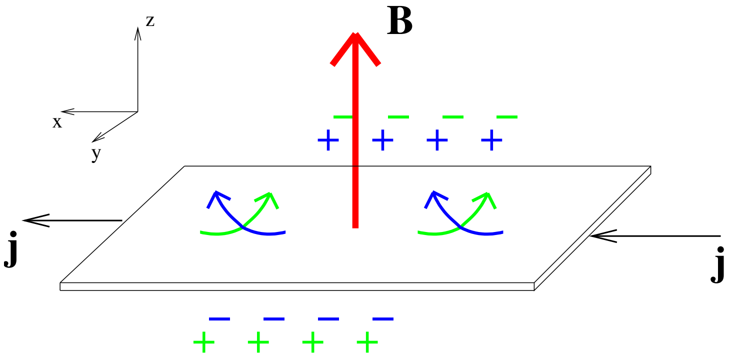
so

$$E_y = -\frac{eB}{m}\tau E_x.$$

The current density is the charge passing unit area in unit time, which is $-e$ times the number of electrons passing unit area in unit time. The latter is $n_e\mathbf{v}$, where n_e is the number of electrons per unit volume, so

$$\mathbf{j} = -en_e\mathbf{v} = \frac{n_e e^2 \tau}{m} E_x \hat{\mathbf{x}}.$$

From this we see that the conductivity $\sigma = \frac{n_e e^2 \tau}{m}$ is just the same as in the $B = 0$ case above, but now we no longer have the vector relation $\mathbf{j} \propto \mathbf{E}$, because $E_y \neq 0$ and there is no component of the current in the y-direction. What is happening here is illustrated below, where the green arrows indicate what the electron trajectories would be if E_y were zero:



As the electrons are forced through the slab by E_x the Lorentz force pushes them in the $-y$ direction and their trajectories are bend toward the edge of the sample, but they cannot escape the confines of the slab so a negative charge builds along the top edge of the slab. At the same time electrons are depleted from the lower edge. This charge generates a voltage in the y-direction which builds up until there is a Coulomb force in the positive y-direction ($E_y \neq 0$) which exactly cancels the Lorentz force due to the magnetic field and the electrons just move in straight lines in the x-direction. This voltage is called the **Hall voltage**. Note that the Hall voltage would have the opposite sign if the particles carrying the current had a positive electric charge (blue arrows above). By measuring the sign of the Hall voltage we can tell that electrons in metals carry a negative charge.

We have

$$j_x = \frac{n_e e^2 \tau}{m} E_x = -\frac{n_e e}{B} E_y \qquad \Rightarrow \qquad E_y = R_H B j_x$$

where

$$\boxed{R_h = -\frac{1}{n_e e}}$$

is intrinsic to the material and is called the **Hall co-efficient**. The sign of the Hall co-efficient depends on the sign of the electric charge on the charge carriers (it is negative for electrons) and we can also obtain n_e directly by measuring R_H. For example silver has $R_H = -9 \times 10^{-11}\ m^3 C^{-1}$ from which $n_e = 7 \times 10^{28}\ m^{-3}$. Some materials do have positive Hall co-efficients, for example aluminium has $R_H = 1.0 \times 10^{-10}\ m^3 C^{-1}$ and so behaves as though the current is being carried by *positive* charges. This apparently anomalous behaviour is explained below in terms of *energy bands*.

Energy Bands and Bloch's Theorem

The crystalline structure of metals puts strong constrains on the form of the electron wave-functions. Again, for simplicity, we consider one-dimension first. Model the motion of an electron moving in a one-dimensional crystal by demanding that the electron moves in a potential that is periodic with period a, $U(x) = U(x + a)$. Denote the electron wave-function by $\psi(x)$. The electron density $n(x) \propto |\psi(x)|^2$ is also periodic, $n(x) = n(x+a)$, so $|\psi(x)|^2 = |\psi(x + a)|^2$, but this does not necessarily mean that $\psi(x)$ itself is periodic with period a, only that it is periodic up to a phase $\psi(x+a) = e^{i\phi}\psi(x)$. However we can impose periodic boundary conditions in ψ over the whole crystal, $\psi(x + \mathcal{N}a) = \psi(x)$ where $\mathcal{N}a$ is the size of the crystal (for large $\mathcal{N}$ the boundary conditions do not affect the behaviour in the interior of the crystal very much).

We now argue that ϕ is independent of position and can only have a discrete set of possible values. The Schrödinger equation for a state with energy E is

$$-\frac{\hbar^2}{2m}\psi''(x) + U(x)\psi(x) = E\psi(x), \tag{41}$$

and lattice periodicity requires

$$\begin{aligned} & -\frac{\hbar^2}{2m}\psi''(x+a) + U(x+a)\psi(x+a) = E\psi(x+a), \\ \Rightarrow \qquad & -\frac{\hbar^2}{2m}\psi''(x+a) + U(x)\psi(x+a) = E\psi(x+a). \end{aligned}$$

In particular $\psi(x)$ and $\psi(x+a)$ satisfy the same second order ODE with the same boundary conditions, hence they are linearly dependent (in general the real and imaginary parts of $\psi(x)$ are linearly independent). Thus $\psi(x+a) = c\psi(x)$ with c a (possibly complex) constant. Since $\psi(x+\mathcal{N}a) = \psi(x)$ we conclude that $c^{\mathcal{N}} = 1$, hence $c = e^{i\phi} = e^{\frac{2\pi i j}{\mathcal{N}}}$ where $j = 0, \ldots \mathcal{N}-1$ is an integer with $\mathcal{N}$ possible values (or, if $\mathcal{N}$ is even, we can use $j = -\frac{\mathcal{N}}{2}, \ldots, \frac{\mathcal{N}}{2}$). The integer j labels different solutions of the Schrödinger equation: let $k = \frac{2\pi j}{\mathcal{N}a}$ then we denote the eigenfunction associate with any particular choice of j by $\psi_k(x)$. These eigenfunctions have the property that

$$\psi_k(x+a) = e^{ika}\psi_k(x). \tag{42}$$

k is a wave-vector and if $j = -\frac{\mathcal{N}}{2}, \ldots, \frac{\mathcal{N}}{2}$ then $-\frac{\pi}{a} \le k \le \frac{\pi}{a}$ and k takes $\mathcal{N}$ discrete values in the first Brillouin zone of the crystal. For large $\mathcal{N}$ the allowed values of k are crowded very close together and approximate a continuum as $\mathcal{N} \to \infty$, but they never stray outside the first Brillouin zone.

An example of a function with the property (42) is e^{ikx}, though this is not a solution of Schrödinger's equation above unless $U = 0$, in which case $E = \frac{\hbar^2 k^2}{2m}$ and $\hbar k$ has the physical interpretation of being the electron's momentum.

It is not possible to solve the Schrödinger equation (41) in closed form for a general periodic potential $U(x)$. Nevertheless the assumed periodicity of the potential allows a simplification and we shall prove a theorem, called **Bloch's theorem**, that the eigenfunctions can always be written in the form[20]

$$\boxed{\psi_k(x) = e^{ikx}B_k(x).} \tag{43}$$

[20] Strictly speaking the one-dimensional version quoted here is **Floquet's theorem**. Bloch proved the three-dimensional version quoted later.

where k lies in the first Brillouin zone and $B_k(x) = B_k(x+a)$ is a periodic function. This automatically satisfies (42).

The proof of Bloch's theorem is instructive as it sheds light on the general structure of the energy spectrum of electrons in a crystal. Let $\psi(x)$ be an eigenfunction of the Schrödinger equation with energy E,

$$\widehat{H}\psi(x) = E\psi(x),$$

where the Hamiltonian operator is $\widehat{H} = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + U(x)$. Any $\psi(x)$ can be expanded as a sum of plane waves

$$\psi(x) = \sum_q b(q)e^{iqx} = \sum_{s=-\infty}^{\infty} b_s e^{\frac{2\pi i s}{\mathcal{N}a}x}, \tag{44}$$

where the co-efficients $b(q) = b_s$ are complex constants. With periodic boundary conditions $\psi(x + \mathcal{N}a) = \psi(x)$, the allowed values of $q = \frac{2\pi s}{\mathcal{N}a}$ are necessarily discrete with s taking on all integer values between $-\infty$ and ∞. Unlike phonons the geometry puts no bound on the electron momentum, electrons can have wavelengths much shorter than the lattice spacing.[21]

In contrast the potential $U(x)$ is periodic with period a, $U(x + a) = U(x)$, so its Fourier decomposition involves only reciprocal lattice vectors $G = \frac{2\pi h}{a}$,

$$U(x) = \sum_{G'} \widetilde{U}(G')e^{iG'x} = \sum_{h'=-\infty}^{\infty} \widetilde{U}_{h'} e^{\frac{2\pi i h'}{a}x}.$$

Let $E^{(0)}(q) = \frac{\hbar^2 q^2}{2m}$ be the energy of a free electron with wave-vector q (*i.e.* the energy for $U = 0$).

The Schrödinger's equation can be written

$$\begin{aligned}
&\sum_q E^{(0)}(q)b(q)e^{iqx} + \sum_{G',q'} \widetilde{U}(G')b(q')e^{i(q'+G')x} = \sum_q E\, b(q)e^{iqx} \\
\Rightarrow \qquad &\sum_q \big(E^{(0)}(q) - E\big)b(q)e^{iqx} + \sum_{G',q} \widetilde{U}(G')b(q-G')e^{iqx} = 0, \qquad \text{where } q = q' + G', \\
\Rightarrow \qquad &\big(E^{(0)}(q) - E\big)b(q) + \sum_{G'} \widetilde{U}(G')b(q-G') = 0.
\end{aligned} \tag{45}$$

Now $\widetilde{U}(G')$ are given complex numbers, they are determined by $U(x)$, so this is a set of linear equations for the unknown numbers $b(q)$. Up till now $q = \frac{2\pi s}{\mathcal{N}a}$ and s could have any integral value but it is often convenient to add a reciprocal lattice vector to q so as to force it into the first Brillouin zone: for any q choose $G = \frac{2\pi h}{a}$ so that $k = q - G$ lies in the first

[21] There is a dynamical limit, though — the smaller the wave-length the more energetic the electrons are and if the wave-length is very small the electrons can be so energetic that they are not confined to the crystal. The happens when the energy of the electron becomes comparable with the work function of the material, typically a few electron volts.

Brillouin zone.[22] To emphasise the distinction between the wavevector k, which lies in the first Brillouin zone, and G, which is a reciprocal lattice vector, we write $b(k+G) = b_k(G)$. With this notation (45) reads

$$\big\{E^{(0)}(k+G) - E\big\}b_k(G) + \sum_{G'} \widetilde{U}(G')b_k(G-G') = 0.$$

Then, with $G'' = G - G'$, we have

$$\boxed{\Big\{E^{(0)}(k+G) - E\Big\}b_k(G) + \sum_{G''} \widetilde{U}(G-G'')b_k(G'') = 0.} \tag{46}$$

This is called the **central equation**, though it is actually a set of coupled linear equations for the $b_k(G)$. Alternatively it can be written, using $G = \frac{2\pi h}{a}$ and the notation $\widetilde{U}(G) = \widetilde{U}_h$,[23]

$$\left\{E^{(0)}\left(k + \frac{2\pi h}{a}\right) - E\right\} b_k(h) + \sum_{h''} \widetilde{U}_{h-h''} b_k(h'') = 0. \tag{47}$$

The sum in (47) is over the integers and this form shows that the equation can be written as an infinite matrix equation

$$\begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & E^{(0)}\big(k+\frac{2\pi}{a}\big) - E(k) & \widetilde{U}_1 & \widetilde{U}_2 & \cdots \\ \cdots & \widetilde{U}_{-1} & E^{(0)}(k) - E(k) & \widetilde{U}_1 & \cdots \\ \cdots & \widetilde{U}_{-2} & \widetilde{U}_{-1} & E^{(0)}\big(k-\frac{2\pi}{a}\big) - E(k) & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ b_k(1) \\ b_k(0) \\ b_k(-1) \\ \vdots \end{pmatrix} = 0. \tag{48}$$

Remember $\widetilde{U}_{-h} = \widetilde{U}^*_h$ for a real potential. For simplicity we have chosen to shift the potential by a constant so that $U_0 = 0$, this simply adds an overall constant to the energy and does not change any physics. In principle this is an infinite matrix equation, but in practice we can cut $|h|$ off at some large but finite value with negligible error. Physically, if the momentum of the electron gets too large it will escape from the crystal anyway and the whole description breaks down.

The eigenvalues are determined by requiring that the matrix has zero determinant, so as to ensure that a non-zero solution, a non-zero eigenvector $\begin{pmatrix} \vdots \\ b_k(1) \\ b_k(0) \\ b_k(-1) \\ \vdots \end{pmatrix}$, exists. For a

[22] For a given q this is a unique decomposition, there is only one k in the first Brillouin zone and one reciprocal lattice vector G that can satisfy this. In 1-dimension this is completely analogous to breaking a fractional number up into an integer and a remainder: h is the integer, obtained by rounding the fraction up or down depending on whether the fractional part is greater than or less than $\frac{1}{2}$, $-\frac{1}{2} < \frac{a}{2\pi}k < \frac{1}{2}$ is the remainder.

[23] With a slight abuse of notation we also write $b_k(G) = b_k(h)$ where $G = \frac{2\pi h}{a}$).

given potential, and hence given Fourier co-efficients $\widetilde{U}_h$, calculating the energy eigenvalues and their associated eigenvectors still involves calculating the determinant of a very large matrix. Fortunately it is often the case that many of the off-diagonal components are very small, in real situations it is usually the case that $|\widetilde{U}_1| >> |\widetilde{U}_2| >> |\widetilde{U}_3| >> \cdots$, but before looking at some explicitly solvable cases we pause to prove Bloch's theorem.

The structure of (48) shows that there are energy eigenfunctions associated with any given value of k (indeed there are many). From (44) and the above analysis we can write an energy eigenfunction associated with any particular value of k as

$$\psi_k(x) = \sum_{h=-\infty}^{\infty} b_k(h) e^{i\left(k+\frac{2\pi h}{a}\right)x} = e^{ikx} \sum_{h=-\infty}^{\infty} b_k(h) e^{\frac{2\pi i h}{a}x} := e^{ikx} B_k(x),$$

where the **Bloch function** $B_k(x) = \sum_{h=-\infty}^{\infty} b_k(h) e^{\frac{2\pi i h}{a}x}$ is, by construction, periodic $B_k(x+a) = B_k(x)$. This is Bloch's theorem, (43).

Of course, since $B_k(x)$ is periodic, it has a Fourier expansion

$$B_k(x) = \sum_G \widetilde{B}_k(G) e^{iGx},$$

with $G = \frac{2\pi h}{a}$ and we see that the wave-function co-efficients, $b_k(G)$ are the Fourier co-efficients of the Bloch function.

Note that $\hbar k$ is *not* the electron momentum, that would correspond to the eigenvalue of the operator $\hat{p} = -i\hbar \frac{d}{dx}$ and $\psi_k(x)$ is not an eigenfunction of $\hat{p}$ in general (unless $U(x) = 0$). Indeed $-i\hbar \frac{d\psi_k(x)}{dx} \neq \hbar k \psi_k(x)$ unless $B_k(x)$ is independent of x. In a crystal electron energy eigenfunctions do not have a specific momentum, a crystal does not have translational invariance under infinitesimal translations, it is only invariant under finite lattice transformations, so momentum is not conserved. The Hamiltonian $\widehat{H}$ does not commute with the momentum operator $\hat{p}$, so these two operators cannot be simultaneously diagonalised — an eigenstate of the Hamiltonian cannot simultaneously be an eigenstate of momentum, it must be a linear combination of different momenta states. Nevertheless $\hbar k$ is a momentum of sorts, it is called the **crystal momentum**, but it is not equal to the electron's momentum in general. We shall give a physical interpretation of the crystal momentum later.

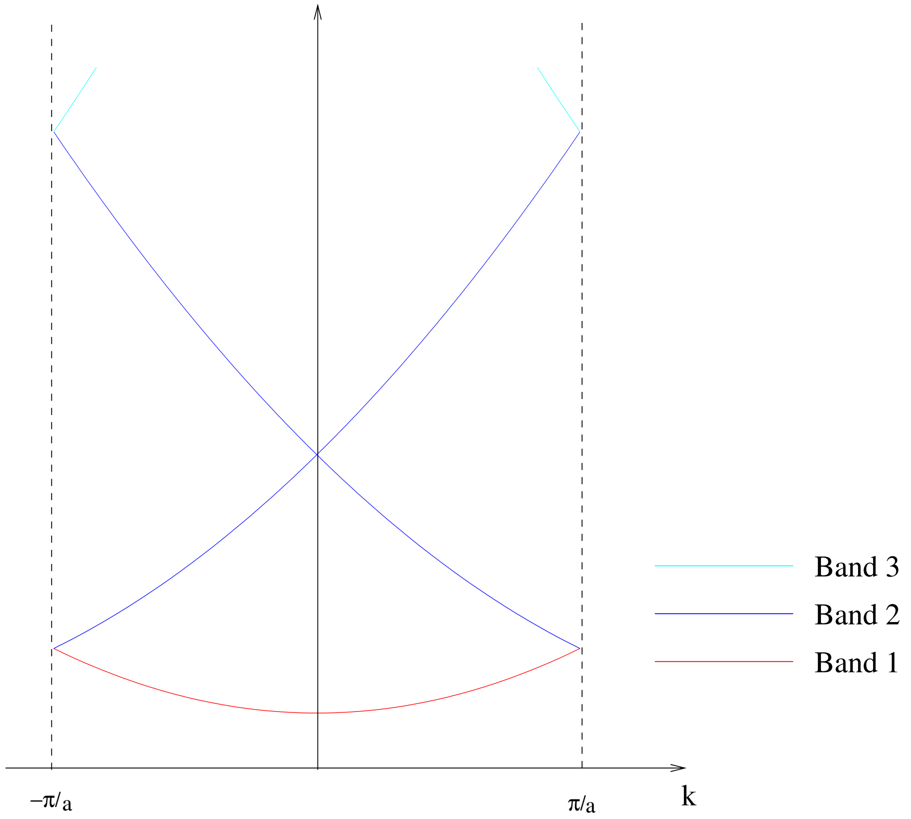
Free Electron Approximation

To get some feeling for the behaviour of the solution of (47) consider the crudest possible approximation, just setting all the $\widetilde{U}_h$ to zero. That is $U(x) = 0$ in the Schrödinger equation and we are dealing with free electrons. To solve (47) choose a specific value of h and k and set

$$E = E^{(0)}\left(k + \frac{2\pi h}{a}\right)^2 = \frac{\hbar^2}{2m}\left(k + \frac{2\pi h}{a}\right)^2,$$

and the eigenvectors for a specific choice of h are $b_k(h) \neq 0$ with all other b_k zero. Remember that k is restricted to lie in the first Brillouin zone $-\frac{\pi}{a} \leq k \leq \frac{\pi}{a}$. The energy spectrum is shown below, the energy is a multivalued function of k and there are different bands of energy labelled by $n = |h|$: band one is $n = 0$, band two is $n = \pm 1$, band three is $n = \pm 2$ etc. The allowed energy in each band is a function of k and the band is identified with a subscript called the band index, $E_n(k)$.

The energy spectrum is the same as that of a single parabola, $E = \frac{\hbar q^2}{2m}$ with $-\infty < q < \infty$, but it is represented as different pieces all with $k = q + \frac{2\pi h}{a}$ and h chosen so that $-\frac{\pi}{a} \leq k \leq \frac{\pi}{a}$. This trick, of knocking q into the first Brillouin zone by adding a reciprocal lattice vector to it, is called the **reduced zone scheme**. It is completely equivalent to a single energy band with $E = \frac{\hbar q^2}{2m}$ and $-\infty < q < \infty$, which is called the **extended zone scheme**. It is a matter of taste which description is used, though the reduced zone scheme is often more convenient.



In three-dimensions the story is essentially the same, but the notation gets a little messier. Solutions of the Shrödinger equation

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{x}) + U(\mathbf{x})\psi(\mathbf{x}) = E\psi(\mathbf{x}), \tag{49}$$

can be expanded in plane waves as

$$\psi(x) = \sum_{\mathbf{q}} b(\mathbf{q}) e^{i\mathbf{q}\cdot\mathbf{x}},$$

with constant co-efficients $b(\mathbf{q})$, and the sum is over three-dimensional wave-vectors $\mathbf{q}$. We can impose periodic boundary conditions in three dimensions, $\psi\Big(\mathbf{x} + \mathcal{N}_i \mathbf{a}_i\Big) = \psi(\mathbf{x})$, where

$i = 1, 2, 3$ and $\mathcal{N}_i$ are three large integers and $\mathbf{a}_i$ are primitive lattice vectors — this is a natural generalisation of periodic boundary conditions in one-dimension and are known as Born - von Karman boundary conditions. In a crystal of volume $V = \mathcal{N}_1\mathcal{N}_2\mathcal{N}_3 V_c = \mathcal{N} V_c$, where V_c is the volume of a primitive cell. These boundary conditions make the allowed values of $\mathbf{q}$ a discrete set, though the allowed values get more and more dense and closer together as $\mathcal{N} \to \infty$.

The Central Equation in 3-dimensions

Let $\psi(\mathbf{x})$ be an eigenfunction of the Schrödinger equation (49) with energy E. The symmetries of the lattice dictate that $U(\mathbf{x})$ is periodic in all three primitive lattice vectors, $U(\mathbf{x}) = U(\mathbf{x}+\mathbf{a}_1) = U(\mathbf{x}+\mathbf{a}_2) = U(\mathbf{x}+\mathbf{a}_3)$, so it can be decomposed into Fourier modes

$$U(\mathbf{x}) = \sum_{\mathbf{G}'} \widetilde{U}(\mathbf{G}') e^{i\mathbf{G}'.\mathbf{x}}$$

where the sum is over all reciprocal lattice vectors and the Fourier co-efficients, $\widetilde{U}(\mathbf{G}')$, are complex numbers. Let $E^{(0)}(\mathbf{q}) = \frac{\hbar^2 \mathbf{q}.\mathbf{q}}{2m}$ be the energy of a free electron with wave-vector $\mathbf{q}$ (*i.e.* the energy for $U=0$).

The Schrödinger equation (49) can be written

$$\sum_{\mathbf{q}} E^{(0)}(\mathbf{q}) b(\mathbf{q}) e^{i\mathbf{q}.\mathbf{x}} + \sum_{\mathbf{G}',\mathbf{q}'} \widetilde{U}(\mathbf{G}') b(\mathbf{q}') e^{i(\mathbf{q}'+\mathbf{G}').\mathbf{x}} = E \sum_{\mathbf{q}} b(\mathbf{q}) e^{i\mathbf{q}.\mathbf{x}}$$

$$\Rightarrow \qquad \sum_{\mathbf{q}} \big(E^{(0)}(\mathbf{q}) - E\big) b(\mathbf{q}) e^{i\mathbf{q}.\mathbf{x}} + \sum_{\mathbf{G}',\mathbf{q}} \widetilde{U}(\mathbf{G}') b(\mathbf{q}-\mathbf{G}') e^{i\mathbf{q}.\mathbf{x}}, = 0 \qquad \text{where } \mathbf{q} = \mathbf{q}'+\mathbf{G}',$$

$$\Rightarrow \qquad \big(E^{(0)}(\mathbf{q}) - E\big) b(\mathbf{q}) + \sum_{\mathbf{G}'} \widetilde{U}(\mathbf{G}') b(\mathbf{q}-\mathbf{G}') = 0. \tag{50}$$

Now $\widetilde{U}(\mathbf{G}')$ are given complex numbers, they are determined by $U(\mathbf{x})$, so this is a set of linear equations for the unknown co-efficients $b(\mathbf{q})$. Up till now $\mathbf{q}$ could be arbitrarily large but, as in the one-dimensional case, it is often convenient to add to it a reciprocal lattice vector so as to force it into the first Brillouin zone: for any $\mathbf{q}$ in our discrete set choose $\mathbf{G}$ so that $\mathbf{k} = \mathbf{q} - \mathbf{G}$ lies in the first Brillouin zone (for a given $\mathbf{q}$ this is a unique decomposition, there is only one $\mathbf{k}$ in the first Brillouin zone and one reciprocal lattice vector $\mathbf{G}$ that can satisfy this). Again to emphasise the distinction between the wave-vector $\mathbf{k}$, which is a wave-vector in the first Brillouin zone, and $\mathbf{G}$, which is a reciprocal lattice vector, we write $b_{\mathbf{k}+\mathbf{G}} = b_{\mathbf{k}}(\mathbf{G})$ and write (50) as

$$\Big\{E^{(0)}(\mathbf{k}+\mathbf{G}) - E\Big\} b_{\mathbf{k}}(\mathbf{G}) + \sum_{\mathbf{G}'} \widetilde{U}(\mathbf{G}') b_{\mathbf{k}}(\mathbf{G}-\mathbf{G}') = 0.$$

Then, with $\mathbf{G}'' = \mathbf{G} - \mathbf{G}'$, we have

$$\boxed{\big\{(E^{(0)}(\mathbf{k}+\mathbf{G}) - E\big\} b_{\mathbf{k}}(\mathbf{G}) + \sum_{\mathbf{G}''} \widetilde{U}(\mathbf{G}-\mathbf{G}'') b_{\mathbf{k}}(\mathbf{G}'') = 0.} \tag{51}$$

This is the three-dimensional **central equation**, the three-dimensional analogue of (46).

The situation in three-dimensions is more complicated, even for free electrons. Then the reciprocal lattice vectors are labelled by three integers h, k and l (there is a clash of notation here, k is an integer in the notation (hkl) denoting a reciprocal lattice vector —

this should not be confused with the length of the wave-vector $|\mathbf{k}|$, hopefully it will be clear which is meant from the context). For a simple cubic lattice, for example,

$$\mathbf{G}_{hkl} = \frac{2\pi}{a}(h\hat{\mathbf{x}} + k\hat{\mathbf{y}} + l\hat{\mathbf{z}})$$

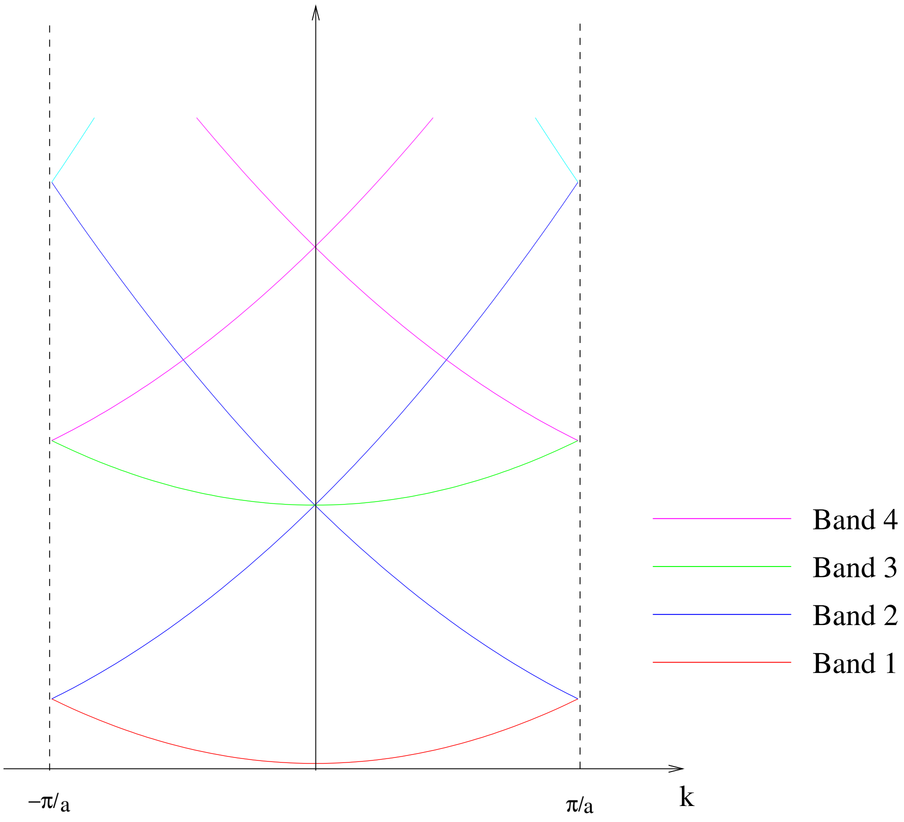
and for a free electron, $U(\mathbf{x}) = 0$, the allowed energies are

$$\begin{aligned} E(\mathbf{k}) = E_{\mathbf{G}}(\mathbf{k}) &= \frac{\hbar^2}{2m}(\mathbf{k}+\mathbf{G})^2 \\ &= \frac{\hbar^2}{2m}\left\{(k_x+G_x)^2 + (k_y+G_y)^2 + (k_z+G_z)^2\right\} \\ &= \frac{\hbar^2}{2m}\left\{\left(k_x+\frac{2\pi h}{a}\right)^2 + \left(k_y+\frac{2\pi k}{a}\right)^2 + \left(k_z+\frac{2\pi l}{a}\right)^2\right\}. \end{aligned}$$

The first few energy bands are listed in the table below, for wave-vectors with $k_y = k_z = 0$:

hkl	$\frac{2m}{\hbar^2}E_{hkl}(0,0,0)$	$\frac{2m}{\hbar^2}E_{hkl}(k_x,0,0)$
000	0	k_x^2
100, $\bar{1}$00	$\left(\frac{2\pi}{a}\right)^2$	$\left(k_x \pm \frac{2\pi}{a}\right)^2$
010, 0$\bar{1}$0, 001, 00$\bar{1}$	$\left(\frac{2\pi}{a}\right)^2$	$k_x^2 + \left(\frac{2\pi}{a}\right)^2$

These energy bands are shown graphically below, in the reduced zone scheme,



You can amuse yourself by including k_y and k_z or by constructing the energy band structure for free electrons in other crystal structures such as FCC or BCC.

Now we shall go beyond the free electron approximation and include the effects of a non-zero potential $U(\mathbf{r})$. For simplicity consider a one-dimensional crystal with central equation (48) and focus on the 2×2 sub-matrix obtained by restricting to $h = 0$ and $h = -1$,

$$\begin{pmatrix} E^{(0)}(k) - E & \widetilde{U}_1 \\ \widetilde{U}_{-1} & E^{(0)}\left(k - \frac{2\pi}{a}\right) - E \end{pmatrix} \begin{pmatrix} b_k(0) \\ b_k(-1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad (52).$$

For a non-trivial solution of the central equation, with $b_k(0)$ and $b_k(-1)$ not both zero, the determinant of the 2×2 matrix must vanish, giving

$$E^2 - \left\{ E^{(0)}(k) + E^{(0)}\left(k - \frac{2\pi}{a}\right)\right\} E + E^{(0)}(k) E^{(0)}\Bigl(k - \frac{2\pi}{a}\Bigr) - \bigl|\widetilde{U}_1\bigr|^2 = 0$$

(remember $\widetilde{U}_{-1} = \widetilde{U}_1^*$ for a real potential). Solving for the eigenvalues E gives $E(k)$, a dispersion relation. Using $E^{(0)}(k) = \frac{\hbar^2}{2m}k^2$, there are two possibilities,

$$E(k) = \frac{\hbar^2}{m}\left\{ \frac{k^2}{2} + \frac{\pi}{a}\left(\frac{\pi}{a} - k\right) \pm \frac{\pi}{a}\sqrt{\left(k - \frac{\pi}{a}\right)^2 + \left(\frac{ma}{\pi\hbar^2}\right)^2 \Bigl|\widetilde{U}_1\Bigr|^2} \right\}. \qquad (53)$$

A consequence of this 2×2 approximation (52) to the full central equation (48) is that the latter is clearly symmetric under $k \to -k$, while (52) is not. The true energy should be an even function of k, but (53) is not. We shall remedy this defect later but, for the moment, just concentrate on $0 \le k \le \frac{\pi}{a}$.

If $0 < \Bigl|\widetilde{U}_1\Bigr| << \frac{\pi^2\hbar^2}{a^2 m}$, then the second term under the square root in (53) is small relative to the first, unless k is close to $\frac{\pi}{a}$, and the two roots are

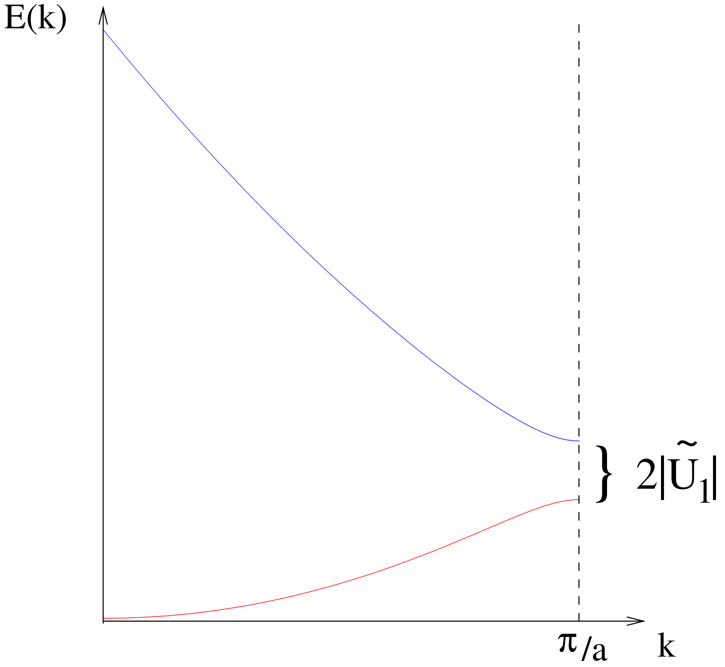
$$E(k) = \begin{cases} \frac{\hbar^2 k^2}{2m} \\ \frac{\hbar^2}{2m}\left(k - \frac{2\pi}{a}\right)^2 + o\Bigl(\frac{|\widetilde{U}_1|}{\pi}\Bigr)^2, & \text{for } k << \frac{\pi}{a}. \end{cases}$$

We see that $\widetilde{U}_1$ doesn't make much difference for $k << \frac{\pi}{a}$, but it has a significant effect when k is on or near the first Brillouin zone boundary. In particular for $k = \frac{\pi}{a}$, equation (53) gives

$$E\Bigl(\frac{\pi}{a}\Bigr) = \frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 \pm \bigl|\widetilde{U}_1\bigr|$$

a 'gap' has opened up in the energy spectrum, at $k = \frac{\pi}{a}$ of magnitude $2\bigl|\widetilde{U}_1\bigr|$. This is called a **band gap** and is a generic feature of solutions of (48), the periodic potential U causes such gaps to open up in the spectrum.

The two solutions (53) are plotted below,



Now focus on the 2×2 sub-matrix of the central equation (48) obtained by restricting to $h=+1$ and $h=-1$.

$$\begin{pmatrix} \frac{\hbar^2}{2m}\left(k+\frac{2\pi}{a}\right)^2 - E & \widetilde{U}_2 \\ \widetilde{U}_{-2} & \frac{\hbar^2}{2m}\left(k-\frac{2\pi}{a}\right)^2 - E \end{pmatrix} \begin{pmatrix} b_k(1) \\ b_k(-1) \end{pmatrix} = 0. \tag{54}$$

Again demanding a non-zero solution for $b_k(1)$ and $b_k(-1)$ requires that the determinant of the associated 2×2 matrix vanishes,

$$\begin{vmatrix} \frac{\hbar^2}{2m}\left(k+\frac{2\pi}{a}\right)^2 - E & \widetilde{U}_2 \\ \widetilde{U}_{-2} & \frac{\hbar^2}{2m}\left(k-\frac{2\pi}{a}\right)^2 - E \end{vmatrix} = 0.$$

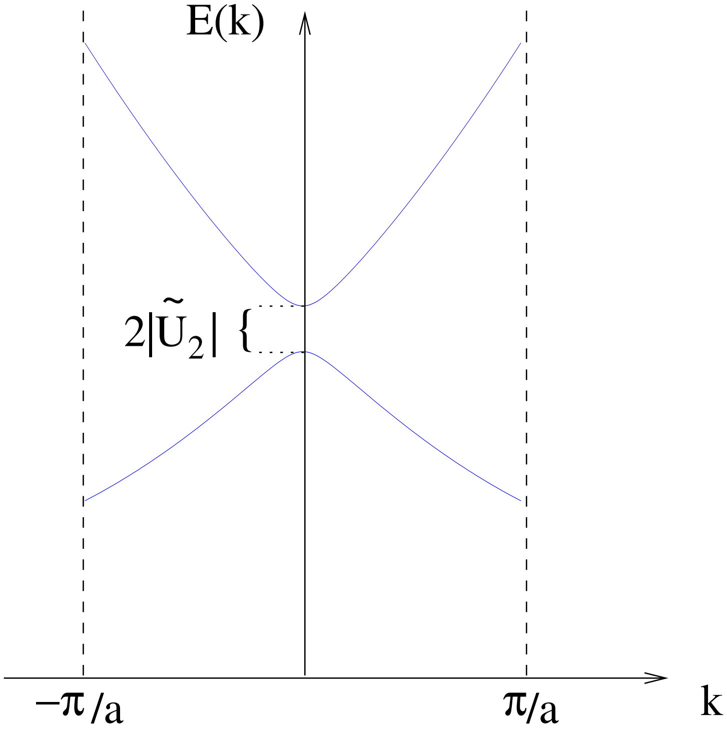
The solutions are

$$E(k) = \frac{\hbar^2}{m}\left\{\frac{k^2}{2} + \frac{2\pi^2}{a^2} \pm \frac{4\pi}{a}\sqrt{k^2 + \left(\frac{ma}{2\pi\hbar^2}\right)^2 |\widetilde{U}_2|^2}\right\}.$$

Now there is a gap between the two bands at $k=0$,

$$E(0) = \frac{2\pi^2\hbar^2}{ma^2} \pm 2|\widetilde{U}_2|.$$

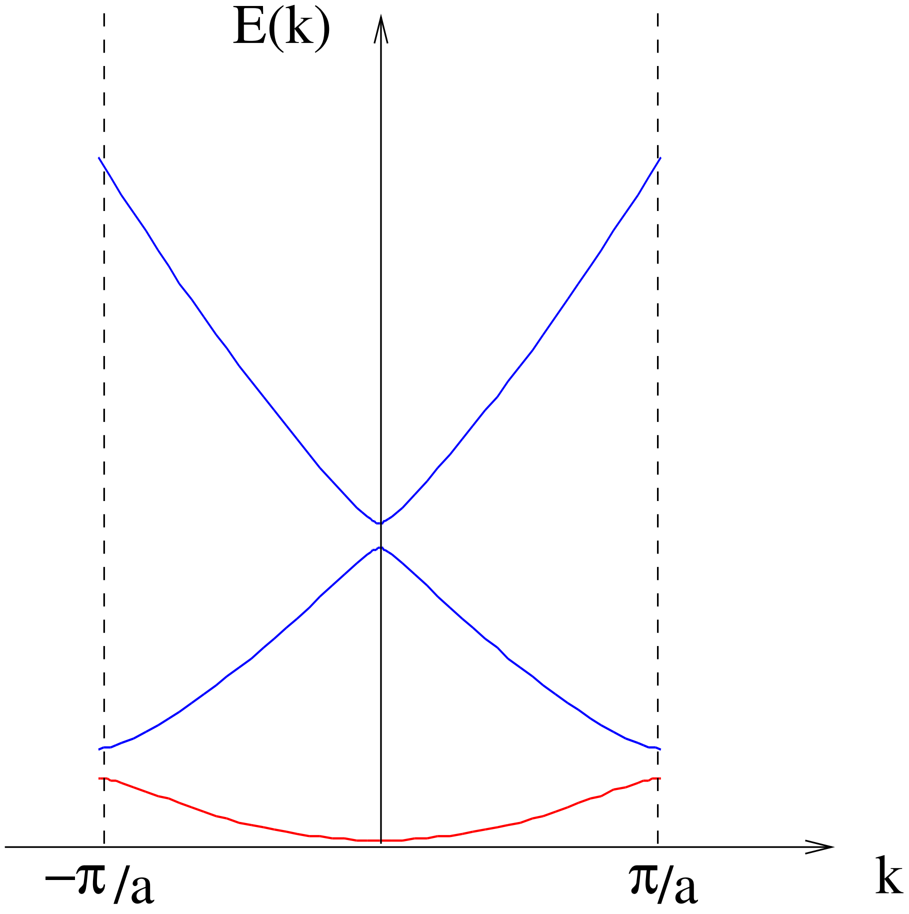
The structure of these two bands is shown below



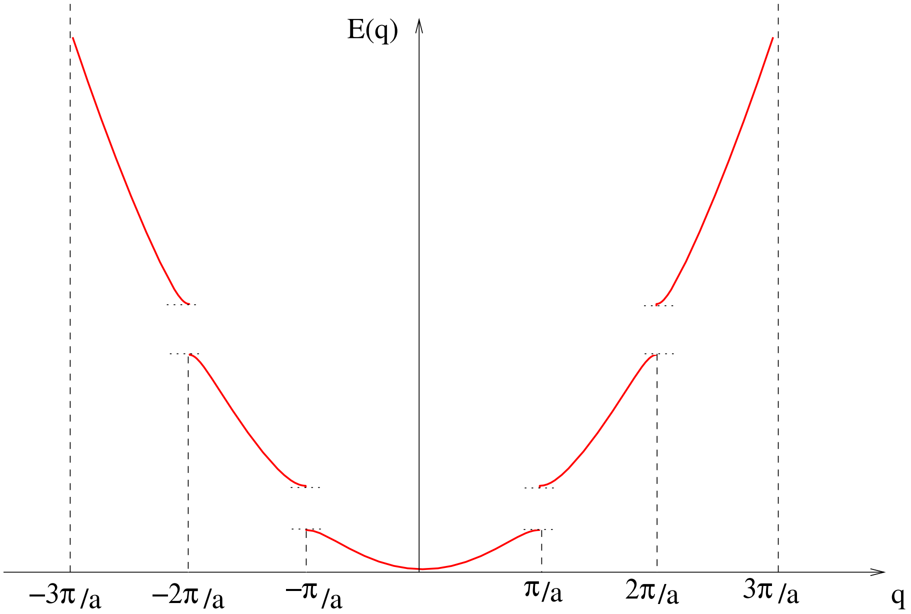
The above results can be combined by focusing on the 3×3 sub-matrix of the central equation (48) obtained by restricting to $h = +1$, 0 and -1. We now demand that the determinant of the matrix

$$\begin{pmatrix} \frac{\hbar^2}{2m}\left(k + \frac{2\pi}{a}\right)^2 - E & \widetilde{U}_1 & \widetilde{U}_2 \\ \widetilde{U}_1^* & \frac{\hbar^2}{2m}k^2 - E & \widetilde{U}_1 \\ \widetilde{U}_2^* & \widetilde{U}_1^* & \frac{\hbar^2}{2m}\left(k - \frac{2\pi}{a}\right)^2 - E \end{pmatrix}$$

must vanish. This gives a cubic equation for $E(k)$, which can be solved analytically to find the three roots, all of which are even functions $E(k) = E(-k)$, but the explicit expressions are not very illuminating. It is more instructive to plot the three functions, using Maple or Mathematica for example. Typically $|\widetilde{U}_2|$ and $|\widetilde{U}_1| << \frac{\pi^2\hbar^2}{a^2m}$ and the spectrum is shown below. There are three bands exhibiting the band gaps found above and, comparing with the figure on page 1, we see that $\widetilde{U}_2$ has split the second energy band of the free spectrum (the blue band on page 1) into two bands here.



In the extended zone scheme the above spectrum looks like a piecewise parabola with jumps at the Brillouin zone boundaries. The gaps can be thought of as being due to reflection of electrons off the zone boundaries.



Sometimes it is convenient to continue the reduced zone scheme to wavevectors outside the first Brillouin zone to get an infinite number of copies of the reduced zone scheme with period $\frac{2\pi}{a}$ — this is called the **periodic zone scheme**.

To gain a deeper understanding of the band structure for electrons in metals consider the energy eigenfunctions in the 2 × 2 matrix equation (52). For simplicity, assume that $\widetilde{U}_1$ is real, $\widetilde{U}_1^* = \widetilde{U}_1$, and negative so that $U(x)$ is has minima at $x = na$, corresponding to attractive ion cores at $x = na$. We can determine the eigenvectors $\begin{pmatrix} b_k(0) \\ b_k(-1) \end{pmatrix}$ by putting the eigenvalues (53) into (52). At the zone boundary, $k = \frac{\pi}{a}$, denote the eigenvalues by $E_\pm$, with $E_+ > E_-$,

$$E_\pm = \frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 \pm |\widetilde{U}_1| \qquad \Rightarrow E^{(0)}\left(\frac{\pi}{a}\right) - E_\pm = \mp|\widetilde{U}_1|$$

and the eigenvectors are determined by

$$\begin{pmatrix} \mp|\widetilde{U}_1| & -|\widetilde{U}_1| \\ -|\widetilde{U}_1| & \mp|\widetilde{U}_1| \end{pmatrix} \begin{pmatrix} b_{\frac{\pi}{a}}(0) \\ b_{\frac{\pi}{a}}(-1) \end{pmatrix} = 0.$$

The solutions are

$$b_{\frac{\pi}{a}}(0) = \mp b_{\frac{\pi}{a}}(-1)$$

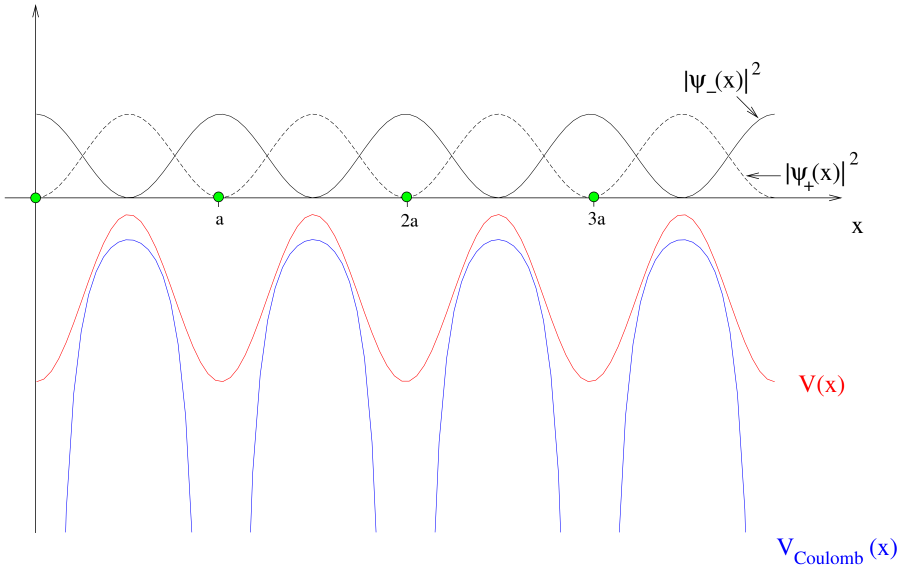
which determines the wave-function $\psi_k(x) = \sum_G b_k(G) e^{i(k+G)x}$at $k = \frac{\pi}{a}$. In this approximation there are only two terms in the sum

$$\psi_{\frac{\pi}{a}}(x) = \sum_{h=-1}^{0} b_{\frac{\pi}{a}}(h) e^{i\left(\frac{\pi}{a} + \frac{2\pi h}{a}\right)x} = b_{\frac{\pi}{a}}(0)\left(e^{\frac{i\pi}{a}} \mp e^{-\frac{i\pi}{a}}\right) := \psi_\pm(x).$$

Hence the energies and associated wave-functions are

$$
\begin{aligned}
E_- &= E^{(0)} - |\widetilde{U}_1|, \qquad \psi_-(x) \propto \cos\left(\frac{\pi x}{a}\right) \\
E_+ &= E^{(0)} + |\widetilde{U}_1|, \qquad \psi_+(x) \propto \sin\left(\frac{\pi x}{a}\right).
\end{aligned}
$$

The energy eigenstates form standing waves as the electrons are reflected off the zone boundaries. The higher energy state, E_+, has the electron wave-function concentrated mid-way between the positive ions, at $x = \left(n + \frac{1}{2}\right)a$, while the lower energy state, E_-, has the electron wave-function concentrated at the positive ions, at $x = na$, where n is an integer. This is because the negatively charged electrons are attracted to the positively charged atomic cores (green in the figure below).



In this illustrative example the attractive Coulomb potential (blue above) is modelled by a periodic potential $V(x) = -V_0 \cos\left(\frac{2\pi x}{a}\right)$ with $\widetilde{U}_i = 0$ for $i \neq \pm 1$ and $\widetilde{U}_1 = -\frac{V_0}{2}$ (red). The lower energy state electron wave-function (solid black line) is attracted to the potential minima at $x = na$ and hence peaks there. The higher energy state wave-function (dotted black line) has a minimum at the potential minima and a maximum at the potential maxima.

In three-dimensions the electron wave-function is periodic in three directions, $\psi(\mathbf{r}) = \psi(\mathbf{r} + \mathcal{N}_1 \boldsymbol{a}_1) = \psi(\mathbf{r} + \mathcal{N}_2 \boldsymbol{a}_2) = \psi(\mathbf{r} + \mathcal{N}_3 \boldsymbol{a}_3)$ and the the details are somewhat more complicated, but the basic concepts are the same. There can be different band structures in different directions.

Fermi surfaces for metals

The dynamics of mobile electrons in crystals is greatly affected by the fact that they are *fermions* and fermions must obey the Pauli exclusion principle — no two fermions can occupy the same quantum state.

Again consider a one-dimensional crystal with length $L = \mathcal{N}a$ and $\mathcal{N}$ sites. The allowed wavevectors in the first Brillouin zone are $k = \frac{2\pi j}{\mathcal{N}a}$, with $j = \pm 1 \ldots, \pm\frac{\mathcal{N}}{2}$ (assume $\mathcal{N}$ is even, when $\mathcal{N}$ is very large this doesn't matter) and they form a discrete set with spacing $\Delta k = \frac{2\pi}{L} = \frac{2\pi}{\mathcal{N}a}$ between allowed states. For free electrons the de Broglie relation between momentum and wavevector in quantum mechanics, $p = \hbar k$, implies that there are also $\mathcal{N}$ allowed momentum states $p = \frac{2\pi\hbar j}{\mathcal{N}a}$, with one momentum state in each interval $\Delta p = \frac{2\pi\hbar}{\mathcal{N}a}$ in momentum space. Since an electron has spin one-half it has two spin states and each momentum state can accommodate at most two electrons, one spin-up and one spin-down. If the energy is an even function of momentum, $E(-p) = E(p)$, and is independent of spin then there are four quantum states for each allowed value of the energy, $\pm p$, spin-up and spin-down.

In a monovalent metal crystal with a monatomic basis (*e.g.* Na or K) only the single electron in the outer electronic shell of the atom is mobile in the metal, so there is one mobile electron per atom or one mobile electron per lattice site. In a divalent metal there would two mobile electrons per atom and, if the basis is monatomic, two mobile electrons per lattice site. Consider a monatomic crystal of a monovalent metal, so there are $\mathcal{N}$ mobile electrons. Imagine starting with no mobile electrons and adding them to the metal one by one. The lowest energy, with four quantum states, is filled first, once that is full, a fifth electron has to go into the next available energy state, that is the next energy level above the lowest one, because it is excluded from the lowest energy by the Pauli exclusion principle. Continue in this way until all $\mathcal{N}$ electrons have been used up and the lowest $\frac{\mathcal{N}}{4}$ energy states are full. The electrons in the topmost filled energy state will have a momentum $|p_F| = \hbar k_F$ with $\frac{k_F}{\Delta k} = \frac{\mathcal{N}}{4}$ where $\Delta k = \frac{2\pi}{\mathcal{N}a}$, so

$$k_F = \frac{\pi}{2a}.$$

The Fermi wavevector, k_F, is half-way to the first Brillouin zone boundary (this factor of two is because of spin degeneracy). The Fermi momentum associated with k_F is

$$p_F := \hbar k_F = \frac{\pi\hbar}{2a},$$

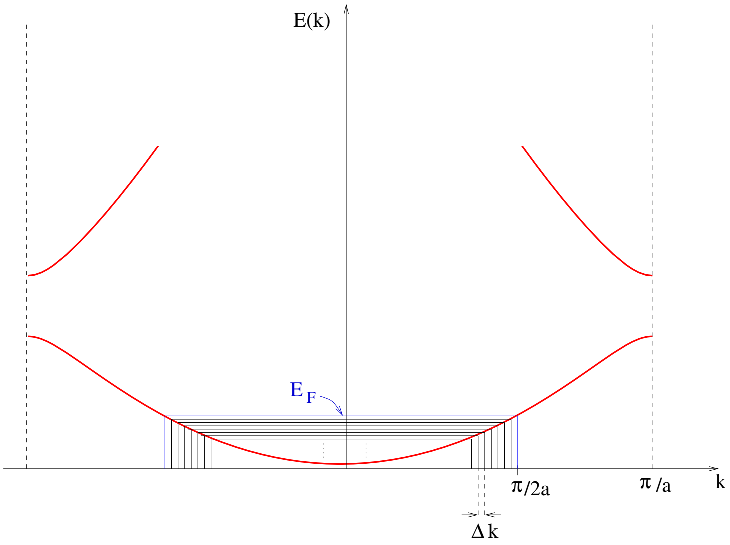
and the Fermi energy is $E_F = E(p_F)$. For free electrons, for example,

$$E_F = \frac{p_F^2}{2m} = \frac{\hbar^2\pi^2}{8ma^2} \tag{55}$$

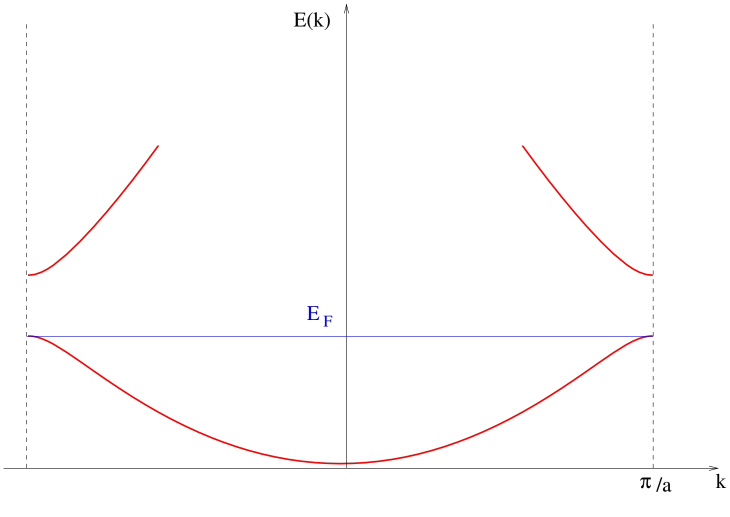
but the concept of the Fermi momentum and the Fermi energy is valid for any dispersion relation.

The energy levels are sketched below. For any finite $\mathcal{N}$ the allowed k-values are always discrete but as $\mathcal{N}$ increases the allowed states crowd closer an closer together as Δk gets smaller and smaller. As the number of atoms, $\mathcal{N}$, increases the number of mobile electrons also increases in just the right way so that the maximum wavenumber, k_F, and the top filled energy level, E_F, are independent of $\mathcal{N}$. We can even let $\mathcal{N} \to \infty$, giving a continuum of states but still with the same Fermi momentum $p_F = \frac{\hbar\pi}{2a}$ and Fermi energy $E_F = E(p_F)$.

The Fermi momentum and Fermi energy are intrinsic to the microscopic crystal structure and are independent of the crystal size — this is an important point, if this were not the case they would not be such useful concepts.

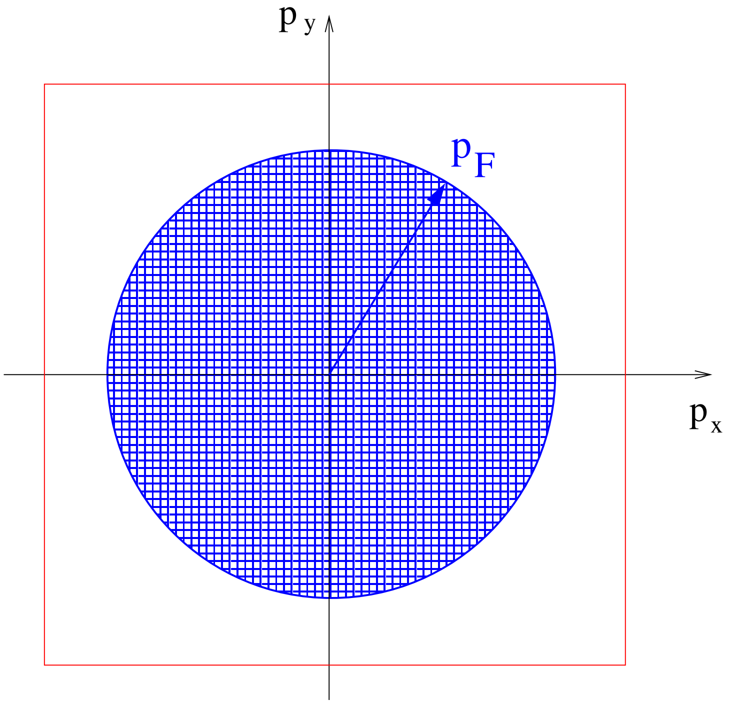


The above sketch is specific to a monovalent, monatomic basis. For a metal with either a divalent monatomic basis or a monovalent diatomic basis there are $2\mathcal{N}$ electrons in a crystal with $\mathcal{N}$ cells, but the same number of momentum states, so the Fermi wavevector reaches all the way to edge of the first Brillouin zone, $k_F = \frac{\pi}{a}$,



In two dimensions there is a whole grid of allowed points in two-dimensional k-space, spanned by the two components of the wavevector, k_x and k_y. Consider a two-dimensional square lattice with lattice spacing a, $\mathcal{N}_1$ cells in the x-direction and $\mathcal{N}_2$ cells in the y-direction, so the total number of cells in the crystal is $\mathcal{N} = \mathcal{N}_1\mathcal{N}_2$ and the area of the crystal is $\mathcal{N}a^2$. For simplicity we shall choose $\mathcal{N}_1 = \mathcal{N}_2$, so $\mathcal{N} = \mathcal{N}_1^2$, but the changes for $\mathcal{N}_1 \neq \mathcal{N}_2$ are pretty straightforward. The spacing of allowed states in both the k_x-direction and the k_y-direction is $\frac{2\pi}{\mathcal{N}_1 a}$, so there is one state per area $\left(\frac{2\pi}{\mathcal{N}_1 a}\right)^2 = \left(\frac{4\pi^2}{\mathcal{N}a^2}\right) := \Delta^2 k$ in k-space. If the dispersion relation is rotationally symmetric and $E(p)$ depends only on the magnitude of the momentum $\mathbf{p}$ and not its direction, then the quantum states will fill up a disc in two-dimensional momentum space. For example for free fermions $E(p) = \frac{p^2}{2m}$ and the filled states form a disc of radius p_F in momentum space. The radius of this disc is independent of $\mathcal{N}$ and, when $\mathcal{N}$ is very large we can think of the distribution of states as a continuum.

For a monovalent monatomic basis the area of this disc (the Fermi disc) in the space of wavevectors is one-half the area of the first Brillouin zone[24] which is $\frac{1}{2}\frac{4\pi^2}{a^2} = \frac{2\pi^2}{a^2}$. Hence the Fermi disc has radius k_F given by $\pi k_F^2 = \frac{2\pi^2}{a^2}$, or $k_F = \frac{\sqrt{2\pi}}{a}$. The reciprocal lattice for a simple cubic lattice with lattice spacing a is another simple cubic lattice with lattice spacing $\frac{2\pi}{a}$ and the first Brillouin zone is a square with $-\frac{\pi}{a} \leq k_x \leq \frac{\pi}{a}$ and $-\frac{\pi}{a} \leq k_y \leq \frac{\pi}{a}$. Since $\frac{\sqrt{2\pi}}{a} = 0.7979\frac{\pi}{a} < \frac{\pi}{a}$ the Fermi disc lies entirely withing the first Brillouin zone.



Note that the number of states with energy E_F is the circumference of the disc divided by the average separation between two states in k-space, which we can take to be $\sqrt{\Delta^2 k}$, and

[24] Again the factor of $\frac{1}{2}$ is due to electron spin — there are two quantum states for every allowed momentum state.

again multiply by two for spin. For large $\mathcal{N}$

$$2\left(\frac{2\pi k_F}{\sqrt{\Delta^2 k}}\right) = 2\left\{\frac{\frac{(2\pi)^{2/3}}{a}}{\left(\frac{2\pi}{\sqrt{\mathcal{N}}a}\right)}\right\} = \frac{2\sqrt{\mathcal{N}}}{(2\pi)^{1/3}},$$

and so grows like $\sqrt{\mathcal{N}} = \frac{L}{a}$, linearly with the size of the crystal. This is in marked contrast to the one-dimensional case were the number of states with energy E_F is always only four, two spin states for $k = \pm\frac{\pi}{a}$.

For a divalent metal there are twice as many mobile electrons for the same number of lattice cells and the area of the Fermi disc is doubled, the radius therefore increases by $\sqrt{2}$ to $k_F = \frac{2\sqrt{\pi}}{a} = 1.128\frac{\pi}{a} > \frac{\pi}{a}$ and the Fermi disc extends outside of the first Brillouin zone into the second zone.

In three dimensions consider a crystal of volume V and a simple cubic lattice structure with lattice spacing a. For simplicity we assume the crystal is a cube of size L, with edges aligned with the crystal axes, so $V = L^3$ and the number of primitive cells is $\mathcal{N} = \frac{L^3}{a^3}$ (for large $\mathcal{N}$ the overall shape of the crystal is not important, it is only $\mathcal{N}$ that matters). There is one momentum state state per volume $\Delta^3 k = \left(\frac{2\pi}{L}\right)^3 = \frac{8\pi^3}{V}$in wavevector space and, for $\mathcal{N}$ large we picture the allowed states as sequentially filling a sphere in k-space of volume $\frac{4\pi}{3}k_F^3$, called the **Fermi sphere**, with

$$\mathcal{N} = 2\frac{\left(\frac{4\pi}{3}k_F^3\right)}{\Delta^3 k} \qquad \Rightarrow \qquad \frac{4\pi k_F^3}{3} = 4\pi^3\frac{\mathcal{N}}{V} \tag{56}$$

(again the factor of 2 is for spin) giving $k_F = (3\pi^2 n_c)^{\frac{1}{3}} = \frac{(3\pi^2)^{\frac{1}{3}}}{a} \approx 0.985\left(\frac{\pi}{a}\right) < \frac{\pi}{a}$ where $n_c := \frac{\mathcal{N}}{V} = \frac{1}{a^3}$ is the number of primitive cells per unit volume. For a free electron dispersion relation, $E = \frac{p^2}{2m}$, this give the Fermi energy

$$E_F = \frac{\hbar^2}{2m}(3\pi^2 n_c)^{\frac{2}{3}} \tag{57}$$

Since $k_F = 0.985\left(\frac{\pi}{a}\right)$ the Fermi surface does not extend as far as the edge of the first Brillouin zone at $k = \frac{\pi}{a}$, but it almost does and any slight distortion of it can easily send part of the Fermi surface into the second Brillouin zone.

The number of states with energy E_F in this case is twice the area of the Fermi sphere divided by the average area of a single state in k-space when projected onto the sphere, which we can take to be $(\Delta^3 k)^{2/3}$ giving

$$2\left\{\frac{\left(\frac{4\pi k_F^2}{3}\right)}{(\Delta^3 k)^{\frac{2}{3}}}\right\} = 2\frac{\left\{\frac{4\pi}{3}\left(\frac{3\pi^2}{a^3}\right)^{2/3}\right\}}{\left(\frac{8\pi^3}{\mathcal{N}a^3}\right)^{2/3}} = 2\left(\frac{\pi}{3}\right)^{\frac{1}{3}}\mathcal{N}^{\frac{2}{3}} = 2\left(\frac{\pi}{3}\right)^{\frac{1}{3}}\left(\frac{L}{a}\right)^2,$$

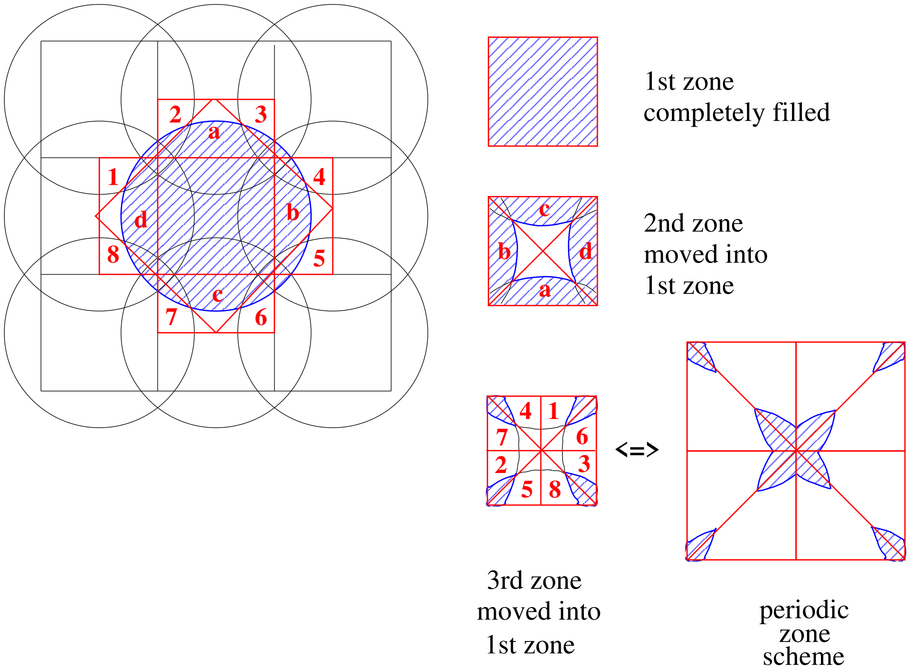
quadratically with the size of the crystal.

For a divalent metal with a monatomic basis, such as magnesium for example, there are two mobile electrons for every primitive cell which doubles the density of electrons and doubles the volume of the Fermi sphere (56), increases k_F by a factor $2^{\frac{1}{3}}$ to $k_F = (6\pi^2 n_c)^{\frac{1}{3}}$ and increases the Fermi energy by a factor $2^{\frac{2}{3}}$.

Since the boundary of the first Brillouin zone for a simple cubic lattice with lattice spacing lies at $k = \frac{\pi}{a}$, the Fermi surface of a monovalent metal $k_F = \frac{(3\pi^2)^{\frac{1}{3}}}{a} \approx \frac{3.09}{a} < \frac{\pi}{a}$ lies inside the first Brillouin zone. For a divalent metal $k_F = \frac{(6\pi^2)^{\frac{1}{3}}}{a} \approx \frac{3.90}{a} > \frac{\pi}{a}$ and the Fermi surface extends into the second Brillouin zone.

The concept of the Fermi surface is of central importance in understanding the dynamics of electrons in metals. A crystal with inter-atomic spacing $a = 4\,\mathring{A}\ = 4 \times 10^{-10}\,m$ has Fermi energy $E_F = 1.5 \times 10^{-17} J \approx 94 eV$, so an electron deep within a Fermi sphere has no empty quantum states near it, its energy must change by at least $\approx 90\,eV$ in order for it to change quantum state. Any interaction with neighbouring electrons, phonons, photons or anything else leaves it completely unaffected unless the energy transfer is of order $90\,eV$. This is a very large energy, for example the thermal energy of an electron at room temperature is of order $k_B T = 4 \times 10^{-21}\,J \approx 0.02\,eV << E_F$. This means that an electron deep within the Fermi sphere is essentially frozen out of all the dynamics. Only electrons near the Fermi surface, in a thin shell of thickness $k_B T$, or about 1% of the radius of the sphere in momentum space, can be thermally excited out of their filled energy state into an available empty energy state nearby, so only about 1% of all electrons are available to transport quantities such as electric current or heat energy.

The shape of the Fermi surface is very important in understanding transport properties of electrons in metals and it is often useful to visualise it in the reduced zone scheme. For a two-dimensional crystal with a monatomic basis of divalent metal atoms $k_F = \frac{2\sqrt{\pi}}{a} > \frac{\pi}{a}$ and the Fermi surface extends into the second, the third and even the fourth Brillouin zone.



In the above picture the Fermi sphere is broken up into pieces lying in different Brillouin zones which are then moved around by reciprocal lattice vector translation to re-assemble the pieces into a single shape in each zone. The four pieces in the second zone are labelled a, b, c and d and the eight pieces in the third zone are labeled $1, \ldots, 8$. In the reduced zone scheme there is a 'hole' of empty states in the middle of the second Brillouin zone states and a star-shaped region of filled states in the third zone. The pictures are drawn assuming that the Fermi surface is a perfect circle, which is a consequence of using the free particle relation between energy and momentum (55). When the periodic potential energy is taken into account band gaps open up but the general shape doesn't change much except that the sharp edges in these pictures are replaced by more rounded edges.

The shape of the Fermi surface can be explored experimentally using magnetic fields. An electron moving in a magnetic field experiences the Lorentz force $\mathbf{F} = -e(\mathbf{v} \times \mathbf{B})$ so $\mathbf{F}.\mathbf{v} = 0$ and the electron's energy does not change, it moves on a surface of constant energy in momentum space. In particular electrons can move around on the Fermi surface, but cannot leave it.

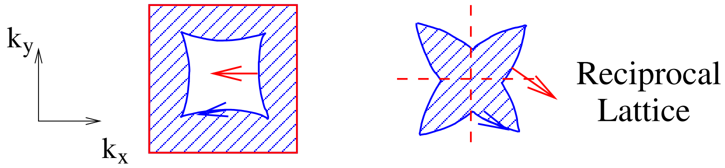
For a given dispersion relation $E(\mathbf{k})$ (not necessarily quadratic) the group velocity of the particles, $\mathbf{v}_g$, has components

$$v_g^i = \frac{1}{\hbar}\frac{\partial E}{\partial k_i} \qquad \text{so} \qquad \mathbf{v}_g = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k}),$$

where $\nabla_{\mathbf{k}}$ denotes the gradient operator in $\mathbf{k}$-space. The Lorentz force is then

$$\mathbf{F} = \dot{\mathbf{p}} = \hbar\dot{\mathbf{k}} = -e\mathbf{v} \times \mathbf{B} = -\frac{e}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k}) \times \mathbf{B},$$

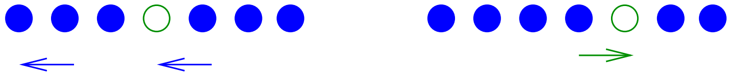
so the force is in a direction perpendicular to both $\mathbf{B}$ and $\nabla_{\mathbf{k}} E(\mathbf{k})$ in $\mathbf{k}$-space. Now the gradient operator acting on $E(\mathbf{k})$ returns a vector $\nabla_{\mathbf{k}} E(\mathbf{k})$ that is normal to surfaces of constant energy. This direction is indicated by the red arrow in the figures below — it points in opposite directions for electrons in a Brillouin zone with a hole in the middle compared to Brillouin zones with a solid island in the middle. This means that, when a magnetic filed is applied in the z-direction, electrons circulate around the holes in a clockwise direction (left-hand figure below, blue arrow) while they circulate around islands in an anti-clockwise (right-hand figure below, blue arrow).



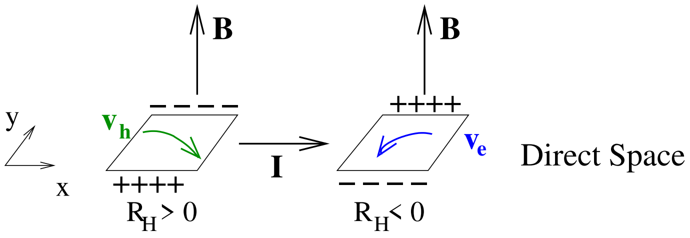
A curious effect here is that, when there is a hole in the middle of the Brillouin zone, the group velocity is in the opposite direction to the wave-vector $\mathbf{k}$ and hence is in the opposite direction to the momentum $\mathbf{p} = \hbar\mathbf{k}$: with $\mathbf{p} = m\mathbf{v}_g$ the electrons behave as though they have a negative mass! Newton's 2nd law can be written as

$$\dot{\mathbf{v}}_g = -\frac{e}{m}(\mathbf{E} + \mathbf{v}_g \times \mathbf{B})$$

with $-\frac{e}{m} = \frac{e}{|m|} > 0$ (remember $e > 0$ and the charge on the electron is $-e$). The forces on an electron with charge $-e < 0$ and mass $m < 0$ are indistinguishable from those on a particle with positive charge $e > 0$ and mass $m > 0$ and the latter picture is a very useful one. Such a particle is called a 'hole' and, when an electric field is applied to drive a current to the right the current can be viewed as being carried by positively charged holes moving to the right, rather than negatively electrons moving to the left. In one dimension the situation looks like the picture below, as the electrons hop one space to the left the hole (open circle) hops to the right. In two dimensions the 'hole' in the Brillouin zone on the left in the figure above moves to the right as all the electrons move to the left.



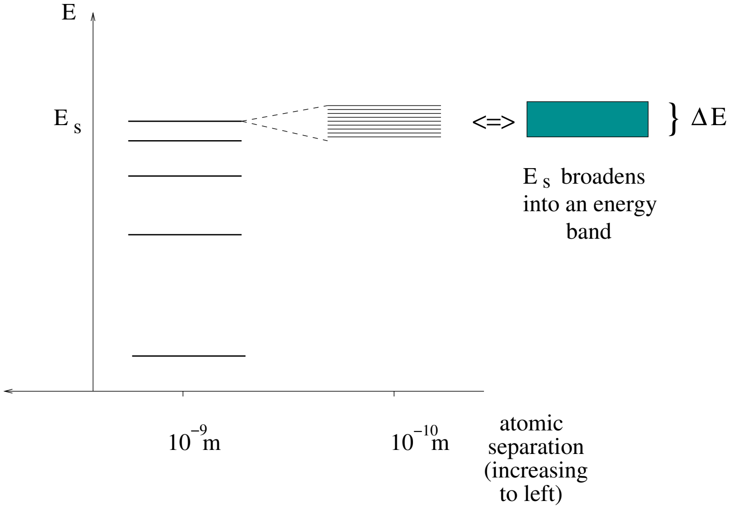
In direct space we get the bottom left hand picture below: the two possibilities, a hole or an island in the Brillouin zone, contribute different signs for the Hall co-efficients.



The overall sign of the Hall co-efficient depends on how many electrons lie on the part of the Fermi surface in the second Brillouin zone and how many lie on the part of the Fermi surface in the third zone, the proportions are the same as just measuring the lengths of the boundaries in each case. Magnesium actually has a negative Hall co-efficient, indicating that electrons win over holes, but aluminium (valence 3), has a positive Hall co-efficient, indicating that there are more electrons orbiting around a hole in the Fermi surface of aluminium than around an island. This explains the mystery of the positive Hall coefficients described on page 1.

Calculation of energy bands — tight binding method

So far we have discovered and analysed the band structure of mobile electrons in metals by considering perturbations of free electrons. We started with free electrons and then modelled the effects of the crystal lattice by introducing a small periodic potential $U(\mathbf{r})$. We can gain further insights by going to the opposite extreme and starting with the electrons tightly bound in $\mathcal{N}$ neutral atoms in free space and then trying to account for perturbations on this model when these atoms are brought into close proximity to each other in a crystal. Suppose the electrons in the outer shell of the free atom are in an s-wave orbital with wave-function $\phi(\mathbf{r})$ and an energy E_s which is non- degenerate. Then this orbital is $\mathcal{N}$-fold degenerate in a system of $\mathcal{N}$ atoms, because there are $\mathcal{N}$ such orbitals each of which has the same energy. When a perturbation is switched on energy states generally tend to have their degeneracy lifted and they split into a family of $\mathcal{N}$ very close energy levels. Suppose we perturb the energy levels of the free atoms by bringing the atoms so close together that the electrons in one atom start to feel forces due to the electrons in ins neighbour — they energy levels are split as shown in the diagram below,


E
E s
<=>
} ΔE
E s broadens into an energy band
10⁻⁹m
10⁻¹⁰m
atomic separation (increasing to left)

In $\mathcal{N}$ is very large, of the order of Avogadro's number 10^{23} in crystals a few millimetres across, then the split energy levels are so close they almost form a continuum, or a band, with a thickness ΔE.

In the **tight binding approximation** we use the free neutral atom orbital wavefunctions for an electron at position $\mathbf{r}$ relative to the centre of the atom, $\phi(\mathbf{r})$, as a basis for wave-functions of the mobile electrons in the metal,

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_j c_{\mathbf{k},j}\phi(\mathbf{r}-\mathbf{r}_j),$$

where the sum is over all the atoms in the crystal (we assume a monatomic basis), $\mathbf{r}_j$ is the position of atom j (lattice sites) and the $c_{\mathbf{k},j}$ are constants. While the $\psi(\mathbf{r}-\mathbf{r}_j)$ are strongly localised around $\mathbf{r} = \mathbf{r}_j$ the electron wave-function, $\psi_k(\mathbf{r})$, extends throughout the whole crystal.

By Bloch's theorem we have

$$\psi_{\mathbf{k}}(\mathbf{r}+\mathbf{L}) = e^{i\mathbf{k}.\mathbf{L}}\psi_{\mathbf{k}}(\mathbf{r})$$

for any lattice vector $\mathbf{L}$. This will be true if the $c_{\mathbf{k},j}$ are of the form $c_{\mathbf{k},j} = \frac{1}{\sqrt{\mathcal{N}}}e^{i\mathbf{k}.\mathbf{r}_j}$, since then

$$\begin{aligned}\psi_{\mathbf{k}}(\mathbf{r}+\mathbf{L}) &= \frac{1}{\sqrt{\mathcal{N}}}\sum_j e^{i\mathbf{k}.\mathbf{r}_j}\phi(\mathbf{r}+\mathbf{L}-\mathbf{r}_j)\\ &= \frac{1}{\sqrt{\mathcal{N}}}\sum_j e^{i\mathbf{k}.(\mathbf{r}_j+\mathbf{L})}\phi(\mathbf{r}-\mathbf{r}_j) \qquad \text{(by Bloch's theorem)}\\ &= e^{i\mathbf{k}.\mathbf{L}}\psi_{\mathbf{k}}(\mathbf{r}).\end{aligned}$$

The energy associated with this wave-function is,
with $\phi_j := \phi(\mathbf{r}-\mathbf{r}_j)$,

$$\begin{aligned}E(\mathbf{k}) =< \psi_{\mathbf{k}}|\,\widehat{H}\,|\psi_{\mathbf{k}} > &= \frac{1}{\mathcal{N}}\sum_{j,j'} e^{i\mathbf{k}.(\mathbf{r}_j-\mathbf{r}_{j'})} < \phi_{j'}|\,\widehat{H}\,|\phi_j >\\ &= \sum_j e^{i\mathbf{k}.(\mathbf{r}_j-\mathbf{r}_0)}\int dV\phi^*(\mathbf{r}-\mathbf{r}_j)H(\mathbf{r})\phi(\mathbf{r}-\mathbf{r}_0), \qquad (58)\end{aligned}$$

where the integral is over the whole crystal and $< \phi_j|\widehat{H}|\phi_i >= \int dV\phi^*(\mathbf{r}-\mathbf{r}_j)H(\mathbf{r})\phi_i(\mathbf{r}-\mathbf{r}_i)$. Assuming only nearest neighbour atoms have a non-zero overlap of their wave-functions define

$$\alpha := -\int dV\phi^*(\mathbf{r}-\mathbf{r}_0)H(\mathbf{r})\phi(\mathbf{r}-\mathbf{r}_0),$$

where α is positive since it is close to the energy of the orbital of an electron when the atoms are free and the electron is bound to the atom, and

$$\int dV\phi^*(\mathbf{r}-\mathbf{r}_j)H(\mathbf{r})\phi(\mathbf{r}-\mathbf{r}_0) := \gamma$$

if $\mathbf{r}_j$ and $\mathbf{r}_0$ are nearest neighbours and all other overlaps are zero. With this approximation the energy eigenvalues (58) are

$$E(\mathbf{k}) = -\alpha - \gamma\sum_j e^{i\mathbf{k}.\mathbf{r}_j} \qquad (59)$$

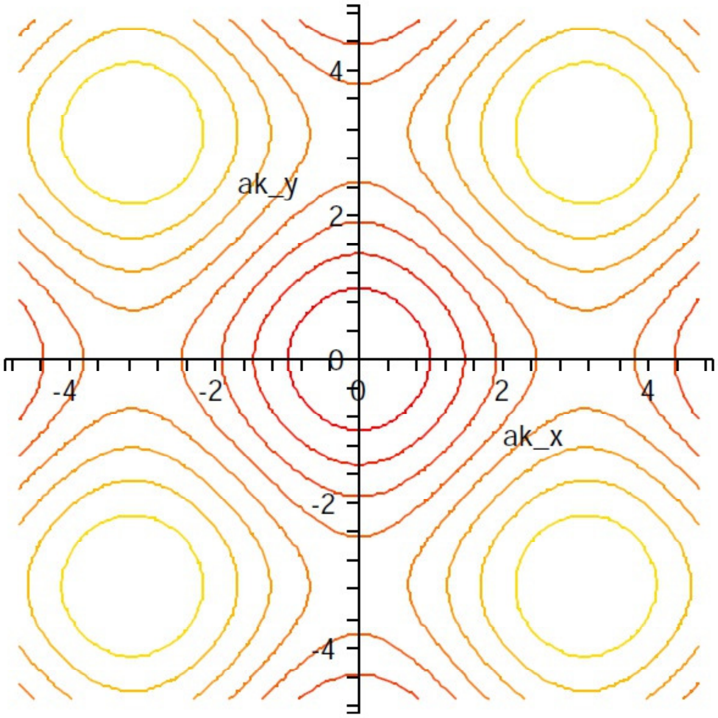
where the sum is only over lattice sites that are nearest neighbours to a reference site at $\mathbf{r}_0$, which can be taken to be the origin. For example in a simple cubic lattice each lattice site has six nearest neighbours, at the origin these are

$$(\pm a, 0), \qquad (0, \pm a, 0) \qquad \text{and} \qquad (0, 0, \pm a)$$

and the energy (59) is

$$E\mathbf{k} = -\alpha - 2\gamma\big(\cos(k_x a) + \cos(k_y a) + \cos(k_z a)\big).$$

A contour plot of $E(\mathbf{k})$ in the $k_x - k_y$ plane, with $k_z = 0$, is shown below and we see that lines of equal energy are distorted from the free electron result, $E(\mathbf{k}) = \frac{\hbar^2}{2m}(k_x^2 + k_y^2)$, for which they would be circles.



In three dimensions the perfect spheres of the free electron approximations are distorted into curved cubical shapes. It was shown above that a spherical Fermi surface for a monatomic basis of monovalent atoms does not extend as far as the edge of the first Brillouin zone, but when it is distorted by including interactions between electronic orbitals in the tight binding approximation it can, as shown below

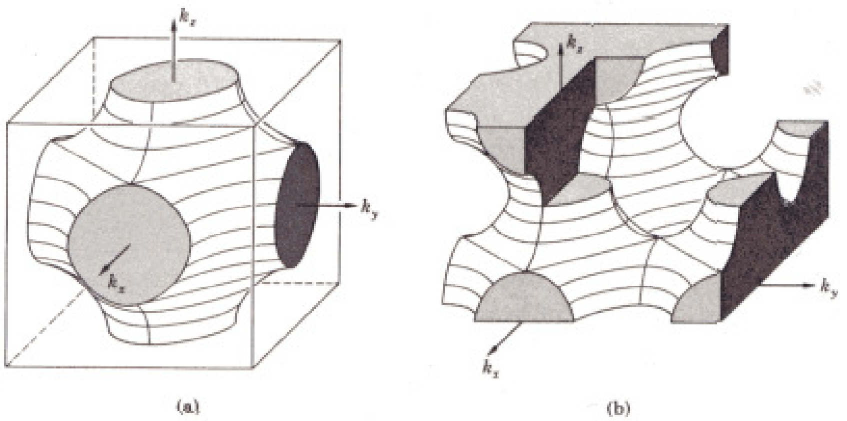
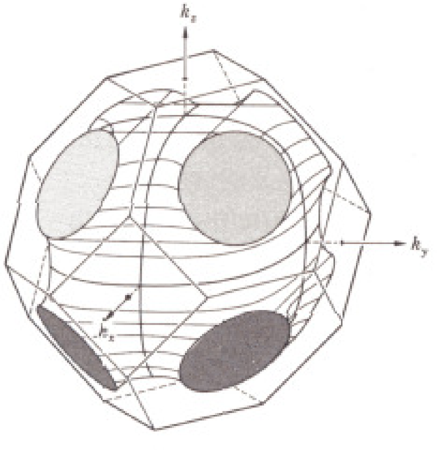


Figure 15 Constant energy surface in the Brillouin zone of a simple cubic lattice, for the assumed energy band $\epsilon_k = -\alpha - 2\gamma(\cos k_xa + \cos k_ya + \cos k_za)$. (a) Constant energy surface $\epsilon = -\alpha$. The filled volume contains one electron per primitive cell. (b) The same surface exhibited in the periodic zone scheme. The connectivity of the orbits is clearly shown. Can you find electron, hole, and open orbits for motion in a magnetic field $B\hat{z}$? (A. Sommerfeld and H. A. Bethe.)

Copper is a valence one metal with a face centred cubic lattice and the Fermi surface looks like this, in relation to a Wigner-Seitz cell of the BCC reciprocal lattice,



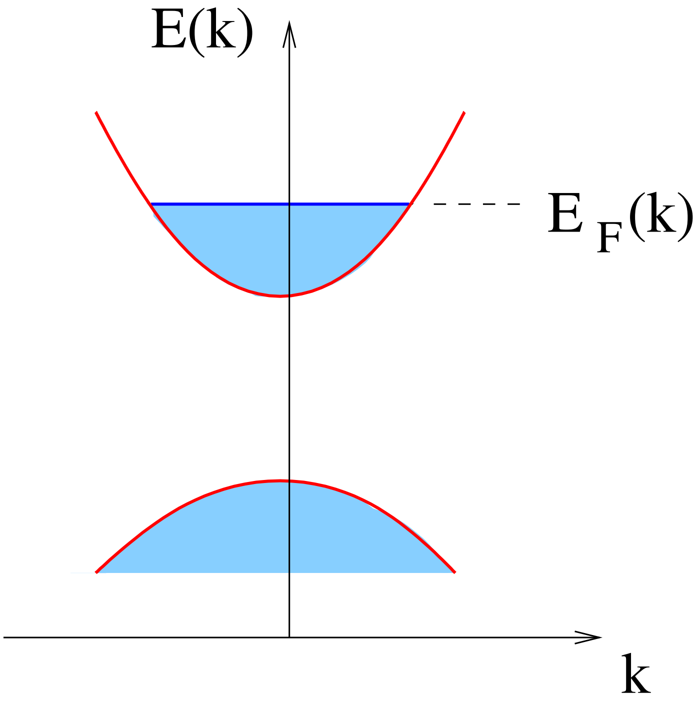
7. Semi-conductors

Broadly speaking conductors are materials that conduct electricity (*e.g.* metals) while insulators do not. More quantitatively conductors have small resistivities (large conductivities) at room temperature: for example silver has $\rho = 1.59 \times 10^{-8}\ \Omega\, m$ and copper has $\rho = 1.68 \times 10^{-8}\ \Omega\, m$. At room temperature the main contributor to resistivity is scattering of electrons off phonons and as the temperature decreases the number of phonons decreases and the resistivity goes down like $\rho \propto T$, but at very low temperatures electrons will scatter off impurities in the metal and the resistivity tends to a constant value, $\rho \to \rho_0$ as $T \to 0$, since the impurity density is independent of T.

Insulators typically have very high resistivities, $\rho \approx 10^{20}\ \Omega\, m$ unless applied voltages get large enough to cause electrical breakdown.

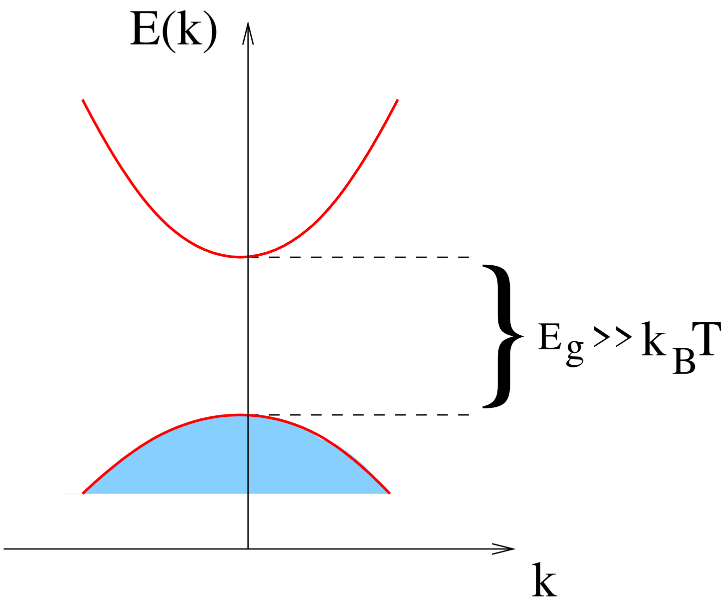
Semi-conductors are materials that have resistivities intermediate between metals and insulators, with a wide range of values $\rho = 10^{-3} \sim 10^{9}\ \Omega\, m$. Furthermore the resistivity in these materials is very sensitive to temperature and impurity density, the resistivity *increases* as the temperature goes down — the opposite behaviour to metals! Examples of semi-conducting materials are silicon and germanium (both valence four elements).

The behaviour of these different types of materials can be understood in terms of their Fermi surfaces and the size of their band gaps.

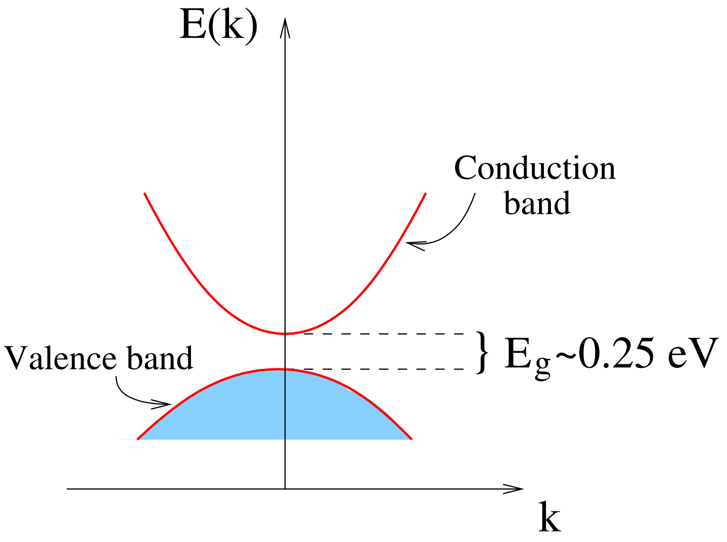


In a metal the Fermi surface is not near a band gap, there are empty states available arbitrarily close to the Fermi surface, as shown above, and there is a large number of states available with the same E_F. Electrons with energy near E_F can move easily when an electric field is applied: indeed we would expect the resistivity to be zero at $T = 0$ were it not for the fact that all real metals inevitably have a certain amount of impurity present, either foreign atoms or imperfect crystal structure, and these impede the electron's progress giving a finite resistivity ρ_0, even at $T = 0$.

At finite temperature there is a strip of width $k_B T$ just above the Fermi surface which electrons can scatter into by thermal excitations, leaving behind an empty state just below the Fermi surface. Thus there is a strip with width of a few $k_B T$ at the Fermi surface in which electrons can scatter off phonons and this scattering also contributes to the resistivity. The width of the strip is $\propto T$ and hence the resistivity is $\propto T$ as mentioned above.

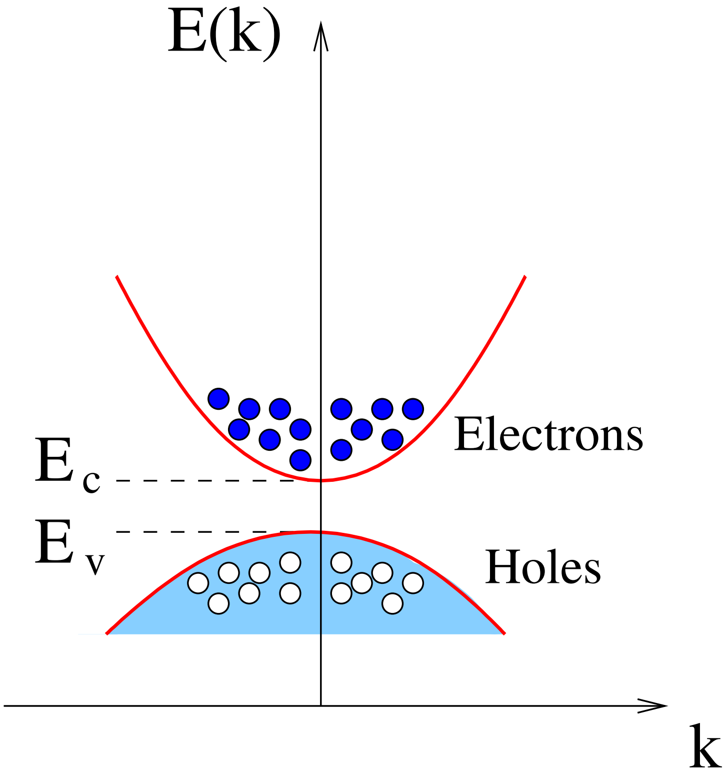


In an insulator, shown above, the Fermi energy coincides with the top of an energy band and the gap above, E_g, is much larger than the thermal energy k_BT for any realistic temperature (*i.e.* below the melting point). There are no nearby empty states available for an electron to move into — the electrons cannot move and the resistivity is essentially infinite. Neither can the electrons be thermally excited into the band above because $E_g >> k_BT$. At room temperature $k_BT \approx 0.025\ eV$ while typical band gaps in insulators are a few eV.



In a semi-conductor the Fermi energy again coincides with the top of an energy band but now the band gap, E_g, is only a few times the thermal energy k_BT. At zero temperature semi-conductors are insulators but at room temperature electrons can be thermally excited into the band above E_F and then there are nearby empty states available and they can move under the influence of an external electric field and carry a current. The higher the temperature the more nearby empty states there are and the lower the resistivity. The upper band is called the **conduction band**, because this is the energy band in which electrons can carry current, while the lower band is called the **valence band**, because this is the band that is exactly filled by virtue of the valence of the material.

When electrons in a semi-conductor are thermally excited into the conduction band they leave behind empty states in the valence band which are called **holes**.



A hole is the absence of an electron. If an electric field is applied to make the electrons in the conduction band move to the left a rightward moving current is generated. When a hole is present the same electric field makes the electron in the valence band which is just to the right of the hole jump into the hole on its left, then the next electron to the right does the same, and so on. The net effect is that the hole moves to the right, contribute to a rightward current — the hole behaves for all practice purposes like a positively charged particle.

Denoting the energy at the bottom of the conduction band by E_c and at the top of the valence band by E_v, so the energy gap is $E_g = E_c - E_v$, we can Taylor expand $E(k)$ around E_c, assuming $E(k) = E(-k)$, for small k

$$E(k) = E_c + \frac{\hbar^2 k^2}{2m_e} + o(k^4).$$

The parameter m_e appearing in the second term of the expansion behaves like a mass for a free particle. In it is natural to *define*

$$m_e := \frac{2}{\hbar^2} \frac{dE}{d(k^2)}\bigg|_{E_c}$$

as being the mass of the electrons in the conduction band, it is called the **effective mass** of the electrons. This depends on the dispersion relation and is often very different to the mass of an isolated free electron. In GaAs, for example, the effective mass $m_e = 0.066 m_{free}$ is a little less the 7% of the free electron mass.

Similarly if we Taylor expand around the top of the valence band

$$E(k) = E_v - \frac{\hbar^2 k^2}{2m_h} + o(k^4),$$

where we have defined

$$m_h := -\left.\frac{2}{\hbar^2}\frac{dE}{d(k^2)}\right|_{E_v}.$$

m_h behaves like an effective mass for holes. Holes and electrons can have different masses in different materials, again in GaAs, for example, $m_h = 0.082 m_{free}$. Notice that the mass of the charge carriers affects the Drude result for the conductivity (40), the contributions of electrons and holes to the conductivity can be different even when their densities are the same.

For small k the density of states for electrons has the free electron form (31), with the substitution of the effective mass and $\varepsilon \to E - E_c$. We also define the density of states for holes,

$$\begin{aligned}\mathcal{D}_e(E) &= \frac{V}{2\pi^2}\frac{(2m_e)^{\frac{3}{2}}}{\hbar^3}(E - E_c)^{\frac{1}{2}}\\ \mathcal{D}_h(E) &= \frac{V}{2\pi^2}\frac{(2m_h)^{\frac{3}{2}}}{\hbar^3}(E - E_v)^{\frac{1}{2}}.\end{aligned}$$

Provided the number of electrons in the conduction band is not too large there will be many more quantum states available at the bottom of the band than there are electrons and we can evaluate the number of electrons in the conduction band with a given energy, as a function of temperature, using Maxwell-Boltzmann statistics

$$f_e = \frac{1}{\exp\left(\frac{E-\mu}{k_B T}\right) + 1} \approx e^{-\frac{(E-\mu)}{k_B T}}, \qquad E > E_c.$$

The electron concentration is then

$$n_e = \frac{N_e}{V} = \frac{1}{V}\int_{E_c}^{\infty} \mathcal{D}_e(E) f_e(E) dE = \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{\frac{3}{2}} e^{\frac{\mu}{k_B T}} \int_{E_c}^{\infty} \sqrt{E - E_c}\exp\left(-\frac{E}{k_B T}\right) dE,$$

where N_e is the total number of electrons. The integral can be evaluated analytically,

$$\int_{E_c}^{\infty} \sqrt{E - E_c}\exp\left(-\frac{E}{k_B T}\right) dE = (k_B T)^{\frac{3}{2}} e^{-\frac{E_c}{k_B T}}\Gamma\Big(\frac{3}{2}\Big) = (k_B T)^{\frac{3}{2}} e^{-\frac{E_c}{k_B T}}\Big(\frac{\sqrt{\pi}}{2}\Big),$$

and

$$n_e = 2\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{\frac{3}{2}}\exp\left(\frac{\mu - E_c}{k_B T}\right).$$

The thermal distribution of holes can be determined by recalling that f_e is the probability of finding an electron with energy E when the temperature is T. For $E < E_v$ every state is either a hole or is filled with an electron so, if f_h is the distribution of holes, $f_e + f_h = 1$ with probability one. Hence

$$f_h = 1 - \frac{1}{\exp\left(\frac{E-\mu}{k_BT}\right)+1} = \frac{1}{\exp\left(\frac{\mu-E}{k_BT}\right)+1} \approx e^{\frac{(E-\mu)}{k_BT}}, \qquad E < E_v.$$

The hole concentration (conventionally denoted p_h, to remind us that holes carry a positive charge) is

$$p_h = \frac{N_e}{V} = \frac{1}{V}\int_{-\infty}^{E_v} \mathcal{D}_h(E) f_h(E) dE = \frac{1}{2\pi^2}\left(\frac{2m_h}{\hbar^2}\right)^{\frac{3}{2}} e^{\frac{\mu}{k_BT}} \int_{-\infty}^{E_v} \sqrt{E_v - E}\exp\left(\frac{E}{k_BT}\right) dE$$
$$= 2\left(\frac{m_h k_BT}{2\pi\hbar^2}\right)^{\frac{3}{2}} \exp\left(\frac{E_v-\mu}{k_BT}\right).$$

The chemical potential disappears from the product

$$\boxed{n_e p_h = 4(m_e m_h)^{\frac{3}{2}}\left(\frac{k_BT}{2\pi\hbar^2}\right)^3 \exp\left(\frac{-E_g}{k_BT}\right),}$$

where $E_g = E_c - E_v$ is the band gap. This is known as the **law of mass action**.[25]

In the semi-conductors described so far $N_e = N_p$, because every electron gives rise to a hole, so $n_e = p_n$ and

$$n_e = p_h = 2(m_e m_h)^{\frac{3}{4}}\left(\frac{k_BT}{2\pi\hbar^2}\right)^{\frac{3}{2}} \exp\left(\frac{-E_g}{2k_BT}\right). \tag{60}$$

Both electrons and holes contribute to the conductivity and the Drude formula (40) gives

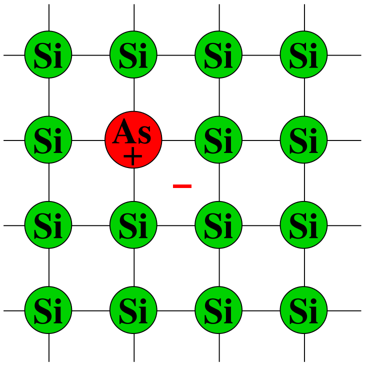
$$\sigma = \frac{n_e e^2 \tau_e}{m_e} + \frac{p_h e^2 \tau_h}{m_h} \tag{61}$$

(electrons and holes can have different scattering times $\tau_e \neq \tau_h$, just as they can have different masses, because their dynamics can be different). The dominant temperature dependence here is the exponential behaviour in (60) and, since $E_g > 0$ by definition, this explains why the conductivity goes up as the temperature goes up — the higher the temperature the more electrons are excited into the conduction band, increasing n_e, and at the same time more holes are created, increasing p_h. The exponential dependence on

[25] In analogy with chemical reactions where the same equation relates the density of two constituent parts of a compound molecule which forms from its constituents with release of energy $\Delta E{=}E_g$.

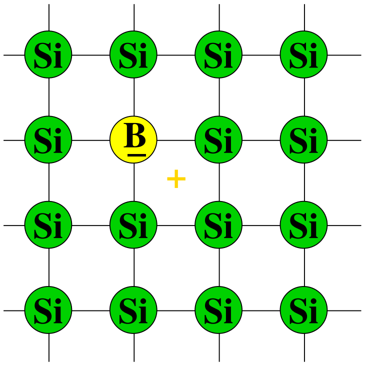
the temperature also explains the conductivity is a very sensitive function of temperature in a semi-conductor.

Semi-conductors with $n_e = p_h$ are called **intrinsic semi-conductors**, but it is also possible to arrange for materials with $n_e \neq p_h$ by deliberately adding impurities, a procedure called **doping**, resulting in **doped semi-conductors**. For example silicon (valence IV) is a semi-conductor. If we replace a silicon atom in a crystal of silicon with an arsenic atom then, since arsenic lies in the column just to the right of silicon in the periodic table and hence has valence V, the arsenic atom has one more electron in its outermost orbital than the silicon atom it replaced had and this electron becomes mobile in the crystal, giving $N_e = N_p + 1$ and contributing to the conductivity . The arsenic atom is called a **donor**, because it donates an electron to become a mobile charge carrier.



If we replace a number of silicon atoms with arsenic atoms, but too many so that the silicon crystal retains its integrity as a crystal, then more generally $N_e > N_p$. By varying the concentration of arsenic we can control the conductivity quite carefully.

Similarly we could replace some silicon atoms with boron atoms, which are in the column immediately to the left of silicon in the periodic table and hence have valence III, then the boron atoms have one electron less in their outer shell than silicon atoms. Mobile electrons from silicon in the crystal then tend to get attracted to the boron and lose their mobility, thus reducing N_e, so that $N_e < N_p$, effectively increasing the number of holes. The boron atoms are called **attractors**.



Semi-conductors doped with donors are called n-type semi-conductors while those doped with acceptors are called p-type.

By growing crystals while controlling the amount of doping it is possible to manufacture semi-conductors with a range of designed conductivities and this is key to the semi-conductor industry.

7. Dielectrics and conductors

If a slab of insulating material is placed in an external electric field the electrons in the outer shell of the atoms will be displaced relative to the positively charged atomic core in the direction of the field, generating a small electric dipole moment $\mathbf{p}$ on each atom: this called *electronic polarisation.* Furthermore, in an ionic crystal like NaCl for example, the positive and negative ions in the crystal will be slightly displaced towards one another, generating another dipole moment, called *ionic polarisation.* Some molecules, *e.g.* water, H_2O, have permanent electric dipole moments due to an asymmetric distribution of the electronic density around the atomic cores, and these tend to line up parallel to any electric field that the atom experiences, giving rise to *dipolar polarisability.* The net effect of any or all of these three types of electric polarisability is that the material will develop an electric polarisation with a dipole moment per unit volume $\mathbf{P}$, simply referred to as the *polarisation.* $\mathbf{P}$ tends to point in the same direction as the total electric field $\mathbf{E}$ inside the material and this tends to reduce the total electric field in the crystal.[26]

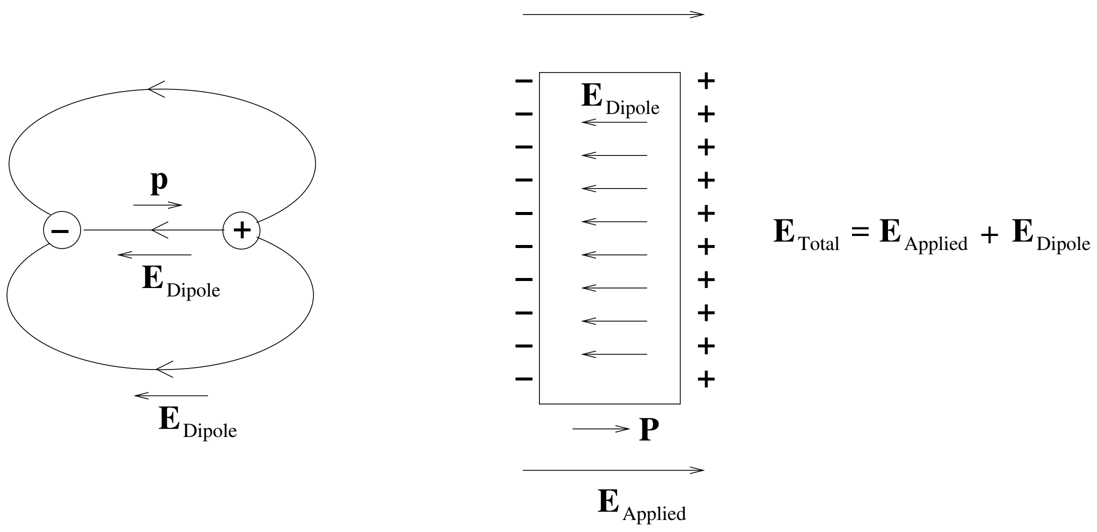
[26] $\mathbf{E}$ is a macroscopic *average* electric field. The microscopic electric field $\mathbf{e}$ is a combination of any externally applied fields $\mathbf{E}_{Applied}$ and any microscopic fields generated by the material in response to the applied field, $\mathbf{e}_{Microscopic}$: $\mathbf{e}$=$\mathbf{E}_{Applied}$+$\mathbf{e}_{Microscopic}$. For the latter the microscopic response field $\mathbf{e}_{Microscopic}$ will vary wildly from place to place and from time to time, as the atoms and molecules in the crystal jiggle around due to their thermal motion (this is obvious for ions, but even if the atoms are electrically neutral they may still have electric dipole moments). Taking the average over a finite time period $\mathcal{T}$, centered around a time t, and many primitive cells, with a total volume $v_{\mathbf{x}}$ centered on a point $\mathbf{x}$, gives an average response field

$$\mathbf{E}_{Medium}(\mathbf{x},t)=\frac{1}{v_{\mathbf{x}}\mathcal{T}}\int_{t'=t-\frac{\mathcal{T}}{2}}^{t'=t+\frac{\mathcal{T}}{2}}\int_{v_{\mathbf{x}}}\mathbf{e}_{Microscopic}(\mathbf{x}',t')d^3\mathbf{x}'dt'$$

is the average of the field generated by the medium's response to $\mathbf{E}_{Applied}$. The average total field is then

$$\mathbf{E}(\mathbf{x},t)=\mathbf{E}_{Applied}(\mathbf{x},t)+\mathbf{E}_{Medium}(\mathbf{x},t).$$

We need to take $v_{\mathbf{x}}$ and T large enough that the fluctuations cancel out and $\mathbf{E}_{Medium}$ does not depend on $v_{\mathbf{x}}$ and T, but keep $v_{\mathbf{x}}$ small enough that it is much less than either the volume of the crystal or the spatial variation of the applied field and keep T much less then the frequency of the applied field.



A slab of such polarisable material responds to an externally applied electric field as though there were a positive surface density on side of the slab and a negative charge density on the other, as shown above. Such a material is called a *dielectric.*

A dipole $\mathbf{p}$ in an electric field has potential energy

$$\mathcal{U} = -\mathbf{p}.\mathbf{E}$$

so having $\mathbf{p}$ parallel to $\mathbf{E}$ is a minimum of the energy. Furthermore if $\mathbf{E}$ is not uniform the potential energy has a gradient

$$\nabla_i \mathcal{U} = -\sum_{j=1}^{3} p_i \nabla_i E_j$$

giving rise to a force

$$F_i = -\nabla_i \mathcal{U} = \sum_{j=1}^{3} p_j \nabla_i E_j$$

which tends to pull the dipole towards regions of higher E_i, dipoles are attracted to regions of stronger electric fields. That dielectrics are attracted to strong electrics fields is famously demonstrated by rubbing a piece of plastic, such as a plastic comb, to generate some electric charge on the plastic and then pieces of paper can be picked up by the electric field generated by the charged plastic. The pieces of paper are electrically neutral but paper is dielectric.

When an external electric field $\mathbf{E}_{Applied}$ is applied the field inside the dielectric is modified by the polarisation of the medium which is the response to $\mathbf{E}_{Applied}$. It is convenient to define the rather boringly named *electric displacement vector*, a better name is the *electric intensity*,[27]

[27] The electric charge inside any 2-dimensional surface S is $\int_S \mathbf{D}.d\mathbf{S}$, not $\epsilon_0 \int_S \mathbf{E}.d\mathbf{S}$, hence electric *intensity.* As shown in any textbook on electromagnetism, an important characteristic of $\mathbf{D}$ is that its normal component is continuous across the interface between any two media — this is not true of $\mathbf{E}$ if there is a charge density on the interface between the media. The tangential component of $\mathbf{E}$ is however continuous.

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \tag{27}$$

(the issue of units is always moot in any discussion of electromagnetic properties in matter, all formulae here are in SI units and a discussion of other (better) systems of units are given in appendix B).

Inside a dielectric we now have a chicken and egg situation, the total field **E** depends on the polarisation **P** but the amount of polarisation depends on the total field **E**. Some assumption about the relation between **E** and **P** is needed to break this impasse. We can work with **D** and **E** rather than **P** and **E** but this does not change the fact that we don't know what **D** is unless we know **P** as a function of **E**. To make progress we need another assumption and in many practical situations it is sufficient to assume that **P** is linear in **E**, this is an acceptable procedure provided **E** is not too strong.

Some materials can sustain a non-zero electric polarisation **P** even in the absence of an external field, such materials are called *pyroelectrics* (lithium niobate, LiNbO_3, is an example of a pyroelectric crystal at room temperature). The name comes from the fact that the permanent dipole moment of such crystals is not immediately obvious, as the surface charges necessary to sustain $\mathbf{P} \neq 0$ tend to get neutralised by charged particles in the atmosphere, from cosmic rays, thunder storms, *etc.* But if the crystal is heated slightly these foreign charged particles evaporate off the surface revealing the underlying dipole moment — hence *pyro*, from the Greek word for heat. If they are heated up too much though they can loose their polarisation and change from pyrolelectrics to ordinary dielectrics at some temperature T_0: this is an example of a phase change, similar in some respects to what happens when water boils. If there is no latent heat associated with this phase change then there can be large fluctuation in the polarisation near T_0 and the material is then called a *ferroelectric*, in analogy with ferromagnets described below (the mineral perovskite that we encountered on page 11 is an example of a ferroelectrict).[28]

In a fluid or a gas it seems reasonable that **P** will be parallel to, and in the same direction as, **E** but this is not necessarily true in a solid, such as a crystal. So for a fluid we write

$$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E} \tag{28}$$

where χ_e is a positive constant known as the *electric susceptibility* of the medium (it is a measure of how susceptible the medium is to being polarised when it is placed in an external electric field). A medium whose polarisation vector satisfies (28) is called a *linear medium.*[29] For such a medium

$$\mathbf{D} = \epsilon_0 (1 + \chi_e) \mathbf{E} = \epsilon \mathbf{E}$$

where

$$\epsilon = \epsilon_0 (1 + \chi_e)$$

[28] Although the applied field is zero this does not mean that $\mathbf{E}$=0, rather $\mathbf{E}$=$\mathbf{E}_{Medium}$. This is discussed in more detail in Appendix C.

[29] The polarisation **P** can depend on temperature and a better definition of the electric susceptibility is $\chi_e = \frac{1}{\epsilon_0} \frac{\partial \mathbf{P}}{\partial E}\Big|_T$, which works even for a non-linear medium, but (28) will be good enough for our purposes.

is called the *electric permittivity of the medium* (in a vacuum $\chi_e = 0$ so $\epsilon = \epsilon_0$ is the same as the electric permittivity of the vacuum).

In a solid $\mathbf{P}$ need not be exactly parallel to $\mathbf{E}$ and it is better to write in components

$$\mathbf{P}_i = \epsilon_0 \sum_{j=1}^{3} (\chi_e)_{ij} \mathbf{E}_j$$

where the electric susceptibility $(\chi_e)_{ij}$ is a 3×3 matrix. In fact $(\chi_e)_{ij}$ is symmetric[30] and like any symmetric matrix can be diagonalised, its eigenvalues representing the three principle directions of the electric susceptibility determined by the underlying crystal structure (for any of the cubic crystals we expect the three principles directions to be equivalent and $(\chi_e)_{ij} = \chi_e \delta_{ij}$, crystals with cubic symmetry have no preferred direction and so cannot be pyroelectrics). For most solid dielectrics the eigenvalues of χ_e are in the range[31] $10 - 200$, for example the static susceptibility for NaCl is $\chi_e = 74$.

In a metal electrons are free to move around inside the material and, as long as the electric field is static and not too strong, they will always conspire to exactly cancel $\mathbf{E}$ inside the material. For a perfect conductor $\chi_e \to \infty$, $\mathbf{D} = \mathbf{P} \neq 0$ and $\mathbf{E} = 0$ inside the material — a conducting material is a perfect dielectric. The electric field inside a conducting medium vanishes even when a static external field is applied because the free electrons in the material arrange themselves on the surface of the conductor so as to generate a non-zero $\mathbf{P}$ which exactly cancels the electric field inside the conductor (this will not be the case for fields that oscillate in time if the frequency is too high for the electrons to respond quickly enough).

In an oscillating electric field the dielectric constant depends on the frequency and a simplistic model of electronic polarisation that gives some insight as to where this frequency comes from (at optical frequencies electronic polarsability gives by far the greatest contribution to the dielectric constant). We view an electron in an atomic orbital as having an energy $\hbar\omega_0$ associated with a characteristic frequency ω_0 and, if it is an excited orbital, will decay with a life-time τ giving rise to damping. We can model the electron bound to the atom as being like a damped harmonic oscillator. For simplicity consider a simple cubic crystal with primitive cells aligned with the x, y and z-axes and an externally generated plane electromagnetic wave passing through it in the x-direction, homogeneous in the y and z-directions. Model an electron in an atom at lattice site sa, with s an integer, as a damped harmonic oscillator with $x_s = sa + \delta x_s(t)$ and $\delta x_s(t) \ll a$ (we assume that $\delta x_s(t)$, the displacement from equilibrium, is much smaller than the size of the atom, otherwise the atom will tear itself apart). In the absence of an electric field the equation of motion for the $\delta_s x(t)$ is

$$m_e \delta\ddot{x}_s = -k\delta x_s - \gamma\delta\dot{x}_s$$

[30] Thermodynamically the free energy density, $f{=}F/V$ where F is the Helmholtz free energy, depends on $\mathbf{E}$ and the temperature. The polarisation per unit volume is $\mathbf{P}_i{=}-\left.\frac{\partial f}{\partial E_i}\right|_T$, so $(\chi_e)_{ij} = -\left.\frac{\partial f}{\partial \mathbf{E}_i \partial \mathbf{E}_j}\right|_T$ is symmetric.

[31] This is in SI units. In cgs units $\chi_{cgs}{=}\frac{1}{4\pi}\chi_{SI}$.

with damping factor $\gamma > 0$. Provided the damping is not too large there is an oscillatory solution $\delta x_s(t) = Re(x_{s,0}e^{-i\tilde{\omega}t})$ with

$$-m_e\tilde{\omega}^2 = -k + i\tilde{\omega}\gamma \qquad \Rightarrow \qquad \tilde{\omega} = \frac{\sqrt{4km_e - \gamma^2} - i\gamma}{2m_e}$$

where $\gamma \ll 2\sqrt{km_e}$ and the sign is chosen so that $\omega_0 = Re(\tilde{\omega}) = \sqrt{\frac{k}{m_e} - \frac{\gamma^2}{4m_e^2}} \approx \sqrt{\frac{k}{m_e}}$ is positive. We see that

$$\delta x_s(t) = Re\big(x_{s,0}e^{-i\omega_0 - \frac{\gamma}{2m_e}t}\big) = x_{s,0}e^{-\frac{\gamma}{2m_e}t}\cos(\omega_0 t)$$

has amplitude $x_{s,0}e^{-\frac{\gamma t}{2m_e}}$, with $x_{s,0} \ll a$, whose square decays with characteristic time-scale $\tau = \frac{m_e}{\gamma}$.

Now include the electric force due to an oscillating electric field in the x-direction $\mathbf{E}(x_s, t) = E_0 e^{-i(\omega t - kx_s)}\hat{\mathbf{x}}$ with E_0 a constant, which can be chosen to be real, and $k = \frac{2\pi}{\lambda}$. For optical frequencies the wavelength of the light is much greater than the size of the atom or the lattice spacing $\lambda \gg a$ and we can replace x_s in the exponent with the average position sa.[32]

The equation of motion for the displacement $\delta x_s(t)$ is now

$$m_e\delta\ddot{x}_s = -m_e\omega_0^2\delta x_s - \gamma\delta\dot{x}_s - eE_0e^{-i(\omega t - ksa)}.$$

There is an oscillating solution of the form $x_s(t) = \delta x_{s,0}e^{-i\omega t}$ with amplitude given by

$$-m_e\omega^2\delta x_{s,0} = -m_e\omega_0^2\delta x_{s,0} + i\omega\gamma\delta x_{x,0} - eE_0e^{iksa}.$$

The amplitude $\delta x_{s,0}$ is complex

$$\delta x_{s,0} = \frac{e}{m_e}\frac{E_0e^{iksa}}{\left\{(\omega^2 - \omega_0^2) + \left(\frac{\gamma}{m_e}\right)i\omega\right\}} = \frac{e}{m_e}\frac{E_0e^{iksa}}{\left\{(\omega^2 - \omega_0^2) + \frac{i\omega}{\tau}\right\}}$$

and the solution is

$$\delta x_s(t) = \delta x_{s,0}e^{-i\omega t} = \frac{e}{m_e}\frac{E_0e^{-i(\omega t - ksa)}}{\left\{(\omega^2 - \omega_0^2) + \frac{i\omega}{\tau}\right\}}.$$

In the presence of the electric field each atom develops a dipole moment

$$p_s = -e(\delta x_s) = -\frac{e^2}{m_e}\frac{E_0e^{-i(\omega t - ksa)}}{\left\{(\omega^2 - \omega_0^2) + \frac{i\omega}{\tau}\right\}}$$

and, if there are n atoms per unit volume, the electric susceptibility is

$$\chi_e = \frac{e^2n}{m_e}\frac{1}{\left\{(\omega_0^2 - \omega^2) - \frac{i\omega}{\tau}\right\}} = \frac{e^2n}{m_e}\frac{(\omega_0^2 - \omega^2) + \frac{i\omega}{\tau}}{\left\{(\omega^2 - \omega_0^2)^2 + \left(\frac{\omega}{\tau}\right)^2\right\}},$$

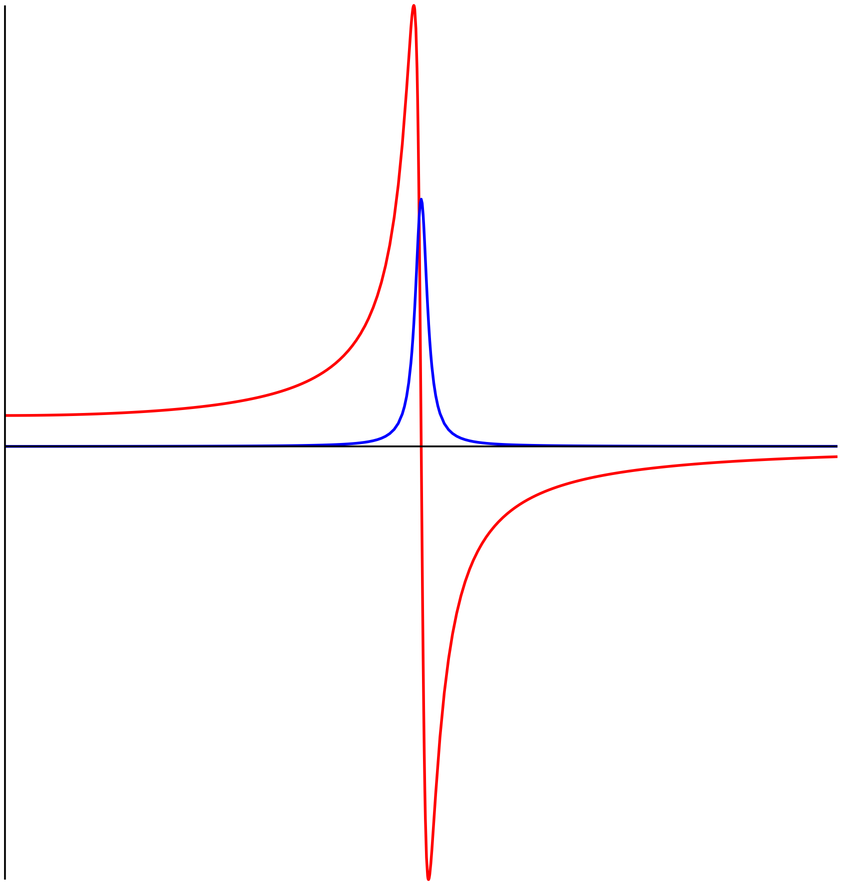
[32] We also cannot let ω become larger than the inverse of the microscopic timescale $\mathcal{T}$ in footnote 28.

it has become *complex*. Before going on the discuss what this means, first consider the static case, $\omega = 0$. Then

$$\chi_e = \frac{e^2 n}{\epsilon_0 m_e \omega_0^2} \qquad \Rightarrow \qquad \epsilon = \epsilon_0 + \frac{e^2 n}{m_e \omega_0^2},$$

the contribution of the electronic polarisation to the electric permittivity depends on how tightly the electrons are bound in the atoms. When ω_0 is large the electrons are tightly bound in the atoms, the atoms are quite rigid and the electronic polarisation susceptibility is low.

For $\omega > 0$ the real and imaginary parts of $\epsilon(\omega)$ are plotted in the figure below for $\tau\omega \gg 1$ (long lived atomic excitations that are only weakly damped), with $Re(\chi_e)$ in red and $Im(\chi_e)$ in blue.



As long as ω is not close to a resonance the real part of $\chi_e(\omega)$ is related to the refractive index of the medium, the speed of light in the medium is given by

$$v^2 = \frac{c^2}{1 + Re(\chi_e)}$$

where $c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$ is the speed of light in a vacuum, hence $\sqrt{1 + Re(\chi_e)}$ is the refractive index. The refractive index, and hence the speed of light, depends on the frequency

— which is why crystals can sometimes exhibit the colours of the rainbow. There is a resonance at $\omega = \omega_0$ where light is absorbed by the atom, its speed drops to a minimum and the light finds it difficult to propagate through the crystal. The change in sign of $Re(\chi_e)$ represents a phase change in the light as it passes through resonance. $Im(\chi_e) > 0$ is an indication of the amount of absorption, near resonance the light loses energy as it excites the atom.

The case $\omega_0 = 0$ is interesting, it corresponds to the electrons not being bound in the atoms at all, they are free to roam around and this is an electrical conductor. For a conductor

$$\epsilon(\omega) = \epsilon_0\big(1 + \chi_e(\omega)\big) = \epsilon_0 + \frac{e^2 n}{m_e}\left(\frac{-\tau^2 + \frac{i\tau}{\omega}}{1 + \omega^2\tau^2}\right).$$

There is a pole at $\omega = 0$ with residue

$$\lim_{\omega \to 0}\big(\omega\chi_e(\omega)\big) = \frac{ie^2 n\tau}{m_e}$$

which is exactly the same form as the Drude conductivity in (40), if we interpret τ as a collision time for ballistic electrons. In fact

$$\sigma(\omega) = -i\omega\chi_e(\omega)$$

is the AC conductivity in general, the real part is the dissipative Ohmic conductivity, the imaginary part is called the refractive conductivity, because the real part of $\chi_e(\omega)$ is responsible for refraction of light.

8. Magnetism and superconductors

Magnetism is a fascinating and very subtle phenomenon in condensed matter. It is fundamentally a quantum mechanical phenomenon which arises from a net magnetic dipole moment due to the electrons in the material, associated with both orbital angular momentum and intrinsic spin. If these magnetic dipole moments can be persuaded to line up then it is possible to generate a macroscopic dipole moment per unit volume, or magnetisation, **M**.

According to classical physics this cannot happen. Maxwell-Boltzmann statistics says that the probability of a system at temperature T being in a state with energy E is proportional to $e^{-E/k_B T}$. An electron with charge $-e$ moving with velocity $\mathbf{v}$ in an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$ experiences the Lorentz force

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

The rate at which it's energy changes is

$$\mathbf{v}.\mathbf{F} = -e\mathbf{v}.\mathbf{E}$$

which is independent of $\mathbf{B}$. Classically magnetic fields do no work and so cannot affect the energy of a particle and cannot affect the thermodynamic state of a system of particles. There is no energetically favourable reason for a magnetisation to develop when a magnetic field is switched on if there was no magnetisation to start with. Nevertheless it is an experimental observation that this often does happen. The explanation requires quantum mechanics.

There are three categories of magnetic materials:

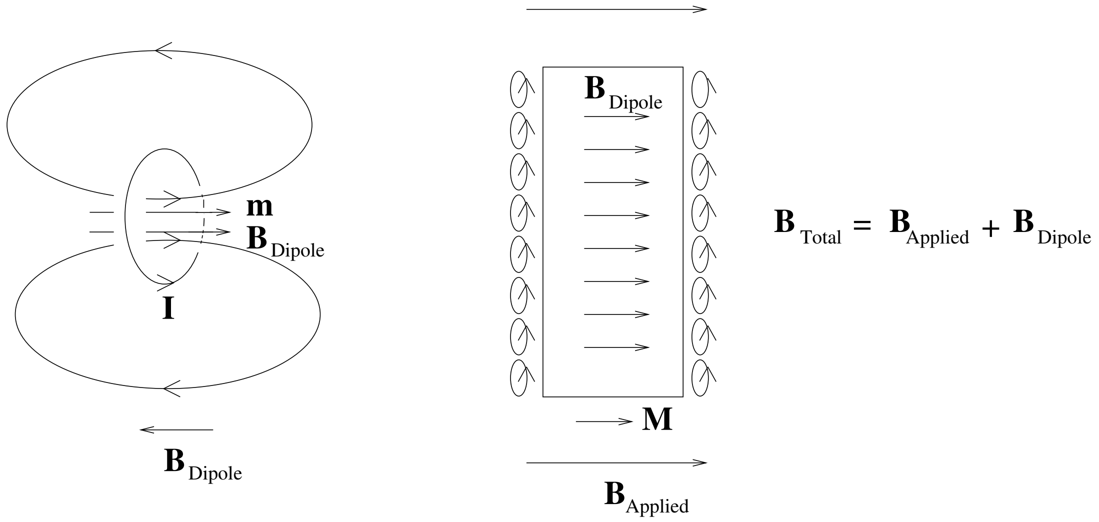
1. Ferromagnets

Perhaps the most familiar aspect of magnetism is a bar magnet, such as a compass needle. These are made of materials that can sustain a permanent macroscopic magnetic moment, or magnetisation **M**, due to a fixed alignment of the microscopic magnetic dipole moments of the electrons in the material even in the absence of any externally applied magnetic field. Such materials are called **ferromagnets**, examples are iron, cobalt and nickel. When ferromagnetic materials are heated up the electron magnetic moments become misaligned and the field disappears at a specific temperature, called the *Curie temperature*, *e.g.* $T_{Curie} = 1043°\mathrm{K}$ for iron.

2. Paramagnets and diamagnets

Even materials that do not have any permanent magnetisation can generate one in the presence of an applied field, due to the microscopic magnetic moments associated with electron motion and intrinsic spin responding to the external field. The magnetisation **M** is defined as the magnetic moment per unit volume generated in the material. If applying an external magnetic field results in a magnetisation **M** in the same direction as the applied field the material is called a **paramagnet**, if it is opposite to the applied field the material is called a **diamagnet**. In a paramagnet the magnetic field in the bulk of the material is larger than the applied field while in a diamagnet it is smaller. Paramagnetic materials tend to be attracted to regions of high magnetic intensity while diamagnetic materials are repelled, though for most materials the effects are much smaller than for the electrical effects in dielectrics.

A classical picture of a paramagnetic is shown below, if we think of the magnetic dipoles as arising from current loops. Note that dipole field intensity $\mathbf{B}_{Dipole}$ inside the loop is in the opposite direction to the field outside it, in contrast to the picture of a dielectric above in which the interior and exterior dipole fields $\mathbf{E}_{Dipole}$ are in the same direction. This is the reason for the sign differences in the magnetic constituent relations below compared to the electric constituent relations above.



Aluminium is paramagnetic as are many salts made with transition metals (but not NaCl, which is diamagnetic). Bismuth is one of the most diamagnetic materials at room temeperature. In the extreme case, when the generated magnetisation completely cancels the applied field, so that $\mathbf{B} = \mathbf{0}$ in the bulk of the material, any applied magnetic intensity is completely expelled from the bulk of the material, we have perfect diamagnetism. This called the *Meissner effect* and when this happens such materials also display zero resistance and are called *superconductors*.

Classically electric currents generate magnetic fields and Ampère's law (in a static situation one of Maxwell's equations) gives

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} \tag{29}$$

where μ_0 is the magnetic permeability of the vacuum. Inside a material the electrons conspire to modify the magnetic permeability and we write[33]

$$\nabla \times \mathbf{B} = \mu \mathbf{j}$$

where μ is the magnetic permeability of the material.

[33] Again $\mathbf{B}$ here is the smoothed average of the total

What is happening here is that individual electron's magnetic dipole moments tend to line up with the magnetic intensity $\mathbf{B}$ and $\mathbf{B}$ can also affect the electron's orbital motion which in turn generates a magnetic dipole moment inside the material. The resulting dipole moment per unit volume $\mathbf{M}$ is called the *magnetisation* of the material. This then modifies $\mathbf{B}$ and, just as for electric fields in a dielectric,

$$\mathbf{B} = \mathbf{B}_{Applied} + \mathbf{B}_{Medium},$$

where $\mathbf{B}_{Applied}$ is the applied magnetic intensity and $\mathbf{B}_{Medium}$ is the field produced by the medium (again $\mathbf{B}_{Medium}$ is a smooth average of the rapidly fluctuating microscopic field, see footnote 29). As for dielectrics it is convenient to define the total *magnetic field* [34]

$$\mathbf{H} = \frac{1}{\mu_0}\mathbf{B} - \mathbf{M}. \tag{30}$$

We have the same chicken and egg situation as for electric polarisation: the magnetisation $\mathbf{M}$ is the response of the medium to the total field $\mathbf{B}$, but the total field depends on both any externally applied field and on $\mathbf{M}$ itself. We need an extra assumption to break this circle and it is often reasonable to assume that $\mathbf{M}$ is proportional to the magnetic field,

$$\mathbf{M} = \chi_m \mathbf{H} \tag{31}$$

where χ_m called the *magnetic susceptibility* of the medium — it is a dimensionless number and is a measure of how susceptible the medium is to being polarised by a magnetic intensity. With this assumption we can eliminate $\mathbf{M}$ in favour of χ_m and write

$$\mathbf{B} = \mu_0(1+\chi_m)\mathbf{H} := \mu\mathbf{H}, \tag{32}$$

where $\mu = \mu_0(1+\chi_m)$ is the magnetic susceptibility of the medium.

Paramagnetic materials have $\chi_m > 0$ and diamagnetic materials have $-1 \leq \chi_m < 0$. For a given $\mathbf{H}$ paramagnetic materials have a larger value of $\mathbf{B}$ than in a vacuum while diamagnetic materials have a smaller value of $\mathbf{B}$. For $\chi_m \to -1$ we have a perfect diamagnet in which $\mathbf{B} = 0$ for any $\mathbf{H}$.[32] We can now avoid all mention of magnetisation and write Ampère's Law as

$$\nabla \times \mathbf{B} = \mu \mathbf{j} \quad \Rightarrow \quad \nabla \times \mathbf{H} = \mathbf{j}$$

[34] The total magnetic flux Φ through any 2-dimensional surface S is $\Phi = \int_S \mathbf{B}.d\mathbf{S}$, not $\frac{1}{\mu_0}\int_S \mathbf{H}.d\mathbf{S}$, hence magnetic *intensity*. An important characteristic of $\mathbf{H}$ is that its tangential component is continuous across the interface between any two media — this is not true of $\mathbf{B}$ if there is a current density on the interface between the media. The normal component of $\mathbf{B}$ is however continuous. Thus the electric intensity $\mathbf{D}$ and the magnetic intensity $\mathbf{B}$ both have continuous normal components to any interface while the electric field $\mathbf{E}$ and the magnetic field $\mathbf{H}$ both have continuous tangential components.

[32] Perfect diamagnetism is associated with zero electrical resistance and this is a superconductor. This requires low temperatures and there is a maximum value of $\mathbf{H}$ above which perfect diamagnatism is destroyed. Some textbooks define $\mathbf{M} = \frac{\chi_m}{\mu_0}\mathbf{B}$, which would be more in line with the electric definition (28), and would give $\mathbf{B} = \frac{\mu_0}{(1-\chi_m)}\mathbf{H}$, but in most substances $|\chi_m| << 1$ so there is no practical difference between this and (32) (superconductors are an exception, perfect diamagnetism would be $\chi_m \to -\infty$ with this convention). Thermodynamically $\chi_M(T)$ can depend on temperature and a better definition is $\chi_m = \frac{\partial \mathbf{M}}{\partial \mathbf{H}}\Big|_T$, which is also good even for a non-linear medium, but we do not require that level of sophistication here.

where

$$\begin{aligned} &\mu > \mu_0 \quad \text{for paramagnetic materials,}\\ &\mu < \mu_0 \quad \text{for diamagnetic materials,}\\ &\mu = 0 \quad\ \ \text{for a superconductor.} \end{aligned}$$

1. Insulators

In an insulating material electrons are bound to atoms and the magnetic properties of the material relate to the spin and orbital characteristics of the electrons in the outermost shell of the atoms or ions in the material.

• *Langevin diamagnetism*

If the outer electron shell is filled (*e.g.* noble gases Ne, Ar, Xe) then the orbital angular momentum $\mathbf{L}$ and the total electron spin $\mathbf{S}$ are zero (all electron spins are paired, up with down, and cancel), hence the total electron angular momentum $\hbar\mathbf{J} = \hbar\mathbf{L}+\hbar\mathbf{S} = 0$. Applying a magnetic intensity will then force the electrons to circulate with the cyclotron frequency

$$\omega_B = \frac{eB}{2m_e}.$$

If there are Z electrons in the outer shell this will generate a circular current

$$\mathbf{I} = (-Ze)\frac{\omega_B}{2\pi} = -\left(\frac{Ze^2}{4\pi m_e}\right)\mathbf{B}.$$

A current circulating around a loop of area $\mathcal{A}$ generates a magnetic dipole moment

$$\mathbf{m} = I\mathcal{A}.$$

Taking $\mathcal{A} = \pi r_e^2$, where r_e is the radius of the outer electronic shell is too naive, but we can replace r_e^2 with the quantum mechanical expectation value $< x^2 > + < y^2 >$ of the electron's position in the orbital plane perpendicular to $\mathbf{B}$. Since

$$< r^2 >=< x^2 > + < y^2 > + < z^2 >$$

we can set

$$< x^2 > + < y^2 >= \frac{2}{3} < r^2 >,$$

where $< r^2 >$ is the quantum mechanical expectation value of the square of the electron's distance from the central atomic nucleus, from which

$$\mathbf{m} = -\frac{Ze^2}{6m_e} < r^2 > \mathbf{B}.$$

Experimentally the molar susceptibility, the magnetic susceptibility of one mole of material, is often quoted. A mole contains Avogadros number $N_A = 6 \times 10^{23}$ of atoms so the magnetisation of one mole, $\mathbf{M}_M$, is

$$\mathbf{M}_M = N_A\mathbf{m} = -\frac{Ze^2 N_A}{6m_e} < r^2 > \mathbf{B}$$

and, assuming $|\chi_m| \ll 1$, (32) then gives the molar magnetic susceptibility

$$\chi_M = -\frac{\mu_0 e^2 N_A}{6m_e} Z < r^2 > \tag{33}$$

which is is negative, the material is expected to be diamagnetic. This is known as *Langevin diamagnetism.*

For most atoms or ions $< r^2 >$ is approximately the Bohr radius $r_0 = \frac{\hbar^2}{me^2} = 5.3 \times 10^{-11}$m. For example solid argon, with $Z = 18$, crystalises at 84°K and equation (33), with $< r^2 >= r_0^2$, gives[33]

$$\chi_M = -1.8 \times 10^{-10} \mathrm{m}^3/\mathrm{mol}.$$

The experimental value is about 30% larger than this at

$$\chi_{Exp} = -2.4 \times 10^{-10} \mathrm{m}^3/\mathrm{mol}.$$

These numbers are very small but they depend on the choice of units for the volume. **M** in (33) is the magnetisation per unit volume and χ_m is the magnetic susceptibility per unit volume (sometimes called the *volume susceptibility*) which, despite its name, is dimensionless,[34]

$$\chi_m = -\frac{\mu_0 e^2 n}{6m_e} Z < r^2 >, \tag{34}$$

where n is the number of atoms per unit volume. Being dimensionless it gives a better idea of the magnitude of the susceptibility. Argon crystallises into a FCC lattice with lattice spacing $a = 5.25$Å so

$$n = 4 \times \left(\frac{1}{5.25 \times 10^{-10}} \right)^2$$

(there are 4 lattice points in each conventional call of a FCC lattice) and equation (34) gives

$$\chi_m = -8.2 \times 10^{-6},$$

diamagnetism is generally a very small effect with $\chi_m \sim 10^{-5}$ (with the exception of superconductors).

• **J** $\neq 0$, *Curie's law*

If the total angular momentum of the outermost electrons $\hbar \mathbf{J} \neq 0$ then a different effect comes into play and susceptibilities tend to be positive (paramagnetism) and larger. The total orbital angular momentum of the outer electrons $\hbar \mathbf{L}$ generates a magnetic dipole moment

$$\mathbf{m_L} = -\gamma\hbar\mathbf{L} = -\mu_B \mathbf{L}$$

[33] This is in SI units, in cgs units χ_M is measured in cm^3/mol and $\chi_{M,cgs} = \frac{10^6}{4\pi}\chi_{M,SI}$.

[34] This is easily seen by writing $\chi_m = -\frac{4\pi\epsilon_0\mu_0}{6m_e}\left(\frac{e^2}{4\pi\epsilon_0 r}\right) Znr{<}r^2{>} = -\frac{2Z}{3}\left(\frac{1}{m_e c^2}\right)\left(\frac{e^2}{4\pi\epsilon_0 r}\right) nr{<}r^2{>}$: $m_e c^2$ and $\frac{e^2}{4\pi\epsilon_0 r}$ are both energies and $nr{<}r^2{>}$ is dimensionless.

where $\gamma = \frac{e^2}{2m}$ is called the *gyromagnetic ratio* and $\mu_B = \frac{e\hbar}{2m_e}$ is the *Bohr magneton.* [35] The total spin $\mathbf{S}$ of the outer electrons generates a magnetic dipole moment

$$\mathbf{m_S} = -2\hbar\mathbf{S} = -2\mu_B\mathbf{S},$$

the intrinsic gyromagnetic ratio of an electron is 2 — a fact whose explanation lies in the relativistic theory of the electron that we do not have time to go into here. The total magnetic moment of the outer electrons is the sum of these

$$\mathbf{m} = \mathbf{m_L} + \mathbf{m_S} = -\mu_B(\mathbf{L} + 2\mathbf{S}).$$

In the presence of an external field $\mathbf{B}$ there is a torque on a magnetic dipole it wants to minimise the classical potential energy

$$\mathcal{U} = -\mathbf{m}.\mathbf{B} \tag{35}$$

by changing the direction of $\mathbf{m}$ so that it lines up with the field. But quantum mechanically $\mathbf{m}$ is related to angular momentum and conservation of angular momentum does not allow the direction of $\mathbf{m}$ to change arbitrarily. Instead, if the magnetic intensity is not too strong, the dipoles precesses around the direction of the angular momentum

$$\mathbf{J} = \mathbf{L} + \mathbf{S}$$

and we should calculate the time averaged value of $\mathbf{m}$ by projecting it onto the direction of $\mathbf{J}$. Of course quantum mechanically $\mathbf{J}$ does not have a specific direction, we cannot simultaneously specify J_x, J_y and J_z because they do not commute as quantum mechanical operators. But the application of an external magnetic intensity explicitly breaks rotational symmetry and specifies a particular direction, which we shall choose to be the z-direction,

$$\mathbf{B} = B\hat{\mathbf{z}}.$$

The average value of the magnetisation in the $\mathbf{J}$-direction is

$$< \mathbf{m} >_{\mathbf{J}} = -\mu_B \frac{(\mathbf{L} + 2\mathbf{S}).\mathbf{J}}{J}$$

and quantum mechanically the energy of the dipole $\mathbf{m}$ in the field $\mathbf{B}$ is

$$\mathcal{U} = - < \mathbf{m} >_{\mathbf{J}} \frac{\mathbf{B}.\mathbf{J}}{J} = \mu_B \frac{\big\{(\mathbf{L} + 2\mathbf{S}).\mathbf{J}\big\} J_z}{J^2} B.$$

The net effect is to give the atom or ion a magnetic dipole moment

$$m = -\mu_B \frac{\big\{(\mathbf{L} + 2\mathbf{S}).\mathbf{J}\big\} J_z}{J^2}$$

[35] μ_B=9.24×10^{-24} m^{2}Coulomb^{-2}=9.24×10^{-24}Joules/Tesla.

which should be evaluated quantum mechanically. Since

$$\begin{aligned} \mathbf{J}^2 &= (\mathbf{L}+\mathbf{S}).(\mathbf{L}+\mathbf{S}) = \mathbf{L}^2 + \mathbf{S}^2 + 2\mathbf{L}.\mathbf{S} \\ \Rightarrow \qquad \mathbf{L}.\mathbf{S} &= \frac{1}{2}(\mathbf{J}^2 - \mathbf{L}^2 - \mathbf{S}^2) \end{aligned}$$

and

$$(\mathbf{L}+2\mathbf{S}).\mathbf{J} = (\mathbf{L}+2\mathbf{S}).(\mathbf{L}+\mathbf{S}) = \mathbf{L}^2 + 2\mathbf{S}^2 + 3\mathbf{L}.\mathbf{S},$$

where $\mathbf{J}^2 = J(J+1)$, $\mathbf{L}^2 = L(L+1)$ and $\mathbf{S}^2 = S(S+1)$, we have

$$m = -g\mu_B J_z$$

with

$$g = \frac{3J(J+1) - L(L+1) + S(S+1)}{2J(J+1)} \tag{36}$$

and J_z is quantised with $2J+1$ values, $J_z = -J, -J+1, \ldots, J-1, J$. g is called the *Landé g-factor*, it reduces to 1 when $S = 0$ and $J = L$ and to 2 when $L = 0$ and $J = S$.

In summary the allowed energies of the atom or ion in a magnetic intensity are

$$\mathcal{U}_{J_z} = g\mu_B J_z B.$$

At finite temperature the lower energy levels will be more populated than the higher energy according to the Boltzmann factors $e^{-\mathcal{U}_{J_z}/k_BT}$ and, if n is the number of atoms or ions per unit volume, the magnetisation per unit volume will be given by

$$M = g\mu_B n \frac{\sum_{J_z=-J}^{J} J_z e^{-\mathcal{U}_J}}{\sum_{J=-J_z}^{J_z} e^{-\mathcal{U}_J}} B.$$

This can be evaluated by noting that the free energy per atom F can be obtained from[36]

$$e^{-F/k_BT} = \sum_{J_z=-J}^{J} e^{-\mathcal{U}_J} = \sum_{J_z=-J}^{J} e^{-g\mu_B J_z B} = \frac{\sinh\left(\frac{(J+1)g\mu_B B}{2k_BT}\right)}{\sinh\left(\frac{g\mu_B B}{2k_BT}\right)}$$

and

$$M = -n\frac{\partial F}{\partial B} = ng\mu_B J B_J\left(\frac{g\mu_B JB}{k_BT}\right) \tag{37}$$

[36] We use $(1 - x^{n+1}) = (1-x)(1 + x + \cdots x^n)$ so

$$\sum_{n=-J}^{J} e^{-nx} = e^{Jx}\sum_{n=0}^{2J} e^{-nx} = e^{Jx}\frac{(1 - e^{-(2J+1)x})}{(1 - e^{-x})} = \frac{e^{(J+\frac{1}{2})x} - e^{-(J+\frac{1}{2})x}}{e^{\frac{x}{2}} - e^{-\frac{x}{2}}} = \frac{\sinh\bigl((J+\frac{1}{2})x\bigr)}{\sinh\bigl(\frac{x}{2}\bigr)}.$$

where

$$B_J(x) = \left(\frac{J+\frac{1}{2}}{J}\right)\coth\left(\frac{(J+\frac{1}{2})x}{J}\right) - \frac{1}{2J}\coth\left(\frac{x}{2J}\right) \tag{38}$$

is known as the *Brillouin function*,[37] with $x = \frac{g\mu_B JB}{k_BT}$.

At very low temperatures $\coth(x) \to 1$ as $T \to 0$ and, again assuming that $\chi_m \ll 1$,

$$M \quad \longrightarrow \quad \mu_0 n g \mu_B J,$$

each atom or ion is aligned and the magnetisation is saturated at its maximum value. At normal temperatures it is more common to have $k_BT \gg g\mu JB$ in which case

$$\coth x = \frac{1}{x} + \frac{x}{3} + O(x^3)$$

and

$$B_J(x) \approx \frac{J+1}{3J}$$

and (37), again with $|\chi_m| \ll 1$, leads to **Curie's law**,

$$\chi_m = \mu_0 n \frac{(g\mu_B)^2}{3}\frac{J(J+1)}{k_BT} := \frac{\mu_0 n p^2 \mu_B^2}{3k_BT}, \tag{39}$$

where $p = g\sqrt{J(J+1)}$. The paramagnetic susceptibility decreases like $\sim 1/T$ as the temperature is increased. In insulators paramagnetic susceptibilities at room temperature are typically of order $10^{-2} \sim 10^{-3}$, two or three orders of magnitude greater that typical diamagnetic susceptibilities. If paramagnetism is present it will dominate, insulating materials generally only exhibit diamagnetism when any paramagnetic effects are completely absent.

p can be calculated from (36) and gives reasonable agreement with experimental measurements for rare earth ions. For triply ionised Cerium, Ce^{3+}, for example, there is one electron in the outer shell with $S = 1/2$, $L = 3$ and $J = 5/2$ ($^4f_2$ in atomic spectroscopy notation) so (36) predicts $p = 2.54$ to be compared to the experimental value of 2.4.

The calculation here has assumed that the outer electrons of the atom or ion are in the same configuration in a crystal as they are in the free atom or ion, but this is not always the case. Sometimes the crystal environment can affect the orbital motion of the electrons and change p. This tends to happen with transition metal ions and is called *quenching*, because the orbital angular momentum is reduced below that of the free atom or ion. Doubly ionised copper, for example, has outer electrons that would be expected to have $L = 2$ for a free ion but behave as though $L = 0$ in crystals, which changes the value of p (Cu^{++} has lost it's outer electrons and is not a metal, it is a common ion in salts that are insulators, such as copper sulphate and copper chloride).

[37] This is a standard notation, but unfortunately B now appears in these formulae representing four different things! Brillouin in the Brillouin function B_J, Bohr in the Bohr magneton μ_B, Boltzmann in Boltzmann's constant k_B and of course the magnetic intensity B itself.

2. Metals

In a metal electrons are not bound to atoms and are free to roam around, like particles in a gas fluid. Magnetism in metals is thus very different to magnetism in insulators.

- *Pauli paramagnetism*

In an external magnetic intensity free electron spins can be either parallel or anti-parallel to the field. If we take B to be in the z-direction then parallel to the field (spin up, $S_z = \frac{1}{2}$) has lower energy than anti-parallel (spin down, $S_z = -\frac{1}{2}$) because of (35). Using $m = 2\mu_B S_z = \pm\mu_B$ for free electrons, the energies of spin up and spin down electrons in the field B are

$$\mathcal{U}_\pm = \mp\mu_B B.$$

A crucial aspect of the Fermi surface is that quantum states well below the Fermi energy ε_F are occupied by electrons are therefore not available to other electrons, due to the Pauli exclusion principle – they are essentially frozen out of the dynamics. At temperature T electrons with energies in the range $\varepsilon_E - k_B T < \varepsilon < \varepsilon_F$ can be thermally excited to states above ε_F and vacate a state with $\varepsilon_F - k_B T < \varepsilon < \varepsilon_F$, making it available to any higher energy electrons and that state can then affect the dynamics. Just as in the calculation of heat capacities in metals that we did earlier only states with energies in the range $\varepsilon_E - k_B T < \varepsilon < \varepsilon_F + k_B T$ contribute to the magnetic properties of a metal because all lower energy state are blocked by the exclusion principle and all higher energy states are beyond the reach of thermal excitations.

If we denote the number density of spin up electrons by n_+ and spin down electrons by n_-, so the total number density of electrons is $n = n_+ + n_-$, we can use (34) to evaluate $n_\pm$ by modifying it to allow for the energy shift. The energy of an otherwise free electron in the external field B is

$$\varepsilon_\pm = \varepsilon \mp \mu_B B = \frac{\hbar^2 k^2}{2m} \mp \mu_B B,$$

so the lowest energies in each case ($k = 0$) are $\mp\mu_B B$. For the number densities $n_\pm$ we can use equation (34) modified to

$$n_\pm = \frac{1}{2V}\int_{\mp\mu_B B}^{\infty} f_F(\varepsilon_\pm)\mathcal{D}(\varepsilon)d\varepsilon$$

where

$$f_F(\varepsilon_\pm) = \frac{1}{e^{\frac{\varepsilon_\pm - \mu}{k_B T}} + 1}.$$

The factor of 1/2 is because, in the absence of any magnetic intensity

$$n_+ = n_- = \frac{1}{2}n = \frac{1}{2V}\int_0^\infty f_F(\varepsilon)\mathcal{D}(\varepsilon)d\varepsilon.$$

Changing the integration variable $\varepsilon \to \varepsilon \pm \mu_B B$

$$n_\pm = \frac{1}{V}\int_0^\infty f_F(\varepsilon)\mathcal{D}(\varepsilon \pm \mu_B B)d\varepsilon$$

where

$$\mathcal{D}_\pm(\varepsilon) = \frac{1}{2}\mathcal{D}(\varepsilon \pm \mu_B B)$$

is the density of state for spin up/down electrons. As long as the magnetic intensity is not too large and $\mu_B B \ll \varepsilon_F$ we can Taylor expand

$$\mathcal{D}_\pm(\varepsilon) = \mathcal{D}_\pm(\varepsilon \pm \mu_B B) = \frac{1}{2}\mathcal{D}(\varepsilon) \pm \frac{1}{2}\mu_B B \mathcal{D}'(\varepsilon) + \cdots \ .$$

The magnetisation per unit volume is

$$M = (n_+ - n_-)\mu_B = \frac{\mu_0 \mu_B^2 B}{V}\int_0^\infty f_F(\varepsilon)\mathcal{D}'(\varepsilon)d\varepsilon = -\frac{\mu_0 \mu_B^2 B}{V}\int_0^\infty f_F'(\varepsilon)\mathcal{D}(\varepsilon)d\varepsilon,$$

where we have integrated by parts in the last equation, with $\mathcal{D}(0) = 0$ and used $\lim_{\varepsilon\to\infty} f_F(\varepsilon) = 0$. From this we get the volume susceptibility

$$\chi_m = -\frac{\mu_0 \mu_B^2}{V}\int_0^\infty f_F'(\varepsilon)\mathcal{D}(\varepsilon)d\varepsilon.$$

At zero temperature $f_F(\varepsilon)$ is a step function,

$$f_F(\varepsilon) = \begin{cases} 1, & \varepsilon < \varepsilon_F; \\ 0, & \varepsilon > \varepsilon_F, \end{cases}$$

and $f_F'(\varepsilon) = -\delta(\varepsilon - \varepsilon_F)$ is a Dirac δ-function, so, at temperatures, $k_B T \ll \varepsilon_F$, we can approximate

$$\chi_m = \frac{\mu_0 \mu_B^2}{V}\int_0^\infty \delta(\varepsilon - \varepsilon_F)\mathcal{D}(\varepsilon)d\varepsilon = \mathcal{D}(\varepsilon_F)\mu_B^2.$$

This is known as *Pauli paramagnetism*, the magnetic susceptibility is independent of temperature (the approximation $k_B T \ll \varepsilon_F$ is valid up to $T \approx 10^{4\circ}$K).

From (31) this can be re-written

$$\chi_m = \frac{3\mu_0}{2}\frac{n}{\varepsilon_F}\mu_B^2.$$

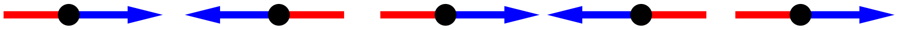
Comparing this this with Curie's Law (39), with $L = 0$, $J = S = 1/2$ and $p = \sqrt{3}$, we see that, apart from a numerical factor of 3/2, $k_B T$ in Curie's Law is replaced with the Fermi energy $\varepsilon_F = k_B T_F$ in a metal. In an insulator the temperature affects the magnetic susceptibility, in a metal it does not as long as the temperature does not approach the Fermi energy. This is due to blocking of the filled low energy states by the Pauli exclusion principle and results in much lower paramagnetic susceptibilities in metals than in insulators.[38]

[38] Just as for specific heats in metals, the low value of the magnetic susceptibility in metals was a mystery until it was realised that electrons obey Fermi-Dirac statistics and the exclusion principle.

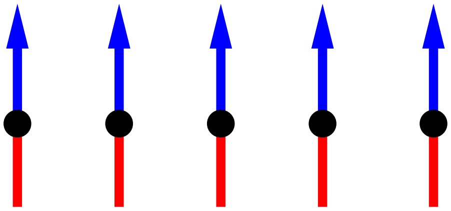
In our discussion so far we have ignored interactions, either electric or magnetic, between electrons. When such interactions become important there are further interesting magnetic phenomena.

3. Ferromagnets

We have been viewing electrons as little magnets with magnetic dipole moments. Classically a line of such magnets free to pivot around fixed centres will line up, as in the figure below, the net magnetisation will be zero.



But if the centres are not fixed the dipoles will be magnetically attracted to each other. Quantum mechanically however there are stranger forces at work: the exclusion principle again dictates that two electrons cannot occupy the same quantum state, so, if their spins are parallel they cannot be at the same point in space while if their spins are parallel they can, at least in principle though their Coulomb repulsion would mitigate against this dynamically. In fact the exclusion principle can work in tandem with Coulomb repulsion, the Coulomb energy is reduced if the electrons keep their distance and the exclusion principle says they cannot be in the same place if their spins are parallel so if their spins are parallel the exclusion principle can help keep them apart, thus reducing their Coulomb energy and indeed this can happen. At low temperatures the exclusion principle in conjunction with the Coulomb energy can ensure that the minimum energy state is that the electrons keep their and their spins are *aligned*, parallel to each other, as in the figure below.



When this happens the material develops a net magnetisation $\mathbf{M}$ *without* having to apply an external field, so $\mathbf{B}_{Applied} = 0$. This is what happens in fridge magnet or a compass needle, the magnetic field generated by a fridge magnet, and famously made visible with iron filings sprinkled onto a piece of paper over the magnet. If the volume of the magnet is V then the total magnetic dipole moment is $\mathbf{M}V$ and the magnetic field outside the magnetic and a distance r well away from it, where $\mu = \mu_0$, is a dipole field with[39]

$$\mathbf{H}(\mathbf{r}) = \frac{1}{\mu_0}\mathbf{B}(\mathbf{r}) = \left(\frac{3(\mathbf{M}.\mathbf{r})\mathbf{r}}{r^5} - \frac{\mathbf{M}}{r^3}\right)V.$$

Inside the material a non-zero magnetisation when $\mathbf{B}_{Applied} = 0$ is only maintained as long as the temperature is not too high. If the temperature increases the thermal energy

[39] The Earth's magnetic field is primarily a magnetic dipole but the Earth is not a ferromagnet. Instead the field is generated by electrical currents due to convection in the molten iron and nickel of the Earth's outer core.

makes the dipoles jiggle around and the order can be destroyed. Typically this happens at a temperature T_C, the Curie temperature mentioned earlier. Above the Curie temperature the medium has no net magnetisation in the absence of an external field, but can develop one when a field is applied and it becomes a paramagnet. This can be understood from equation (37) by treating the electrons as free, with $L = 0$, $J = S = 1/2$ and $g = 2$. Then (37) gives[40]

$$M = n\mu_B \tanh\left(\frac{\mu_B B}{k_B T}\right).$$

B here is the total magnetic intensity inside the material, which is non-zero in a ferromagnet even when no external field is applied, a non-zero B is generated by M and, assuming that B is proportional to M, $B = \lambda M$, we get a relation between M and T,

$$M = n\mu_B \tanh\left(\frac{\mu_B \lambda M}{k_B T}\right), \tag{40}$$

which gives $M(T)$ implicitly in terms of T. When $\mu_B\lambda M \ll k_B T$ we can approximate $\tanh x \approx x - \frac{1}{3}x^3 + \cdots$ and, at lowest order,

$$M \approx \frac{\mu_B^2 n\lambda M}{k_B T} \qquad \Rightarrow \qquad T \approx \frac{\mu_B^2 n\lambda}{k_B} \qquad \text{when} \quad \mu_B\lambda M \ll k_B T$$

and M vanishes at the Curie temperature

$$T_C = \frac{\mu_B^2 n\lambda}{k_B}.$$

We can deduce the functional form of $M(T)$ near T_C by including the x^3 term. For T below T_C, but close to it, let $T = T_C(1-t)$ with $0 \leq t \ll 1$. Then

$$\frac{\mu_B\lambda M}{k_B T} = \left(\frac{\mu_B n\lambda M}{k_B T_C}\right)(1-t) = \left(\frac{M}{\mu_B n}\right)\frac{1}{1-t} = \frac{M}{\mu_B n}(1+t+\cdots)$$

and (40) gives

$$\begin{aligned} M &= n\mu_B\left(\frac{M}{\mu_B n}(1+t) - \frac{1}{3}\left(\frac{M}{\mu_B n}\right)^3(1+3t) + O(t^3)\right) \\ &\approx M\left(1 + t - \frac{1}{3}\left(\frac{M}{\mu_B n}\right)^2(1+3t)\right) \\ \Rightarrow \qquad 0 &\approx M\left(t - \frac{1}{3}\left(\frac{M}{\mu_B n}\right)^2(1+3t)\right). \end{aligned}$$

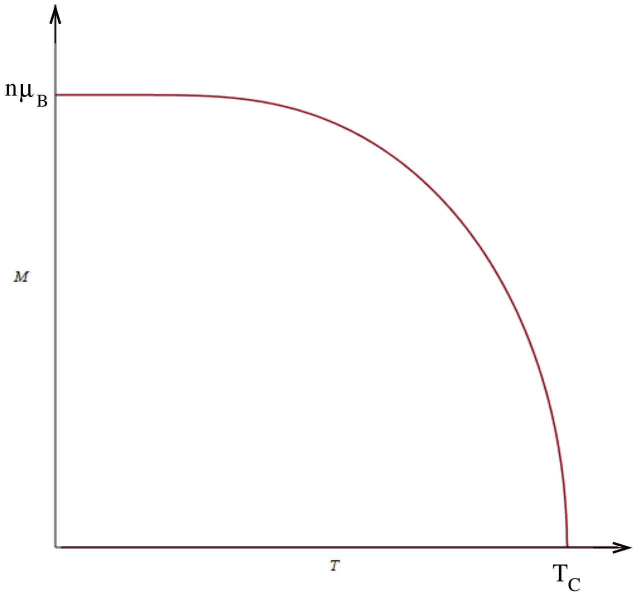
[40] From (38), with J=1/2 and using $\cosh(2x)=(\cosh x)^2+(\sinh x)^2$ together with $\sinh(2x)=2\sinh x\cosh x$, we have $B_2=2\coth(2x)-\coth x=\frac{2\cosh(2x)}{\sinh(2x)}-\frac{\cosh x}{\sinh x}=\frac{2(\cosh x)^2+2(\sinh x)^2}{2\sinh x\cosh x}-\frac{\cosh x}{\sinh x}=\tanh x$.

M vanishes when $t = 0$ so, when $1 \gg t > 0$, we can assume that $\frac{M}{\mu_B n}$ is also small and we can ignore the $M^2 t$ term, giving

$$t \approx \frac{1}{3}\left(\frac{M}{\mu_B n}\right)^2$$

and $M \propto \sqrt{t}$.

The function $M(T)$ defined implicitly in (40) does not have a simple analytic expression, but it can be plotted numerically and shown below is a plot of the magnetisation of a ferromagnet as a function of temperature, with the $\sqrt{t}$ behaviour near T_C. Also $\lim_{x\to\infty} \tanh x = 1$, so $M \to \mu_B n$ as $T \to 0$.



Experimentally it is found that $M \propto t^{\nu}$ with $\nu \approx 0.33$, rather than $\sqrt{t}$ and this is universal behaviour for all ferromagnets (in 3-dimensions, for 2-dimensional layers $M \propto t^{1/8}$). The reason for this discrepancy is that thermal fluctuations, which are usually very small for macroscopic systems, become significant near T_C and the local magnetisation experiences large fluctuations — we need a more sophisticated theory, at least near T_C. The exponent ν, called a *critical exponent*, can be determined numerically within the framework of the theory of critical phenomena and is quite well understood but the derivation is beyond the scope of this exposition.

This change from the ferromagnetic state to the paramagnetic state as the temperature is increased is an example of a phase transition, very similar in many of its details to what happens when water boils. The Curie temperature in the ferromagnetic-paramagnetic phase transition is very like the critical temperature at the critical point of water. When water boils at atmospheric pressure at 100°C the volume increase by a factor of 10^4 as liquid water turns to water vapour, but as the pressure increases volume of the vapour decreases while the liquid is incompressible, at the same time the the boiling point goes up. At a pressure of 218 atmospheres the boiling point is 374°c and the volume of the vapour is identical to that of the liquid phase. At higher temperatures there is no real distinction between liquid and vapour phase, we just have a very hot fluid. Water does not boil at temperatures above 374°c at any pressure — this is called the *critical temperature*. The magnetisation of a ferromagnet is like the difference in density between the liquid and the gaseous phase of water and the Curie temperature for a ferromagnetic is analogous to the critical temperature of water.

In contrast to ferromagnets in some materials the electron magnetic moments prefer to line up at low temperatures as in the first figure in this section §8.3, with the dipoles in a horizontal line as would be expected classically. While such materials do not exhibit any permanent magnetisation they are nevertheless in a very ordered state at low temperatures and are called *anti-ferromagnets*. Chromium and manganese oxide are examples of anti-ferromagnetic materials.

4. Superconductors

For most conductors the conductivity increases as the temperature is decreased, because there are fewer phonons for the electrons to scatter off, but the conductivity remains finite as $T \to 0$ because there are always imperfections in the crystal structure, impurities or atoms missing from some lattice sites. In principle in an absolutely pure, perfectly ordered, crystal the conductivity would become infinite at zero temperature. But this is not what a superconductor is. While superconductors do display zero resistivity at zero temperature they also have the characteristic of expelling any magnetic intensity: the defining features of a superconductor are zero resistivity *and* $\mathbf{B} = 0$ inside the material. From (30) this is equivalent to $\mathbf{H} = -\mathbf{M}$: it is not the magnetic field $\mathbf{H}$ that is expelled from a superconductor it is the magnetic intensity $\mathbf{B}$.

In a perfect conductor there is no scattering of electrons and an electric field will accelerate electrons indefinitely, Newton's 2nd law gives

$$m_e \frac{d\mathbf{v}}{dt} = -e\mathbf{E}$$

where $\mathbf{v}$ is the velocity of the electron. If all electrons have the same velocity and there are n electrons per unit volume the current density is

$$\begin{aligned} & \mathbf{j} = -en\mathbf{v} \\ \Rightarrow \qquad & \frac{d\mathbf{j}}{dt} = -en\frac{d\mathbf{v}}{dt} = \frac{e^2 n}{m_e}\mathbf{E}. \end{aligned} \tag{41}$$

where we assume that n is independent of time. For an AC current $\mathbf{j} = e^{-i\omega t}\tilde{\mathbf{j}}(\omega)$, generated by an oscillating electric field $\mathbf{E} = e^{-i\omega t}\widetilde{\mathbf{E}}(\omega)$,

$$\tilde{\mathbf{j}}(\omega) = \frac{ie^2 n}{m_e \omega}\widetilde{\mathbf{E}}(\omega) = \sigma(\omega)\widetilde{\mathbf{E}}(\omega)$$

and we have a frequency dependent conductivity

$$\sigma(\omega) = \frac{ie^2 n}{m_e \omega}$$

which is infinite at zero frequency, the assumption of no scattering has led to an infinite conductivity.[41]

Taking the curl of (41) and using Faraday' law of electromagnetic induction, that a time varying magnetic intensity generates and electric field according to

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E},$$

we deduce that

$$\frac{\partial}{\partial t}\left(\nabla \times \mathbf{j} + \frac{e^2 n}{m_e}\mathbf{B}\right) = 0,$$

hence $\nabla \times \mathbf{j} + \frac{e^2 n}{m_e}\mathbf{B}$ is independent of time in a perfect conductor. In a superconductor we postulate that

$$\nabla \times \mathbf{j} + \frac{e^2 n}{m_e}\mathbf{B} = 0 \tag{42}$$

and deduce that $\mathbf{B} = \mathbf{j} = 0$ inside the superconductor (a justification of the London equation will be given below). To see this we use the two other Maxwell's equations, when $\mathbf{E}$ is independent of time,[42]

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j}, \qquad \nabla . \mathbf{B} = 0 \qquad \Rightarrow \qquad -\nabla^2 \mathbf{B} = \mu_0 (\nabla \times \mathbf{j}),$$

to give, with (42),

$$\nabla^2 \mathbf{B} = \frac{\mu_0 e^2 n}{m_e}\mathbf{B}. \tag{43}$$

We can also take the curl of (42), and use current conservation $\nabla . \mathbf{j} = 0$ for a time independent current, to arrive at

$$\nabla^2 \mathbf{j} = \frac{\mu_0 e^2 n}{m_e}\mathbf{j}. \tag{44}$$

Equation (43) and (44) have exponentially growing and exponentially damped solutions. The equation $y''(x) = \lambda^2 y(x)$ has solutions $y \propto e^{\pm \lambda x}$. With λ and x positive, $e^{\lambda x}$ grows indefinitely with x while $e^{-\lambda x}$ is exponentially damped and falls off rapidly as x increases. Rejecting the exponentially growing solutions, this means that both $\mathbf{B}$ and $\mathbf{j}$ fall rapidly to zero as we go inside a superconductor with characteristic length

$$\lambda_L = \sqrt{\frac{m_e}{\mu_0 e^2 n}}.$$

Thus $\mathbf{B}$ can only penetrate a distance λ_L into the superconductor and any current carried by the superconductor is constrained to a thin layer of thickness λ_L at the surface. This

[41] The conductivity is complex and the imaginary part is associated with optical absorption, rather than Ohmic conductivity, but there is a relation between the real and imaginary parts dictated by complex analyticity, called the *Kramers-Kronig relation*, which says that a pole in the imaginary part at ω=0 is related to infinite DC Ohmic conductivity.

[42] We use μ_0 here because $\mathbf{j}$ is the naïve current, without yet taking into account the medium's reaction.

is a consequence of (42) which is called the *London equation* and λ_L is called the *London penetration depth.* Typically λ_L is a few hundred Å, in lead for example $\lambda_L = 3.4 \times 10^{-8}$m.

A very successful microscopic mathematical model of superconductivity was constructed by Bardeen, Cooper and Schrieffer in 1957, called the BCS theory, which won them the Nobel prize in 1972, but the full description of BCS theory is beyond the scope of these notes. The basic idea is that the positive ion cores in a metal attract electrons and sometimes the conditions are such that the electrons actually experience an overall attraction which can result in a bound state consisting of two electrons, called a *Cooper pair.* The electrons like to pair up in an s-wave with spins in opposite directions (if the spins are in the same direction the exclusion principle would make it harder for the electrons to pair) so they form a bound state with electric charge $-2e$ and spin zero, they form charged bosons. At low temperatures all the Bosons fall into the same quantum state, there is macroscopic quantum coherence and a Bose-Einstein condensate is formed which is the superconducting state. The attraction between the electrons is very weak, they are easily disrupted as the temperature is raised, the Cooper pairs evaporate and the superconductor reverts to being a normal metal, superconductivity requires low temperatures. The Cooper pairs are also quite large, the wavefunction of the pair typically spreads over $\xi \sim 10^3$Å, covering a volume containing $\sim 10^9$ ions and overlapping with $\sim 10^9$ other pairs (ξ is called the *correlation length*, it is a measure of the distance over which the electron's wave-functions are correlated). The reason for the superconductivity is that it costs an energy Δ to break up the pairs and there are no quantum states between the lowest energy state and the first excited state, as long as $\kappa_B T \ll \Delta$ there can be no interactions and no dissipation, the resistivity is zero. Δ is called an *energy gap* and typically Δ is of order $10^{-3} \sim 10^{-4}\varepsilon_F$, about 10^{-3}eV (10^{-22}J) and the superconducting state requires temperatures some three or four magnitudes less than the Fermi temperature, *i.e.*10°K or less.[43] As always in a metal electrons deep below the Fermi surface are frozen out of the dynamics by the exclusion principle, it is only a small fraction of electrons near the Fermi surface that can form Cooper pairs.

Just as for ferromagnets there is critical temperature T_c above which superconductivity is destroyed, though it is usually orders of magnitude below the Curie temperature for a ferromagnetic.

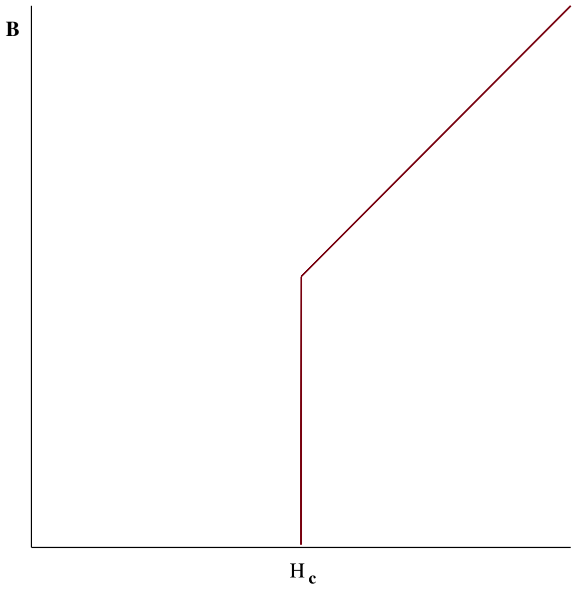
Very strong magnetic fields also destroy superconductivity. If a metal that can superconduct at low temperatures is placed in an external magnetic field at room temperature and the temperature is lowered until it becomes a superconductor the magnetic intensity will be expelled from the bulk of the material only if the external field is not too large. At any given temperature below T_c there is a critical magnetic field $\mathbf{H}_c(T)$ that will destroy the superconducting state and render the metal normal. $\mathbf{H}_c(T_c) = 0$ at $T = T_c$ and increases as T decreases, tending to a finite value $\mathbf{H}_c(0)$ as $T \to 0$. Typically $\mathbf{H}_c(0)$ is of order of a Tesla, about 10^4 times the strength of the Earth's magnetic field.

It takes some energy to expel the $\mathbf{B}$-field from a superconductor as it is cooled, there is latent heat associated with the phase transition between the normal and superconducting

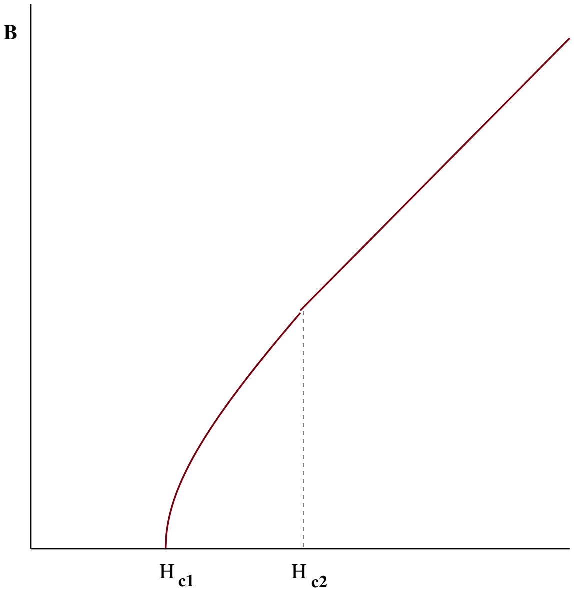
[43] An exception to this are the high-T_c superconductors discovered in 1986, which can superconduct above 77°K, the temperature at which nitrogen liquidises. These have a 2-dimensional layered structure and the Bosons are believed to be spin 1. The microscopic theory of these superconductors is not yet fully understood.

states of matter. Conversely if a conductor transitions from a superconducting state to its normal state energy is released, a phenomenon known as *quenching*. For a strong superconductor this can be quite dramatic. The Large Hadron Collider, the most powerful particle accelerator in the world, at CERN in Geneva uses very strong superconducting magnets to contain and focus the beams of protons circulating round it and it experienced a quenching event in 2008, caused by a faulty electrical connection, which resulted in a number of magnets having to be replaced.

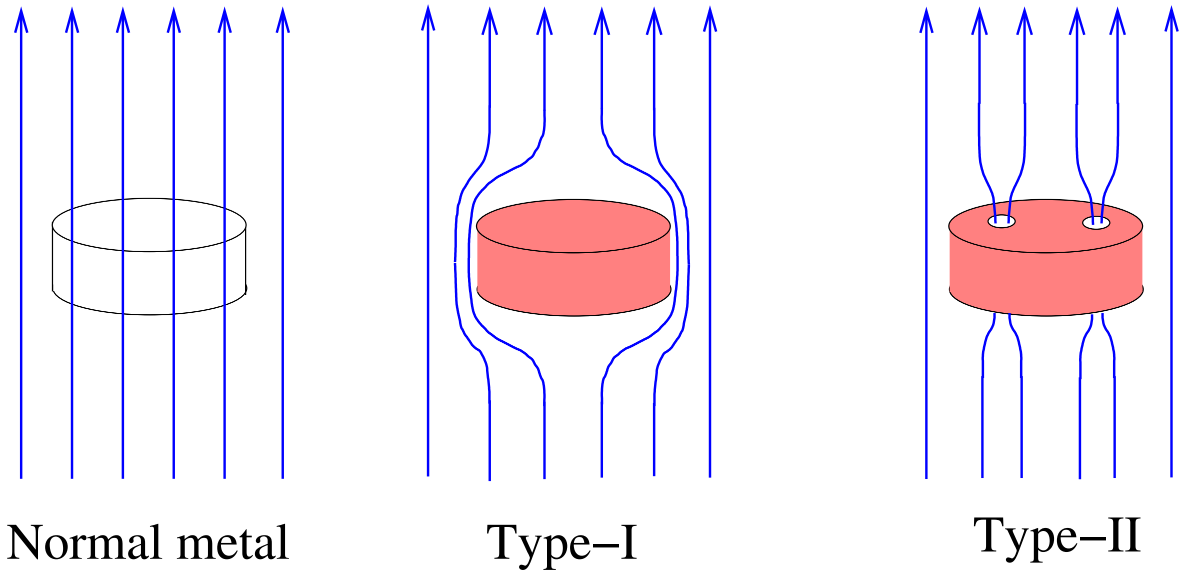
The way in which the Meissner effect is implemented depends very much on whether the London penetration depth λ_L is greater than or less than the coherence length ξ. If $\xi > \lambda_L$ the superconductor is called type-I and if $\xi < \lambda_L$ it is called type-II. There is energy associated with the interface between the normal and the superconducting phases and this is positive when $\xi > \lambda_L$ and negative when $\xi < \lambda_L$. Type-I superconductors like to minimise the area between the two phases and type-II like to maximise it. In a type-I superconductor $\mathbf{B}$ is just excluded discontinuously, at any $T < T_c$ there is a finite jump from $\mathbf{B} = 0$ to $\mathbf{B} \neq 0$ as $\mathbf{H}$ is increased from below $\mathbf{H}_c(T)$ to above it. In a type-II superconductor the transition is smoother and there is a lower critical field $\mathbf{H}_{c1}$ below there is no penetration of magnetic flux through the superconductor and an upper critical field $\mathbf{H}_{c2}$ above which the sample is a normal metal with no Cooper pairs. In between, for $\mathbf{H}_{c1} < \mathbf{H} < \mathbf{H}_{c2}$ there is some penetration of $\mathbf{B}$ into the sample, but it is confined to narrow tubes. The situation is illustrated in the figures below



Type–I



Type–II



Whether or not a superconductor is type-I or type-II is determined by the ratio of the London penetration depth λ_L and the coherence length ξ.

Type-II superconductors are fascinating materials. In a type-II superconductor the magnetic intensity threading through the superconductor is restricted to thin lines passing through the material in the direction of $\mathbf{H}$ called *magnetic vortices*. These thin vortices are non-superconducting, they are thin threads inside about the thickness of a Cooper pair ξ inside of which the material is in the normal state. Remarkably the magnetic flux through these vortices is quantised.

This is understood by the following argument. In quantum mechanics momentum, and hence velocity, is an operator. For a particle with charge q and mass m moving in a magnetic field with magnetic intensity $\mathbf{B} = \nabla \times \mathbf{A}$,

$$\mathbf{v} = \frac{1}{m}(\mathbf{p} - q\mathbf{A}) = \frac{1}{m}(-i\hbar\nabla - q\mathbf{A}).$$

If

$$\Psi = \sqrt{n}e^{i\theta}$$

is the multi-particle wave-function, normalised so that the number of particles per unit volume n is

$$n = \Psi^*\Psi,$$

which we shall assume is uniform and independent of position, then the current density is

$$\mathbf{j} = q(\Psi^*\mathbf{v}\Psi) = \frac{nq}{m}(\hbar\nabla\theta - q\mathbf{A}).$$

This immediately gives the London equation (42)

$$\nabla \times \mathbf{j} = -\frac{nq^2}{m}\nabla \times \mathbf{A} = -\frac{nq^2}{m}\mathbf{B}.$$

Now consider a region of superconducting material with a hole in the middle, $\mathbf{B} = 0$ in the superconductor but can be non-zero in the hole. Integrating round a closed curve C in the superconductor that surrounds the hole (see figure below) gives

$$\int_C \mathbf{j}.d\mathbf{l} = \frac{nq}{m}\int_C (\hbar\nabla\theta - q\mathbf{A}).d\mathbf{l} = \frac{nq\hbar}{m}(\theta_2 - \theta_1) - \frac{nq^2}{m}\int_S (\nabla\times\mathbf{A}).d\mathbf{S} = -\frac{nq^2}{m}\int_S \mathbf{B}.d\mathbf{S},$$

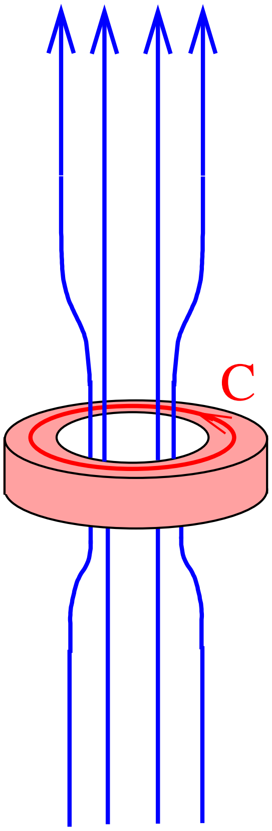
where we have used Stokes' theorem with S a surface stretched across the hole whose boundary is C and θ_1 and θ_2 are the initial and final values of the phase θ as we integrate around C. Now it is not necessarily the case that $\theta_2 = \theta_1$, but for single-valuedness of Ψ it must be that they can only differ by 2π times an integer r,

$$\theta_1 - \theta_2 = 2\pi r.$$

We conclude that the total magnetic flux through the hole

$$\Phi = \int_S \mathbf{B}.d\mathbf{S} = \frac{2\pi\hbar r}{q} = \frac{hr}{q}$$

is quantised in units of $\frac{h}{q}$.



In a superconductor the current is carried by Cooper pairs consisting of pairs of electrons and $q = -2e$, so we see that the vortices penetrating a type-II superconductor are quantised in units of

$$\Phi_0 = \frac{h}{2e} = 2.067833848\ldots \times 10^{-15} \text{ Tesla m}^2.$$

This is a truly remarkable result, the flux quantisation is given purely in terms of the fundamental constants h and e and is completely independent of the properties of the material, it is the same for any superconductor!

Appendix A: Sommerfeld expansion

When calculating the heat capacity of a metal as a function of temperature we expanded in the small parameter $\frac{k_BT}{\varepsilon_F}$, this is called the Sommerfeld expansion. To see how the Sommerfeld expansion works first start with N rather than U, (34),

$$N = \int_0^\infty f_F(\varepsilon)\mathcal{D}(\varepsilon)d\varepsilon = \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}}\int_0^\infty \varepsilon^{\frac{1}{2}} f_F(\varepsilon)d\varepsilon,$$

from (31). The most important region of the integral is around $\varepsilon = \mu$, where the integrand falls off steeply (see last figure), and we can focus on this region by integrating by parts:

$$N = \frac{2}{3}\frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}}\left[\varepsilon^{\frac{3}{2}} f_F(\varepsilon)\right]_0^\infty - \frac{2}{3}\frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}}\int_0^\infty \varepsilon^{\frac{3}{2}}\frac{df_F(\varepsilon)}{d\varepsilon}d\varepsilon.$$

The first term vanishes at both limits while, at least for $\frac{T}{T_F} << 1$, the second is dominated by the region around $\varepsilon \approx \mu$, where $-\frac{df_F(\varepsilon)}{d\varepsilon}$ is large. Write

$$-\frac{df_F(\varepsilon)}{d\varepsilon} = -\frac{d}{d\varepsilon}\left(e^{(\varepsilon-\mu)/k_BT}+1\right)^{-1} = \frac{1}{k_BT}\frac{e^x}{(e^x+1)^2},$$

where $x := \frac{\varepsilon-\mu}{k_BT}$. Now

$$N = \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}}\int_0^\infty \varepsilon^{\frac{3}{2}}\frac{e^x}{(e^x+1)^2}\frac{d\varepsilon}{k_BT} = \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}}\int_{-\frac{\mu}{k_BT}}^\infty \varepsilon^{\frac{3}{2}}\frac{e^x}{(e^x+1)^2}dx,$$

where the integration variable has been changed from ε to x in the last expression. Now we do two things: first observe that for $x < -\frac{\mu}{k_BT}$, that is $\varepsilon < 0$, the integrand is exponentially suppressed[44] so we can extend the lower limit of integration down to $-\infty$ and this has a negligible effect on the integral; second we Taylor expand $\varepsilon^{\frac{3}{2}}$ about $\varepsilon = \mu$,

$$\begin{aligned}\varepsilon^{\frac{3}{2}} &= \mu^{\frac{3}{2}} + (\varepsilon-\mu)\left.\frac{d}{d\varepsilon}\varepsilon^{\frac{3}{2}}\right|_{\varepsilon=\mu} + \frac{1}{2}(\varepsilon-\mu)^2\left.\frac{d^2}{d\varepsilon^2}\varepsilon^{\frac{3}{2}}\right|_{\varepsilon=\mu} + \dots\\ &= \mu^{\frac{3}{2}} + \frac{3}{2}(\varepsilon-\mu)\mu^{\frac{1}{2}} + \frac{3}{8}(\varepsilon-\mu)^2\mu^{-\frac{1}{2}} + \dots .\end{aligned}$$

This allow us to re-write

$$\begin{aligned}N &= \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}}\int_{-\infty}^\infty \varepsilon^{\frac{3}{2}}\frac{e^x}{(e^x+1)^2}dx\\ &= \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}}\int_{-\infty}^\infty \frac{e^x}{(e^x+1)^2}\left(\mu^{\frac{3}{2}} + \frac{3}{2}k_BTx\mu^{\frac{1}{2}} + \frac{3}{8}(k_BTx)^2\mu^{-\frac{1}{2}} + \dots\right)dx,\end{aligned}$$

[44] Putting in some numbers, with $\mu{\approx}\varepsilon_F$, $x < -\frac{\mu}{k_BT} \quad \Rightarrow \quad x < -\frac{\varepsilon_F}{k_BT}{\approx}{-100}$ and e^{-x} is tiny.

since $\varepsilon - \mu = k_B T x$.

Each individual integral over x on the right hand side can now be evaluated:

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{e^x}{(e^x+1)^2} dx &= \int_{-\infty}^{\infty} -\frac{d}{dx}\left(\frac{1}{e^x+1}\right) dx = -\left[\frac{1}{(e^x+1)}\right]_{-\infty}^{\infty} = 1, \\
\int_{-\infty}^{\infty} \frac{xe^x}{(e^x+1)^2} dx &= \int_{-\infty}^{\infty} \frac{x}{(e^x+1)(1+e^{-x})} dx = 0, \\
\int_{-\infty}^{\infty} \frac{x^2 e^x}{(e^x+1)^2} dx &= \int_{-\infty}^{\infty} \frac{x^2}{(e^x+1)(e^{-x}+1)} dx = 2\int_0^{\infty} \frac{x^2}{(e^x+1)(e^{-x}+1)} dx \\
&= -2\int_0^{\infty} x^2 \frac{d}{dx}\left(\frac{1}{e^x+1}\right) dx = 4\int_0^{\infty} \frac{x}{e^x+1} dx = \frac{\pi^2}{3}.
\end{aligned}$$

The second integral vanishes because the integrand is an odd function of x and the third is left as an exercise.[45] We now have

$$N = \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}} \mu^{\frac{3}{2}} + \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}} \frac{\pi^2}{8}(k_B T)^2 \mu^{-\frac{1}{2}} + \dots .$$

In terms of the Fermi energy (57) this is

$$\begin{aligned}
N &= N\left(\frac{\mu}{\varepsilon_F}\right)^{\frac{3}{2}} + N\frac{\pi^2}{8}\frac{(k_B T)^2}{\varepsilon_F^{\frac{3}{2}}\mu^{\frac{1}{2}}} + \dots \\
\Rightarrow \qquad 1 &= \left(\frac{\mu}{\varepsilon_F}\right)^{\frac{3}{2}} + \frac{\pi^2}{8}\left(\frac{k_B T}{\varepsilon_F}\right)^2 \left(\frac{\varepsilon_F}{\mu}\right)^{\frac{1}{2}} + \dots ,
\end{aligned}$$

where the dots indicate terms of order $\left(\frac{k_B T}{\varepsilon_F}\right)^4$ and higher. This implies that $\frac{\mu}{\varepsilon_F} = 1 + o\left(\frac{k_B T}{\varepsilon_F}\right)^2$, so we can replace the $\left(\frac{\varepsilon_F}{\mu}\right)^{\frac{1}{2}}$ factor in the second term on the right hand side above with unity, absorbing the difference into the dots, giving

$$\frac{\mu}{\varepsilon_F} = \left[1 - \frac{\pi^2}{8}\left(\frac{k_B T}{\varepsilon_F}\right)^2\right]^{\frac{2}{3}} + \dots = 1 - \frac{2}{3}\frac{\pi^2}{8}\left(\frac{k_B T}{\varepsilon_F}\right)^2 + \dots .$$

[45] Hint: write $\frac{1}{e^x+1} = \frac{e^{-x}}{1+e^{-x}} = e^x \sum_{n=0}^{\infty} (-1)^n e^{-nx}$ and use Gamma functions, $\Gamma(n) = \int_0^{\infty} x^{(n-1)} e^{-x} dx = (n-1)!$.

Appendix B: Electromagnetic units

The question of what system of units to use in describing the electric and magnetic properties of matter is notoriously moot. SI units are not good for describing electric and magnetic properties of materials, not only do they mess up some otherwise rather elegant formulae with ugly things with horrible sounding names (like the electric permittivity of the vacuum ϵ_0 and the magnetic permeability of the vacuum μ_0) but they also give **E** and **B** different units.

In SI units ϵ_0 and μ_0 are the electric permittivity of the vacuum and the magnetic permeability of the vacuum respectively, with the product $\epsilon_0\mu_0 = \frac{1}{c^2}$ with $c = 2.99792458 \times 10^8$m/s being the speed of light in a vacuum. Electric charge is measured in Coloumb (C), with the magnitude of the electric charge $e = 1.6 \times 10^{-19}$C. **B** and **E** have different dimensions!

Unfortunately SI units are nowadays the standard for pedagogical expositions so they are used in the text. cgs units are better, at least they give **E** and **B** the same dimensions, with ϵ_0 replaced by $\frac{1}{4\pi}$ and μ_0 replaced by $\frac{4\pi}{c^2}$, but these still suffer from 4π's all over the place. The most elegant units are Lorentz-Heavyside units in which $\epsilon_0 = 1$ and $\mu_0 = \frac{1}{c^2}$ and **E** and **B** have the same dimensions. If I ruled the world all electromagnetic formulae would be written in Lorentz-Heavyside units. Below the equations of electromagnetism are exhibited in all three systems of units for comparison.

SI units

Maxwell's Equations:

$$\begin{aligned} \nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} &= 0, \qquad \nabla . \mathbf{B} = 0 \\ \nabla \times \mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} &= \mathbf{j}, \qquad \nabla . \mathbf{D} = \rho \end{aligned}$$

Lorentz force law: Coulomb's law:

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \qquad \mathbf{F} = \frac{q_1 q_2}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}}$$

Constituent relations:

$$\begin{aligned} \mathbf{D} &= \epsilon_0 \mathbf{E} + \mathbf{P}, \qquad \mathbf{H} = \frac{1}{\mu_0}\mathbf{B} - \mathbf{M} \\ \mathbf{P} &= \epsilon_0 \chi_e \mathbf{E}. \qquad \mathbf{M} = \frac{\chi_m}{\mu_0}\mathbf{H} \\ \epsilon &= \epsilon_0(1 + \chi_e), \qquad \mu = \mu_0(1 + \chi_m) \end{aligned}$$

Only four dimensional quantities are need and in SI units these are metres (m) for length, kilogrammes (kg) for mass, seconds (s) for time and Coulombs (C) for electric charge, the magnitude of the charge on the electron being 1.60219×10^{-19}C. The units of $\mathbf{E}$ are kg m $\mathrm{C}^{-1}\,\mathrm{s}^{-2}$ those of $\mathbf{B}$ are kg $\mathrm{C}^{-1}\,\mathrm{s}^{-1}$ (also called a Tesla). The units of ϵ_0 are $\mathrm{C}^2\,\mathrm{s}^2\,\mathrm{kg}^{-1}\,\mathrm{m}^{-3}$ and those of μ_0 are kg m C^{-2},

$$\begin{aligned} \epsilon_0 &= \frac{10^7}{4\pi}\frac{1}{(2.997925 \times 10^8)^2}\mathrm{C}^2\mathrm{s}^2\mathrm{kg}^{-1}m^{-3} = 8.854185 \times 10^{-12}\,\mathrm{C}^2\mathrm{s}^2\mathrm{kg}^{-1}m^{-3} \\ \mu_0 &= 4\pi \times \ 10^{-7}\mathrm{kg}\,\mathrm{m}\,\mathrm{C}^{-2}. \end{aligned}$$

cgs units

Maxwell's Equations:

$$\nabla\times\mathbf{E}+\frac{1}{c}\frac{\partial\mathbf{B}}{\partial t}=0,\qquad \nabla.\mathbf{B}=0$$

$$\nabla\times\mathbf{H}-\frac{1}{c}\frac{\partial\mathbf{D}}{\partial t}=\frac{4\pi}{c}\mathbf{j},\qquad \nabla.\mathbf{D}=4\pi\rho$$

Lorentz force law: Coulomb's law:

$$\mathbf{F}=q\left(\mathbf{E}+\frac{\mathbf{v}}{c}\times\mathbf{B}\right)),\qquad \mathbf{F}=\frac{q_1q_2}{r^2}\hat{\mathbf{r}}$$

Constituent relations:

$$\mathbf{D}=\mathbf{E}+4\pi\mathbf{P},\qquad \mathbf{H}=\mathbf{B}-4\pi\mathbf{M}$$

$$\mathbf{P}=\chi_e\mathbf{E}.\qquad \mathbf{M}=\chi_m\mathbf{H}$$

$$\epsilon=1+4\pi\chi_e,\qquad \mu=1+4\pi\chi_m$$

Units are centimetres (cm) for length, grammes (g) for mass, seconds (s) for time and the electrostatic unit (esu) for electric charge, $1C = 2.997924580\times10^9$Fr. An esu is also known as a statcoulomb (statC) or also as a Franklin (Fr). **E** and **B** are both measured in g cm esu^{-1} s^{-2} while ϵ and μ are dimensionless. By setting $\epsilon_0 = 1$ the esu is actually redundant as it has dimensions of g$^{1/2}$ cm$^{3/2}$ s^{-1}.

Lorentz-Heavyside units

Maxwell's Equations:

$$\nabla \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = 0, \qquad \nabla.\mathbf{B} = 0$$

$$\nabla \times \mathbf{H} - \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t} = \frac{1}{c}\mathbf{j}, \qquad \nabla.\mathbf{D} = \rho$$

Lorentz force law: Coulomb's law:

$$\mathbf{F} = q\left(\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B}\right)), \qquad \mathbf{F} = \frac{q_1 q_2}{4\pi r^2}\hat{\mathbf{r}}$$

Constituent relations:

$$\mathbf{D} = \mathbf{E} + \mathbf{P}, \qquad \mathbf{H} = \mathbf{B} - \mathbf{M}$$

$$\mathbf{P} = \chi_e \mathbf{E}. \qquad \mathbf{M} = \chi_m \mathbf{H}$$

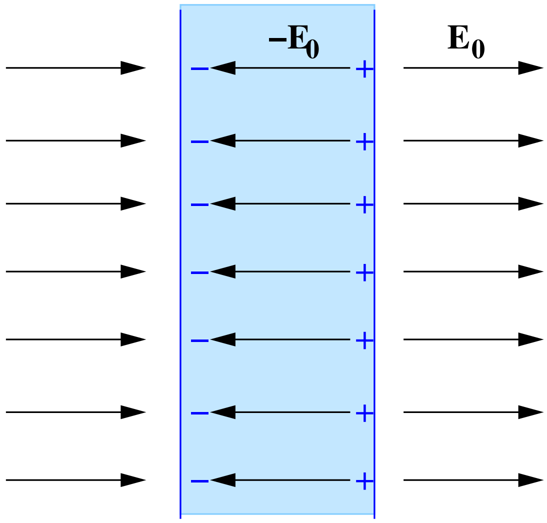
$$\epsilon = 1 + \chi_e, \qquad \mu = 1 + \chi_m$$

We can use metres for length, kilogrammes for mass, and seconds for time. Again setting $\epsilon_0 = 1$ means that charge has units of $\mathrm{kg}^{1/2}\,\mathrm{m}^{3/2}\,\mathrm{s}^{-1}$. $\mathbf{E}$ and $\mathbf{B}$ are both measured in $\mathrm{kg}^{1/2}\,\mathrm{m}^{-1/2}\,\mathrm{s}^{-1}$. Alternatively we could use cm and g for length and mass.

Although not appropriate for condensed matter we note that Maxwell's equations are much tidier if we use units with $c = 1$ and measure lengths in light-seconds (ls), $1\ \mathrm{ls} = 2.997924580 \times 10^8\,\mathrm{m}$. Lorentz-Heavyside units are also ideally suited to relativistic physics and are usually the units of choice in that domain.

Appendix B: Ferroelectrics and ferromagnets

A ferroelectric sustains a non-zero electric dipole moment even when there is no external electric field applied. To understand this consider a uniform slab of a ferroelectric material, perpendicular to the x-direction and extending infinitely far in the transverse y and z-directions, as shown below.



There is no external field here, $\mathbf{E} = \mathbf{E}_0$ is generated purely by the medium itself, by a permanent positive surface charge density σ_0 on one side and a negative surface charge density $-\sigma_0$ on the other. A simple application of Gauss' law tells us that the magnitude of $\mathbf{E}_0$ is

$$E_0 = \frac{\sigma_0}{2\epsilon_0}.$$

It has the same magnitude inside the material as outside but it points in the opposite direction inside. $\mathbf{E}_0$ is generated by a permanent intrinsic polarisation $\mathbf{P} = \mathbf{P}_0$ inside the slab, which is uniform for an infinite slab. Outside the slab $\mathbf{P} = 0$ so

$$\begin{aligned} \mathbf{D} &= \epsilon_0\mathbf{E} + \mathbf{P} = -\epsilon_0\mathbf{E}_0 + \mathbf{P} \qquad && \text{inside the slab} \\ \mathbf{D} &= \epsilon_0\mathbf{E} = \epsilon_0\mathbf{E}_0, && \text{outside the slab.} \end{aligned}$$

There are no free or external charges, so $\mathbf{D}$ is continuous across the surface, it is the same both inside and outside the slab,

$$D = \epsilon_0 E_0 = -\epsilon_0 E_0 + P_0 \qquad \Rightarrow \qquad P_0 = 2\epsilon_0 E_0.$$

It is tempting to conclude that $\chi_e = 2$, but this is misleading. Let us now apply a constant external field in the x-direction, $\mathbf{E}_{Applied} = E_{Applied}\hat{\mathbf{x}}$, so that $\mathbf{E} = \mathbf{E}_{Applied} + \mathbf{E}_{Medium}$. $\mathbf{E}_{Applied}$ will cause additional polarisation in the medium and tends to increase P. This increase can be modelled by an increase in the surface charge density $\sigma_0 \to \sigma = \sigma_0 + \delta\sigma$ giving an extra contribution $\delta E = \frac{\delta\sigma}{2\epsilon_0}$ to the total field exterior to the slab $E_{(e)}$. As long as $E_{Applied}$ is not too large we expect that $\delta\sigma \propto E_{Applied}$, so we can write $\delta\sigma = 2\lambda\epsilon_0 E_A$ with $0 < \lambda < 1$ ($\lambda < 1$ because $\delta\sigma$ cannot be responsible for *all* of E_A, which is being applied externally). This in turn will increase the polarisation inside the medium from $P_0 = \sigma_0$ to

$$P = \sigma_0 + \delta\sigma = P_0 + 2\epsilon_0\lambda E_A = 2\epsilon_0(E_0 + \lambda E_A).$$

The total electric field inside the slab is now

$$E = -E_0 - \delta E + E_A = -E_0 + (1-\lambda)E_A \qquad \Rightarrow \qquad E_A = \frac{E + E_0}{1-\lambda}$$

hence

$$\begin{aligned} P(E) &= 2\epsilon_0\left(E_0 + \frac{\lambda}{1-\lambda}(E + E_0)\right) \\ &= \frac{2\epsilon_0 E_0}{(1-\lambda)} + \frac{2\lambda\epsilon_0}{(1-\lambda)}E \\ &= P(0) + \epsilon_0\chi_e E \end{aligned}$$

where

$$P(0) = \frac{P_0}{1-\lambda} \qquad \text{and} \qquad \chi_e = \frac{2\lambda}{1-\lambda}.$$

$P(0)$ is defined to be the polarisation when the *total* field $E = 0$, it is *not* the same as the intrinsic pyroelectric polarisation P_0 which is defined to be the polarisation when the *applied* field $E_A = 0$.

A pyroelectric is described by a polarisation that is linear in the total field E but with a non-zero constant term. A parallel discussion can be given for a ferromagnet, but magnetic susceptibilities are so small that $\lambda \ll 1$ and $M(0)$ is almost the same as M_0.