FoP3B Part II Lecture 1: Band gaps in Semiconductors

Without semiconductor materials modern day electronics would be impossible. A key physical property of a semiconductor is its **band gap**. The band gap is often used to determine if a given semiconductor is suitable for device applications, such as, for example, **solar cells** or **light emitting diodes (LED)**[1]. Here we will explore the origin of band gaps in semiconductors, how it affects device applications and introduce the concept of **effective mass** of charge carriers (e.g. electrons).

Band gap and its origin

Consider a **free electron material** where an electron experiences a **uniform potential** (we are essentially ignoring the periodic potential of the atomic nuclei). The electron energy varies as k^2, where k is the magnitude of the wavevector. When the periodic potential due to the atomic nuclei is introduced energy gaps appear at the **Brillouin zone boundaries** (Figure 1a). This is variously described as being due to Bragg reflection and interference of the $\pm\mathbf{G}/2$ Fourier components or lifting of the degeneracy of electronic states (FoP3B Part 1).

Now due to **Bloch's theorem** (FoP3B Part 1) the electronic structure can be fully described using only wavevectors within the first Brillouin zone. This gives rise to the **Reduced Zone scheme** and **electronic bands** for a given wavevector (Figure 1b).

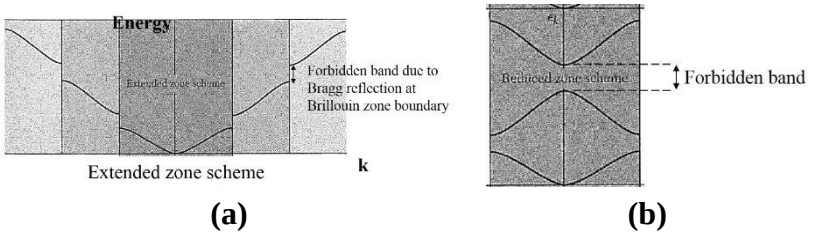


Figure 1: (a) Energy vs. wavenumber k for a periodic crystal and (b) the same diagram in the Reduced Zone scheme. The forbidden band is the semiconductor band gap.

The electronic bands are filled such that the energy is minimised, i.e. from the bottom up. Importantly it can be shown that **each primitive cell can only have two of its electrons in a given energy band**. Consider a typical semiconductor such as silicon. Silicon has an atomic number of 14 and therefore 14 electrons. With 2 atoms in a primitive cell this means that the first 14 electronic bands are completely full. The next available band is empty and separated by an energy gap. We call the highest *occupied* band the **valence band** and the lowest *unoccupied* band the **conduction band**. The **band gap** is the *minimum* energy separation between the two bands, and corresponds to the energy gap between the *valence band maximum* and *conduction band minimum*. **There are no electronic states within the band gap**. **Typical values for semiconductor band gaps are in the range 1-4 eV.**

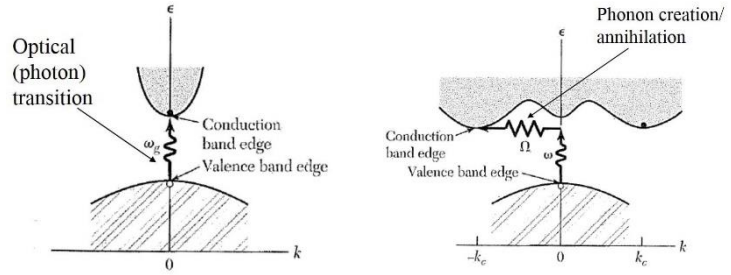
Interaction of light with a semiconductor

The presence of a band gap means that a semiconductor interacts with light in a unique way. The key point to bear in mind is that **a photon has energy but negligible momentum**. When light is *absorbed* an electron in the solid is promoted to a higher energy level that must also be

[1] A solar cell converts sunlight into electricity while an LED does the opposite.

unoccupied. This means that t**he minimum photon energy for light absorption is equal to the band gap**. If photons have lower energy then the light can only be *transmitted* through the solid or *reflected*. In a **direct band gap** semiconductor the valence band maximum has the same wavevector as the conduction band minimum (Figure 2a). For an electronic transition to take place in a direct band gap semiconductor only energy needs to be supplied, since the electron momentum is unchanged. These conditions are readily satisfied by photons of the appropriate energy, and consequently light absorption (as well as the reverse process of light emission) is strong in direct band gap semiconductors. This makes them useful for optoelectronic applications, such as solar cells and LEDs.

Not all semiconductors have a direct band gap. In some cases the wavevector for the conduction band minimum does not overlap with the valence band maximum (Figure 2b). These are known as **indirect band gap** semiconductors[2]. An electronic transition here involves a change in electron momentum. To conserve momentum during light absorption **phonons** from the crystal must therefore be either created or destroyed. **Phonons have momentum but negligible energy (meV)**, which is why the phonon interaction is represented by a horizontal transition in Figure 2b (conversely the light interaction is a vertical transition). Since photons *and* phonons are involved light absorption and light emission are relatively weak in indirect bandgap semiconductors.



(a) **(b)**

Figure 2: Light absorption in (a) direct and (b) indirect band gap semiconductor.

The band gap of semiconductor materials can usually be tuned to the desired energy through **alloying**, which is the process of mixing two different components to form a **solid solution**. For example, the alloy $Al_xGa_{1-x}As$ can be formed by combining x fraction of AlAs with (1-x) fraction of GaAs. The band gap (E_g) for an alloy of two components 'A' and 'B' usually has the following empirical relationship:

$$E_{g,AB} = xE_{g,A} + (1-x)E_{g,B} - bx(1-x)$$

where $E_{g,A}$ and $E_{g,B}$ are the band gaps for the pure components 'A' and 'B' and b is a constant known as the **bowing parameter**.

A further consideration in device applications is the **lattice parameter** of the semiconductor material. Frequently the active semiconductor material must be deposited as a nanometre or micrometre thick film on a support substrate (this is often due to the financial cost of producing such devices, but not always). If there is a large difference in lattice parameter between

[2] You might be wondering why the band structure in Figure 2b looks different to Figure 1b. The answer is that the latter assumes a weak periodic potential and is therefore typically accurate at only small values of k. The band structures of real materials are more complex!

substrate and film then significant strain can build up in the latter, leading to poorer quality devices. The lattice parameter (a) for an AB-alloy frequently obeys **Vegard's law**:

$$a_{AB} = xa_A + (1 - x)a_B$$

For an ideal optoelectronic device the semiconductor must have a direct band gap with energy that matches the intended application. A suitable substrate must also be found that is lattice matched to the semiconductor.

Effective mass of electrons

Charge transport is another key property governing device applications. An important concept is the **effective mass** of the charge carrying electrons. A simple way to think about effective mass is as follows. Consider first an electron in free space. We apply a force F and measure the acceleration a. The mass can then be determined from $F = ma$. Now consider an electron in a real crystal subject to the same external force. As the electron moves through the crystal it will undergo Coulomb scattering with other electrons as well as atomic nuclei. The presence of these *internal* forces means that $F \neq ma$ (recall that F is the *external* applied force). To satisfy Newton's second law we can therefore write $F = m^*a$, where m^* is the *effective* mass. It can be shown that for an isotropic crystal (see DUO supplementary material):

$$m^* = \frac{\hbar^2}{(d^2E/dk^2)}$$

where d^2E/dk^2 is the curvature of electronic band (note that for anisotropic crystals we have to define an effective mass *tensor*).

FoP3B Part II Lecture 2: Electrons and holes

The primary application for semiconductor materials is electronic devices, such as computers, mobile phones etc. It is therefore important to understand how a semiconductor conducts electricity at a fundamental level. Here we explain how **electrons** and **holes** carry charge in a material (see below if you are not familiar with holes). We shall do this using electronic band structure diagrams, relying on **three key equations**:

$$\mathbf{j} = ne\mathbf{v} \quad \ldots (1)$$

$$\mathrm{v} = \frac{1}{\hbar}\frac{dE}{dk} \quad \ldots (2)$$

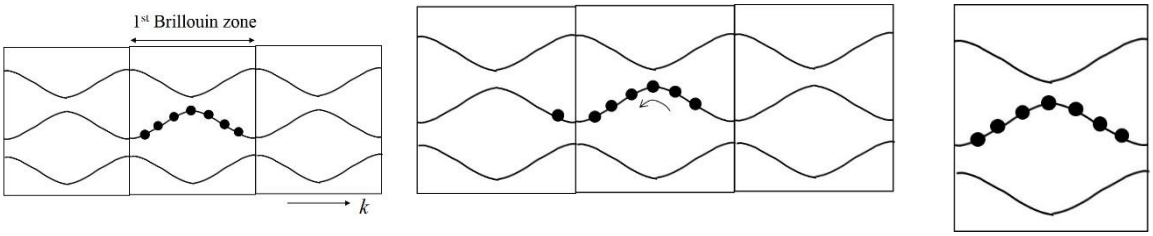
$$\mathbf{F} = \hbar\frac{d\mathbf{k}}{dt} \quad \ldots (3)$$

Equation (1) is the current density **j** expressed as a vector, with *n* being the number density of charge carriers of charge *e* per unit volume that have velocity **v**. Equations (2) and (3) represent the group velocity and force on electrons respectively (you came across the last two equations in Lecture 1). The velocity of an electron is proportional to the gradient of the *E*-*k* diagram.

From Equation (1) it follows that for net electrical conduction $\Sigma\mathbf{v} \neq 0$, *where the summation is over all electrons*. *Similarly we must have* $\Sigma\mathbf{k} \neq 0$, *where* **k** *is the electron wavevector; this means that electrons have a net momentum when conducting electricity.*

Conduction in a completely full band

Let us apply the above principles to electrons in a completely full band under an applied electric field (the driving force for producing an electric current). *Since the electron is negatively charged the force will be in the opposite direction to the electric field*. Before the electric field is applied the entire band is full and therefore $\Sigma\mathbf{v} = 0$ and $\Sigma\mathbf{k} = 0$, i.e. there is no net current as required (Figure 1a). Assume the electric field is applied in the +*k* direction, so that from Equation (3) the electron **k**-vector would decrease (Figure 1b). Because we only need to consider electrons in the first Brillouin zone the Reduced Zone scheme still gives a full band with $\Sigma\mathbf{v} = 0$ and $\Sigma\mathbf{k} = 0$ (Figure 1c). **A completely full band cannot therefore conduct electricity**. **Insulators (e.g. wood, plastic) do not conduct electricity due to filled bands.**



(a) **(b)** **(c)**

Figure 1: Completely full band (a) before and (b) after applying an electric field in +*k* direction. (c) is the Reduced Zone scheme representation of (b)

Conduction in a nearly empty band

Consider the case where only a few electronic states in a band are occupied (Figure 2a). Examples are (i) metals and (ii) semiconductor conduction band after light absorption. Going through the same arguments described previously it should be clear that even in the Reduced Zone scheme $\Sigma\mathbf{v} \neq 0$ and $\Sigma\mathbf{k} \neq 0$ under an applied electric field, which is the condition for conduction. However, there is a subtlety. If Equation (3) is integrated over time the electron wavevector will cycle through all **k**-values within the first Brillouin zone. Now from Equation (2) the velocity of electrons change sign in opposite halves of the first Brillouin zone (Figure 2b). *The time integrated electron velocity and hence electric current is therefore zero!* We do however know that metals are good conductors, so something is missing from our model.

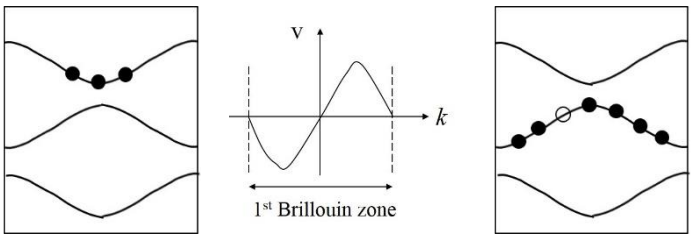


Figure 2: (a) nearly empty band and (b) the electron velocity within the first Brillouin zone. (c) shows a nearly full band, with the missing electron denoted by an empty circle.

The answer lies in electron **scattering**. *In real materials thermal energy cause atoms to vibrate, and electrons can scatter off them such that there is no change in the time averaged position (Figure 3a).* This is however not the case when an electric field is applied (Figure 3b). Then the electrons acquire a **drift velocity** ($\mathbf{v}_d$) given by:

$$\mathbf{v}_d = -\mu\mathbf{E}$$

where μ is the **mobility**. *The negative sign indicates that the negatively charged electrons drift in the opposite direction to the electric field* **E**. **The fact that electrons acquire a drift velocity and hence net momentum means that the Fermi surface shifts when an electric field is applied (Figure 3c)**. *Hence* $\Sigma\langle\mathbf{v}\rangle \neq 0$ *and* $\Sigma\langle\mathbf{k}\rangle \neq 0$, *where the* $\langle\ \rangle$ *sign represents a time averaged quantity*. **We conclude that electrons in a nearly empty band can conduct electricity with the aid of scattering**.

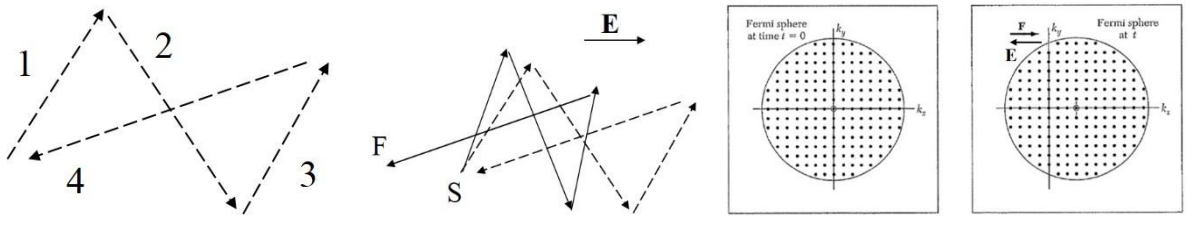


Figure 3: (a) Electron scattering in a material with no electric field. The numeral represents successive trajectories between scattering events. The solid lines in (b) show how the same trajectories are modified under an applied electric field **E**. 'S' and 'F' denote the electron start and finish positions. The shifting of the Fermi surface under an electric field is shown in (c).

Conduction in a nearly full band- the concept of holes

A full band cannot conduct electricity. Let us however consider a nearly full band (Figure 2c). This may occur in (i) metals and (ii) semiconductor valence band after light absorption. Inspecting Figure 2c it is clear that $\Sigma\mathbf{v} \neq 0$ and $\Sigma\mathbf{k} \neq 0$ under an applied field, so that a nearly full band can conduct electricity with the aid of scattering. However, **instead of describing conduction in terms of *all* electrons in the band it is possible to model it using a *single* pseudo-particle called a hole**.

This is best illustrated using Figure 4. The missing electron is initially at the top of the band when an electric field is applied in the $+k$ direction. From Equation (3) this results in the electrons shifting to lower k-values (recall that electrons are negatively charged). The same is also true for the missing electron. *We have* $\Sigma\mathbf{k} = -\mathbf{k}_e$, *where* $\mathbf{k}_e$ *is the wavevector of the missing electron*. Since the hole describes motion of all electrons in the band its wavevector will be $\Sigma\mathbf{k}$ or $-\mathbf{k}_e$. We now consider the energy of the hole so that we may construct a (fictitious) E-k diagram for the **hole band**. **A band with a missing electron is in an excited state and the fact that the missing electron moves into progressively lower energy states with applied field means that the hole band must be a mirror reflection of the nearly full electron band**. This is illustrated in Figure 4. It is also clear from the figure that the hole moves in the opposite direction to electrons under an applied field, i.e. a hole has *positive* charge.

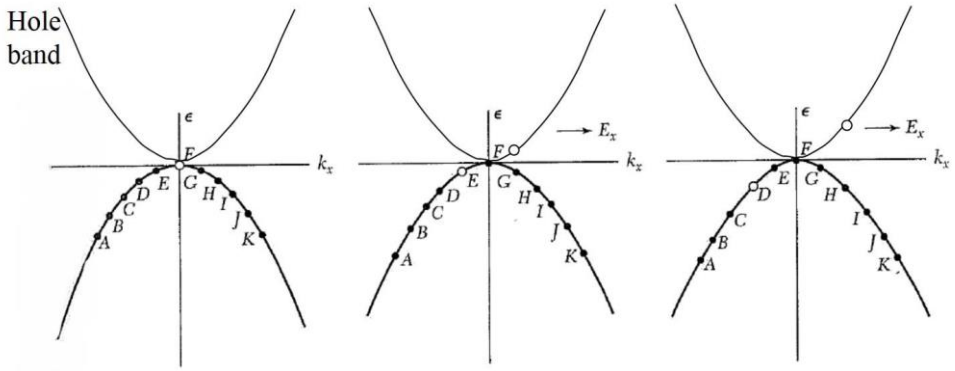


Figure 4: Electrons in a nearly full band under an electric field applied in the $+k$ direction and their description using a pseudo-particle hole band.

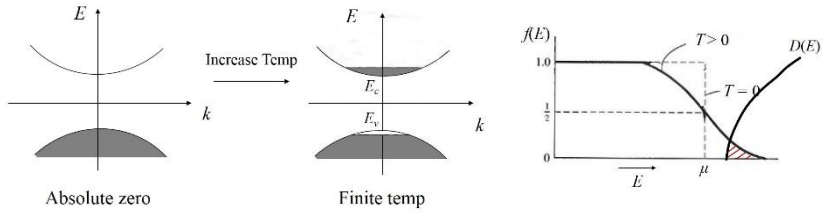
NB: A hole arises due to a missing electron in a nearly full electron band. However, it is NOT the missing electron itself. Instead it is a pseudo-particle of positive charge that we use for our own convenience to describe the net motion of all other electrons in the band.

For this reason it should be clear to you that a completely empty band does not consist of holes and therefore is not conducting.

FoP3B Part II Lecture 3: Statistical Physics of Semiconductors

Both **electrons** *in a nearly empty band* and **holes** *in a nearly full band* conduct electricity. Clearly the more electrons and holes there are the better the conductivity. However, **what determines the electron, hole concentrations in a semiconductor?** Here we will restrict our attention to semiconductors in **thermal equilibrium** with their surroundings.

The problem is illustrated schematically in Figure 1a. At 0K temperature the **valence band** is completely full and the **conduction band** completely empty. The semiconductor therefore shows insulating behaviour. As the temperature is raised a few valence electrons gain enough thermal energy to be promoted across the **band gap** into the conduction band. *Since the semiconductor is in equilibrium the promoted conduction band electrons will occupy the lowest energy states*. **The smaller the band gap and the higher the temperature the larger the electron and hole concentrations**.



(a) **(b)**

Figure 1: (a) thermal excitation of valence band electrons into the conduction band. (b) The Fermi-Dirac function *f(E)* for absolute zero and non-zero temperatures. Also shown is the density of states curve. The shaded area determines the electron concentration.

Electron concentrations

Since the semiconductor is in equilibrium, the electron concentration can be calculated using statistical physics. **We need to know (i) the number of energy states in the conduction band and (ii) the probability of occupying a given energy state**. The number of states per unit volume between E and $E+dE$ is simply $g(E)dE$, where $g(E)$ is the **density of states**. Let $f(E)$ denote the occupation probability of states with energy E. **Multiplying $g(E)dE$ by $f(E)$ gives the number of electrons between E and $E+dE$**.

Let us now derive expressions for $g(E)$ and $f(E)$. The density of states for a **free electron solid** can be shown to be (see FoP3B Part 1):

$$g(E) = \frac{1}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E} \qquad \ldots (1)$$

We assume that the density of states for the conduction band in a semiconductor has the same general form as Equation (1). This might seem strange since the semiconductor is not a free electron solid. *Nevertheless this is a reasonable approximation since in most cases there are only few electrons in the conduction band*[1], *meaning that only states near the bottom of the conduction band are occupied*. *The fact that the Brillouin zone edges are avoided means that*

[1] kT at room temperature is 25 meV, while the semiconductor band gap is in the range 1-4 eV. Hence there is a large barrier for thermal excitation and consequently only a few electrons are promoted across the band gap.

we can treat the electrons as being effectively free. However, some modifications to Equation (1) are necessary. In fact **for a conduction band the density of states $g_e(E)$**:

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

Note that the free electron mass *m* has been replaced by the **effective mass** m_e* and the energy scale has been modified to take into account that the bottom of the conduction band is at energy E_c (Figure 1a).

Since electrons are **fermions** the occupation probability is given by the **Fermi-Dirac distribution function**:

$$f(E) = \frac{1}{1 + \exp\left(\frac{E - \mu}{kT}\right)} \quad \ldots (3)$$

Here μ is the **chemical potential**, which at most temperatures is close to the **Fermi energy**. *From Equation (3) μ can be defined as the energy at which the electron occupation probability is ½.* $f(E)$ is shown schematically in Figure 1b. At 0 K it is a step function such that all states below μ are occupied, while all states above μ are unoccupied. At higher temperatures the step becomes increasingly 'smeared', such that the transition from occupied to unoccupied states is no longer abrupt.

To determine the electron concentration 'n' we integrate $g_e(E)f(E)$ over the conduction band:

$$n = \int_{E_c}^{E_{max}} g_e(E) f(E) dE \quad \ldots (4)$$

where E_{max} is the maximum energy of the conduction band. To solve Equation (4) *analytically* we make **two assumptions**:

(i) **we extend the upper limit of the integral from E_{max} to infinity**. This might seem a bad approximation at first, but note that $f(E)$ decreases exponentially beyond μ. Hence the term $g_e(E)f(E)$ tends to zero for large energy.

(ii) **we assume $(E_c - \mu) >> kT$** (i.e. the chemical potential is not too close to the conduction band minimum), **so that from Equation (3), $f(E) \approx \exp[-(E-\mu)/kT]$**.

Substituting Equations (2), (3) into (4) and making a change of variable $u = (E-E_c)/kT$ gives:

$$n = \frac{1}{2\pi^2}\left(\frac{2m_e^* kT}{\hbar^2}\right)^{3/2} \exp\left[-\frac{(E_c - \mu)}{kT}\right] \int_0^\infty \sqrt{u} \exp(-u) du$$

The integral is equal to $(\sqrt{\pi})/2$, so that:

$$\boldsymbol{n = N_c \exp\left[-\frac{(E_c - \mu)}{kT}\right]} \;\; ; \;\; N_c = 2\left(\frac{m_e^* kT}{2\pi\hbar^2}\right)^{3/2} \quad \ldots (5)$$

N_c is the **conduction band effective density of states**. *Equation (5) indicates that the electron concentration increases with temperature T, consistent with what we would expect*. Furthermore, the *electron concentration increases as the separation between the chemical potential μ and conduction band minimum E_c decreases*. This should be clear from Figure 1b, where *n* is determined by the overlap region of the $g_e(E)$ and $f(E)$ curves (Equation 4).

Hole concentrations

To obtain the hole concentration we multiply the valence band density of states by the probability of states being <u>*unoccupied*</u>. **The hole concentration between E and $E+dE$ is therefore $g_h(E)[1-f(E)]dE$, where $g_h(E)$ is the valence band density of states and $[1-f(E)]$ is the probability of a state being unoccupied**. Using a similar derivation to that of the electron concentration we obtain for the **hole concentration p**:

$$\boldsymbol{p = N_v \exp\left[-\frac{(\mu - E_v)}{kT}\right]} \;\; ; \;\; N_v = 2\left(\frac{m_h^* kT}{2\pi\hbar^2}\right)^{3/2} \quad \ldots (6)$$

where N_v is the **valence band effective density of states** and E_v is the valence band maximum. You should attempt deriving Equation (6) yourself. Similar to electrons, the *hole concentration increases with temperature and decreasing separation between chemical potential μ and valence band maximum E_v.*

Law of mass action

Taking the product of *n* and *p* gives:

$$np = N_c N_v \exp\left(-\frac{E_g}{kT}\right) \propto T^3 \exp\left(-\frac{E_g}{kT}\right) \quad \ldots (7)$$

where $E_g = E_c - E_v$ is the band gap. **Equation (7) is the law of mass action**. It states that the product *np* is a material constant at a given temperature. Furthermore, ***np increases monotonically with decreasing band gap and increasing temperature***. In fact, the temperature dependence is strong, due to the exponential term. <u>*In semiconductors the increase in electrical conductivity with temperature is due to the rapid increase in electron, hole concentrations, which offsets the negative effects of increased scattering at higher temperatures*</u>. *Compare this with metals where the electron, hole concentration is approximately independent of temperature, so that conductivity decreases with temperature due to increased scattering*.

Fermi level position

Thus far nothing has been said about the Fermi energy level or chemical potential μ. Consider an **intrinsic semiconductor**, i.e. a semiconductor with no *impurities* (e.g pure Si, GaAs etc). *For an intrinsic semiconductor all electrons and holes are generated by thermal excitation across the band gap, and consequently n = p.* From Equations (5) and (6) it therefore follows that:

$$\mu = E_{mid-gap} + \frac{3}{4} kT ln\left(\frac{m_h^*}{m_e^*}\right) \quad \ldots (8)$$

where $E_{\text{mid-gap}} = (E_c+E_v)/2$ is the middle of the band gap. Since m_h* is typically of the same order as m_e* the **chemical potential μ for an intrinsic semiconductor is close to the band gap mid-point**. Since $E_F \approx \mu$ this is also approximately the position of the Fermi energy level.

FoP3B Part II Lecture 4: Extrinsic Semiconductors

The electrical conductivity of a semiconductor depends on its electron and hole concentrations, which from the **law of mass action** can be controlled by varying the band gap and/or temperature. However, these two parameters are fixed according to the device application (e.g. a blue light emitting diode operating at room temperature). An alternative way of changing the charge carrier concentration is therefore required. This is achieved through the process of **doping**, i.e. *the controlled addition of 'impurities' or **dopants** to a pure **intrinsic** semiconductor*. A doped semiconductor is called an **extrinsic** semiconductor. Dopants generate either an excess of electrons or an excess of holes. *Dopants that provide excess electrons are called **n-type** dopants or **donors**. **p-type** dopants (or **acceptors**) provide excess holes.*

n-type doping (excess electrons)

Consider silicon as an example of an intrinsic semiconductor. **Silicon is a group IV element**. *In order to complete an octet of electrons and form a stable compound a single silicon atom must covalently bond with four neighbouring atoms* (Figure 1a). This results in the familiar **diamond cubic** crystal structure of silicon. Now assume a Group V element (e.g. nitrogen, phosphorous) is added to the silicon lattice. *The extra electron from the Group V element will orbit around the impurity atom (Figure 1b)*. The energy of this extra electron and the orbiting radius can be calculated approximately using Bohr's model of the atom. *Due to the high* **dielectric constant ε_r** *for a semiconductor (e.g. ε_r = 11.7 for silicon) the energy is small (meV) and the orbit radius large (several nm).* **The extra electron is weakly bound compared to valence electrons, but not as free as conduction electrons, and so occupies an energy level slightly below the conduction band (Figure 1c)**. This is known as the **donor energy level**.

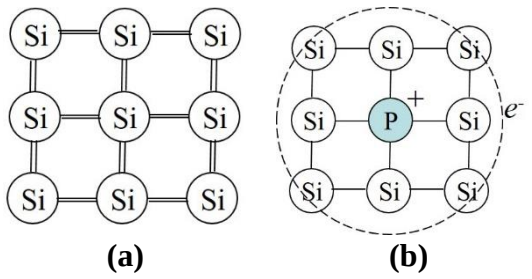
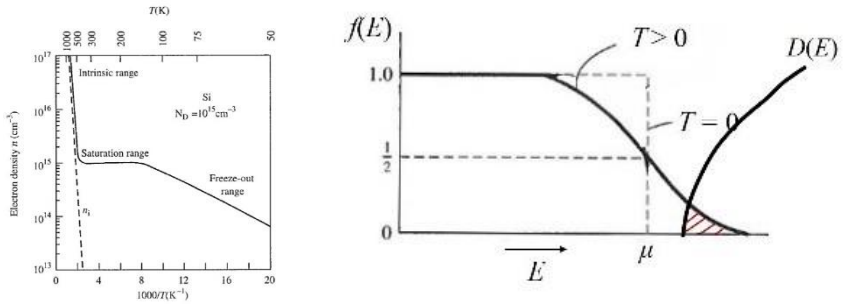


Fig 1: (a) 2D projection of a perfect Si crystal and (b) with a Group V impurity. The extra electron occupies the donor energy level E_d below the conduction band (c).

Let us examine the effect of temperature on Group V doped silicon. **The key point to bear in mind is that the binding energy of donor electrons is only meV and so can easily be promoted into the conduction band via thermal excitation[1]**. At absolute zero all donor electrons occupy the donor energy level, i.e. they remain loosely bound to the Group V impurity. This is the so-called **freeze-out regime**. As the temperature is raised more and more donor electrons can be promoted to the conduction band leaving the Group V donor atoms **ionised** (i.e. they acquire positive charge due to the missing electron). Eventually all donor atoms are ionised and the material is in the **saturation regime**. *At room temperature most semiconductors are in the saturation regime, where the donor energy level is largely empty*. As the temperature is increased still further the only possible electronic transition is from the valence band to the conduction band. This process is comparatively inefficient since the

[1] Recall kT at room temperature is ~25 meV.

electronic transition must take place across the (moderately large) 1-4 eV **band gap**. Nevertheless provided the temperature is sufficiently high (e.g. several 100°C) the material enters the **intrinsic regime**, where the number of electrons generated by thermal excitation across the band gap far outnumber the excess electrons due to dopant atoms. *Due to the nature of excitation across the band gap the electron and hole concentrations are approximately equal in the intrinsic regime.* The different regimes for an *n*-type semiconductor are shown in Figure 2a.



(a) **(b)**

Fig 2: (a) Electron concentration as a function of temperature for a *n*-type semiconductor. (b) The Fermi-Dirac distribution function and conduction band density of states. The overlap region (shaded) determines the electron concentration.

It is possible to calculate the electron and hole concentrations in the saturation regime due to *n*-type doping (we focus on the saturation regime since this corresponds to room temperature). From the law of mass action the electron (*n*) and hole (*p*) product *np* is given by:

$$np = N_c N_v \exp\left(-\frac{E_g}{kT}\right) \quad \ldots (1)$$

(see Lecture 3 summary for a definition of terms). <u>*np* is therefore constant for a given semiconductor at fixed temperature</u>. *Equation (1) was derived from statistical physics and is valid irrespective of whether the semiconductor is pure (i.e. intrinsic) or doped (i.e. extrinsic).* If the semiconductor is intrinsic $n = p$, since electrons and holes are generated equally by thermally exciting a valence band electron into the conduction band. Let us denote by n_i the **intrinsic electron (or intrinsic hole) concentration** for the undoped semiconductor. Then for the doped, extrinsic semiconductor we can write:

$$np = n_i^2 \quad \ldots (2)$$

Comparing (1) and (2) it is clear that $n_i = \sqrt{N_c N_v}\exp(-\frac{E_g}{2kT})$.

The semiconductor must also be charge neutral. In an *n*-type semiconductor the positive charges are due to holes (*p*) and ionised donors. In the saturation regime the latter is equal to the Group V **donor atom concentration N_D**. The negative charge is due to conduction band electrons (*n*). Therefore:

$$n = p + N_D \quad \ldots (3)$$

Using (2) and (3) to solve for n and assuming[2] $N_D >> n_i$ gives $n \sim N_D$. Hence from (2) $p \sim n_i^2/N_D$. We have for an *n*-type semiconductor $n > n_i$ and $p = n_i(n_i/N_D) < n_i$, so that $n > p$. In an *n*-type extrinsic semiconductor electrons are **majority carriers** and holes are **minority carriers** (cf. an intrinsic semiconductor where $n = p$).

It is also possible to calculate the **chemical potential** μ for an *n*-type semiconductor in the saturation regime. The electron concentration n is given by:

$$n = N_c \exp\left[-\frac{(E_c - \mu)}{kT}\right] \quad \ldots (4)$$

(see Lecture 3 summary). Substituting $n = N_D$ and re-arranging for μ gives:

$$\mu = E_c - kT \ln\left(\frac{N_c}{N_D}\right) \quad \ldots (5)$$

This indicates that the chemical potential μ lies below the conduction band minimum E_c. In fact for an intrinsic semiconductor, where $n = p$, μ is located close to the mid-gap position. Therefore for an *n*-type extrinsic semiconductor, where $n > p$, μ must move closer to the conduction band, since from Figure 2b the electron concentration is determined over the energy range where the conduction band **density of states** and **Fermi-Dirac distribution function** overlap (see Lecture 3 summary). From Equation (5) the higher the doping concentration N_D the larger the value of n and consequently the closer μ is to E_c.

p-type doping (excess holes)

Consider adding a Group III element (e.g. boron, aluminium) to a silicon lattice. The Group III electron is now missing an electron in order to complete its full octet. This missing electron can be provided by a neighbouring silicon bond (Figure 3a). *However, this would result in the Group III atom being negatively charged (i.e. 'ionised') and hence the extra electron will be bound but higher in energy than a valence electron*. The electron therefore occupies an **acceptor energy level** that is above the valence band (Figure 3b).

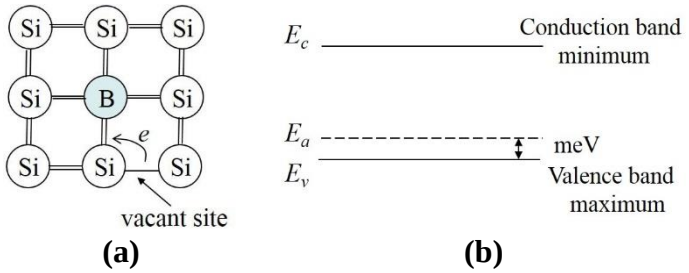


Fig 3: (a) 2D projection of a Si crystal with a Group III impurity. The extra electron required to complete the octet occupies the acceptor energy level E_a above the valence band.

[2] This assumption is justified since n_i is relatively small (recall that thermal excitation across the band gap at room temperature is not an efficient process). Dopant concentrations even at parts per million is therefore sufficient to satisfy the condition $N_D >> n_i$.

The effect of temperature on the hole concentration is similar to the arguments presented for an *n*-type semiconductor. *At absolute zero the valence band is completely full and there are no electrons in the acceptor energy level, i.e. the Group III impurities have not acquired an electron from neighbouring bonds*. This is the freeze-out regime. *As the temperature is raised first valence band electrons are promoted to the acceptor energy level*, i.e. the Group III impurities gain an electron and become 'ionised'. *The partially full valence band can now conduct electricity due to holes*. *With increasing temperature all acceptor energy states become occupied resulting in the saturation regime*. *Increasing the temperature still further results in the intrinsic regime where the electrons and holes are primarily due to thermal excitation across the band gap*.

The hole concentration for a *p*-type semiconductor in the saturation regime is derived using the same arguments of law of mass action and charge conservation. Assuming $N_A >> n_i$ we find that $p \sim N_A$ and $n \sim n_i^2/N_A$, where N_A is the Group III **acceptor atom concentration**. *Thus for a p-type semiconductor holes are the majority carriers and electrons the minority carriers*.

From the equation for the hole concentration p the chemical potential μ is found to be:

$$\mu = E_v + kT\ln\left(\frac{N_v}{N_A}\right) \qquad \ldots (6)$$

Thus μ is above the valence band maximum E_v. This is easily understood by noting that $p > n$ and that the hole concentration is due to overlap of the valence band density of states and probability of *unoccupation* (i.e. [1-$f(E)$]) curves.

FoP3B Part II Lecture 5: pn junction (part I)

The **pn junction** is the basic building block of semiconductor devices. It is used in solar cells, light emitting diodes as well as part of more complex device architectures, such as bipolar transistors. In this and the next lecture we will examine the structure of a pn junction and how this influences electric current transport.

A pn junction is formed by bringing together *n*- and *p*-type material to form an **interface** (Figure 1a)[1]. The electron and hole concentration either side of the interface is different, which results in **majority carrier** electrons in the *n*-layer flowing into the *p*-layer, a process called **diffusion**. Similarly, majority carrier holes will diffuse from the *p*-layer into the *n*-layer. During this process *ionised* **donor** and **acceptor** atoms are left behind. For example, removal of the donor electron from the *n*-layer means that the layer acquires a net *positive* charge due to the Group V ionised donors. Similarly, removal of a hole from the *p*-layer, or equivalently acceptance of an electron from the n-layer, results in *negative* charge due to ionised Group III acceptors. The diffusion process takes place over a region close to the interface and is known as the **space charge region** or **depletion region** (Figure 1b).

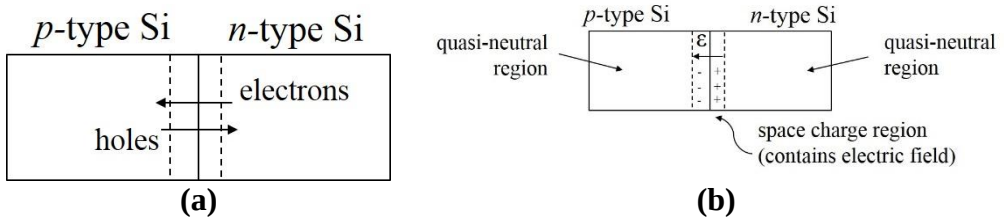


Fig 1: (a) and (b) is a schematic of a silicon pn junction.

The presence of (unscreened) ionised donors and acceptors means that there is an internal electric field or equivalently potential difference within the space charge region. The electric field prevents further diffusion of electrons and holes. The region outside the space charge region is called the **quasi-neutral region**. The space charge region is an important feature of a pn junction. Its properties, i.e. electric field and potential, can be calculated using standard electrostatic methods provided we know the charge distribution. However, this is quite a complex problem, since the charge carrying particles consist of conduction band electrons, valence band holes as well as ionised donors and acceptors. The **depletion approximation** is therefore used to simplify the problem. *This states that the space charge region is free of electrons and holes. The only charged particles are therefore ionised acceptors on the p-side and ionised donors on the n-side. Furthermore, we assume that all donors and acceptors in the space charge region are ionised. This is reasonable since at room temperature the material is in the* ***saturation regime****.*

Calculation of Electric Field

Let us now calculate the electric field distribution within the space charge region. This can be done using **Gauss' Law**:

[1] In reality pn junctions are formed by starting with one material, say p-type silicon, and doping the surface n-type by either ion implantation or diffusion of donor 'impurities'.

$$\vec{\nabla} \cdot \vec{\varepsilon} = \frac{d\varepsilon}{dx} = \frac{\rho(x)}{\epsilon_r \epsilon_0} \quad \ldots (1)$$

where ε is the electric field, ϵ_r is the **dielectric constant** or relative permittivity, ϵ_0 is the permittivity of free space and ρ is the charge density. Due to the depletion approximation ρ takes the form:

$$\rho(x) = \begin{Bmatrix} -eN_A \\ eN_D \end{Bmatrix} \begin{matrix} -w_p < x < 0 \ (p-\text{side}) \\ 0 < x < w_n \ (n-\text{side}) \end{matrix} \quad \ldots (2)$$

where $-w_p$ and w_n are the space charge boundaries for *p*- and *n*-sides (Figure 2a). N_A and N_D are acceptor and donor concentrations respectively (note that the ionised acceptors on the *p*-side are negatively charged and hence $\rho(x) = -eN_A$ in this region). The electric field in the *p*-side is determined by integrating Equation (1):

$$\varepsilon(x) = \int \frac{-eN_A}{\epsilon_r \epsilon_0} dx = \frac{-eN_A}{\epsilon_r \epsilon_0} x + A \quad \ldots (3)$$

where *A* is the constant of integration. To determine its value we make use of the **boundary condition** that *the electric field at the space charge edge is zero*, i.e. $\varepsilon(-w_p) = 0$. Using this condition we find:

$$\varepsilon(x) = \frac{-eN_A}{\epsilon_r \epsilon_0}(x + w_p) \quad \ldots (4)$$

Using the boundary condition $\varepsilon(w_n) = 0$, the electric field on the *n*-side is:

$$\varepsilon(x) = \frac{eN_D}{\epsilon_r \epsilon_0}(x - w_n) \quad \ldots (5)$$

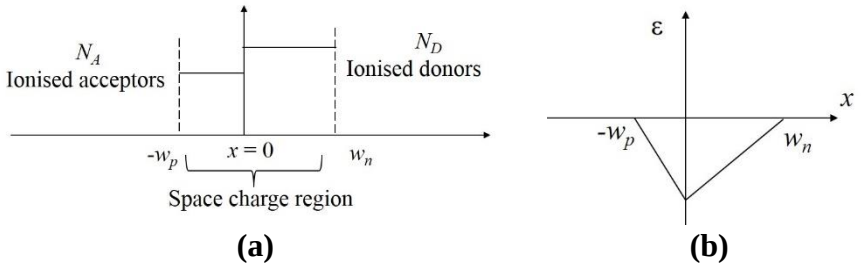


Fig 2: (a) Notation used for calculating space charge region properties and (b) the electric field distribution.

The electric field distribution is sketched in Figure 2b and shows a linear variation either side of the pn junction. It has a negative value since the electric field points in the $-x$ direction (Figure 1b). Note that the electric field must be continuous at the pn junction interface in order

to be consistent with **Maxwell's boundary conditions**[2] for the electric field. Equating (4) and (5) for $x = 0$ gives:

$$N_A w_p = N_D w_n \quad \ldots (6)$$

This is the condition for **charge neutrality**, i.e. *it states that the negative charge due to ionised acceptors is equal to the positive charge due to ionised donors* (recall that we are using the depletion approximation, where it is assumed that there are no electrons or holes within the space charge region).

Calculation of Potential Field

The electric potential ϕ is easily calculated from the electric field using:

$$\vec{\varepsilon} = -\vec{\nabla}\phi \quad \ldots (7)$$

For example, the *p*-side potential is obtained by integrating the electric field in Equation (4):

$$\phi(x) = \frac{eN_A}{\epsilon_r \epsilon_0} \int (x + w_p) dx = \frac{eN_A}{2\epsilon_r \epsilon_0} (x^2 + 2w_p x) + B \quad \ldots (8)$$

where B is a constant of integration. We can *arbitrarily* set the potential at $x = -w_p$ to zero and therefore:

$$\phi(x) = \frac{eN_A}{2\epsilon_r \epsilon_0} (x + w_p)^2 \quad \ldots (9)$$

Similarly, the *n*-side potential can be shown to be:

$$\phi(x) = \phi_{bi} - \frac{eN_D}{2\epsilon_r \epsilon_0} (x - w_n)^2 \quad \ldots (10)$$

where ϕ_{bi} is the **built-in potential** at the $x = w_n$ space charge edge on the *n*-side. Figure 3a is a schematic of the potential field. The potential energy E of an electron in a potential field ϕ is given by $E = -e\phi$, i.e. the energy profile is inverted w.r.t the potential (Figure 3b).

[2] Strictly speaking this is a special case of Maxwell's boundary condition and is only valid when the *p*- and *n*-sides are of the same type of material (e.g. silicon).

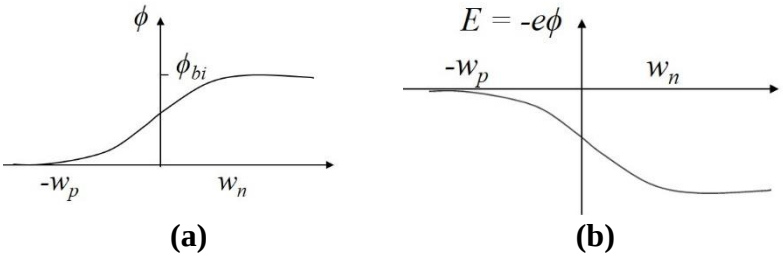


Fig 3: (a) Potential field ϕ and (b) energy E of an electron in the space charge region.

Using Figure 3b we can plot the shape of the conduction band minimum E_c and valence band maximum E_v before and after the *p*- and *n*-side make contact to form the pn junction (see Figure 4). *Before making contact the chemical potential μ of the p- and n-sides are different, but once the pn junction has formed the band bending due to the potential field creates a constant chemical potential throughout the system.* This can be understood by noting that the chemical potential is the free energy (G) change due to the addition of one mole of particles to the system, i.e. $\mu = (dG/dn)$. For the system to be in equilibrium the chemical potential must therefore be constant everywhere, otherwise electrons and holes will move from regions of high to low chemical potential in order to minimise their energy. In fact, the initial process of electron and hole diffusion to create the space charge region can also be thought of as being driven by the difference in chemical potential, forcing the system into a lower energy configuration.

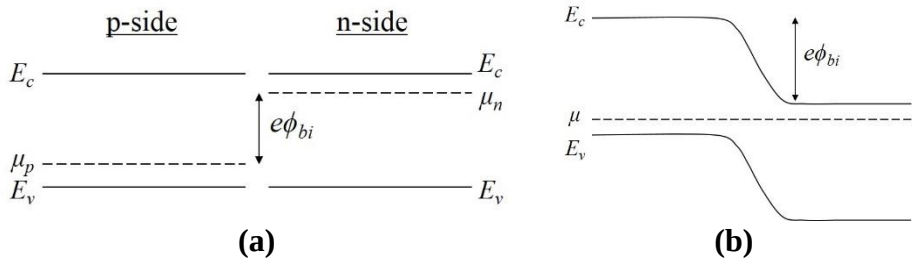


Fig 4: Chemical potential μ and band edges (a) before and (b) after the *p*- and *n*-sides come together to form a pn junction.

FoP3B Part II Lecture 6: pn junction (part II)

In this lecture we will calculate the **built-in potential** ϕ_{bi} and **space charge widths** to complete our discussion on the *equilibrium* pn junction. We will then move onto *non-equilibrium* pn junctions, specifically the application of an external voltage, i.e. **electrical biasing**. It will be shown that current flows for only one voltage polarity, a process known as **rectification**. Finally we will briefly describe the working principles of two pn junction devices, i.e. **solar cells** and **light emitting diodes**.

Calculation of Built-in potential (equilibrium case)

The semiconductor band diagram before and after forming the pn junction is shown in Figure 1. Initially the chemical potentials are at different energy levels (i.e. μ_p for *p*-type and μ_n for *n*-type). For the equilibrium pn junction the chemical potential μ must be uniform everywhere, which gives rise to the built-in potential ϕ_{bi} and band bending. It therefore follows that the energy $e\phi_{bi}$ must be equal to the difference in chemical potential (μ_n-μ_p).

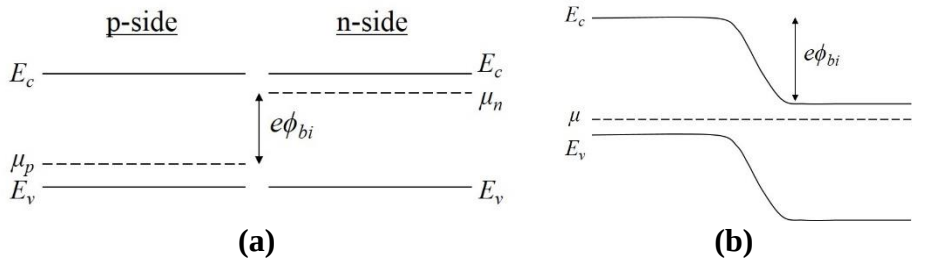


Fig 1: Chemical potential μ and band edges (a) before and (b) after the *p*- and *n*-sides come together to form a pn junction.

From Lecture 4:

$$\mu_p = E_v + kT\ln\left(\frac{N_v}{N_A}\right) \quad \ldots (1)$$

$$\mu_n = E_c - kT\ln\left(\frac{N_c}{N_D}\right) \quad \ldots (2)$$

Therefore:

$$\phi_{bi} = \frac{(\mu_n - \mu_p)}{e} = \frac{E_g}{e} - \frac{kT}{e}\ln\left(\frac{N_v N_c}{N_A N_D}\right) \quad \ldots (3)$$

where $E_g = (E_c\text{-}E_v)$ is the **band gap**. From the **law of mass action**:

$$np = n_i^2 = N_c N_v \exp\left(-\frac{E_g}{kT}\right) \quad \ldots (4)$$

Substituting the expression for E_g derived from (4) in (3) gives:

$$\phi_{bi} = \frac{kT}{e} \ln\left(\frac{N_A N_D}{n_i^2}\right) \quad \ldots (5)$$

Calculation of Space Charge widths (equilibrium case)

From Lecture 5 the potential $\phi(x)$ for the *p*-side was shown to be:

$$\phi(x) = \frac{eN_A}{2\epsilon_r\epsilon_0}\left(x + w_p\right)^2 \quad \ldots (6)$$

And for the *n*-side:

$$\phi(x) = \phi_{bi} - \frac{eN_D}{2\epsilon_r\epsilon_0}(x - w_n)^2 \quad \ldots (7)$$

The potential must be continuous at the pn junction $x = 0$ (Figure 2), so that from (6) and (7):

$$\frac{eN_A}{2\epsilon_r\epsilon_0} w_p^2 = \phi_{bi} - \frac{eN_D}{2\epsilon_r\epsilon_0} w_n^2 \quad \ldots (8)$$

Equation (8) can be solved for either w_n or w_p using the **charge conservation** condition $N_A w_p = N_D w_n$. The final result is:

$$w_p = \left[\frac{2\epsilon_r\epsilon_0\phi_{bi}}{e}\left(\frac{N_D}{N_A}\right)\left(\frac{1}{N_A + N_D}\right)\right]^{1/2}$$
$$w_n = \left[\frac{2\epsilon_r\epsilon_0\phi_{bi}}{e}\left(\frac{N_A}{N_D}\right)\left(\frac{1}{N_A + N_D}\right)\right]^{1/2} \quad \ldots (9)$$

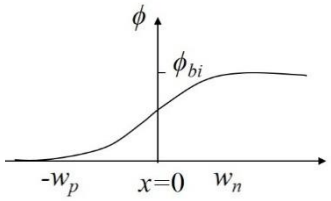


Fig 2: Potential variation across the pn junction. The space charge edges are at $x = -w_p$ (*p*-side) and $x = w_n$ (*n*-side)

Non-equilibrium pn junction: electrical biasing

Consider connecting a pn junction device to a battery. In the **forward bias** configuration the positive terminal of the battery is connected to the *p*-side (Figure 3a). *The external electric field due to the battery is confined to the space charge region and opposes the built-in electric field*[1]. This results in a smaller net electric field and consequently less band bending within the space charge region (Figure 3b). *Furthermore, the pn junction is no longer in equilibrium and consequently the chemical potential is not constant*. For forward bias the chemical potential on

[1] Note that the quasi-neutral region cannot contain the external electric field, since otherwise any electrons in the conduction band or holes can drift out of this region creating a new space charge region.

the *p*-side is lower than the *n*-side. This simply means that electrons in the *p*-side have lower energy due to the positive potential of the battery.

The situation is reversed when the positive terminal of the battery is connected to the *n*-side, which is the **reverse bias** condition (Figure 3c). *Here the external electric field is in the same direction as the built-in electric field and consequently the net electric field is larger, leading to more prominent band bending in the space charge region* (Figure 3d). Furthermore, the chemical potential on the *n*-side is now lower than the *p*-side due to the change in voltage polarity.

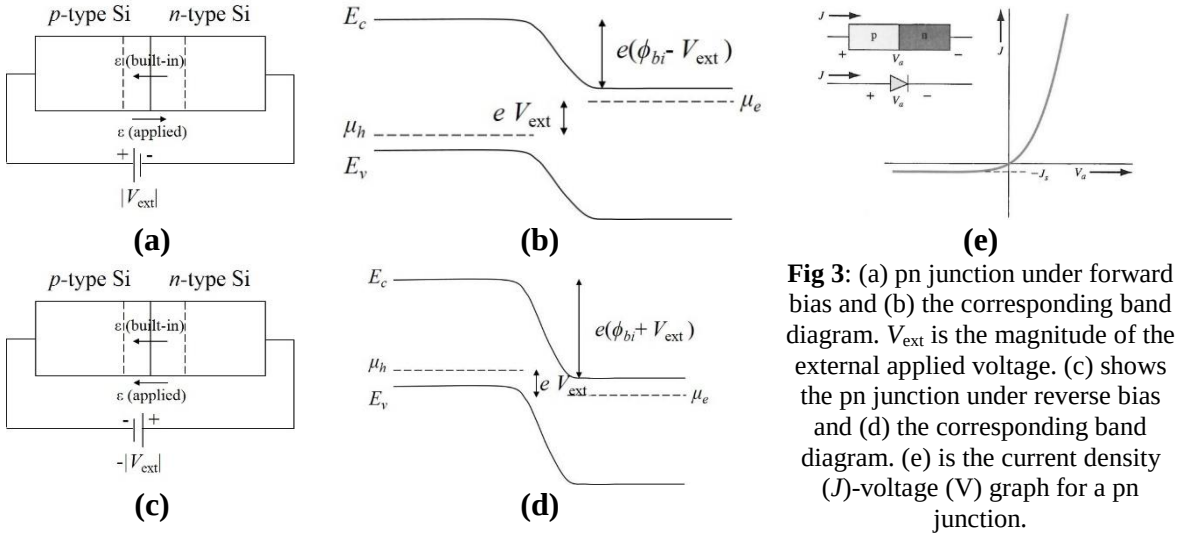


Fig 3: (a) pn junction under forward bias and (b) the corresponding band diagram. V_{ext} is the magnitude of the external applied voltage. (c) shows the pn junction under reverse bias and (d) the corresponding band diagram. (e) is the current density (*J*)-voltage (V) graph for a pn junction.

In an equilibrium semiconductor the diffusion current for electrons or holes at the space charge edge is equal and opposite to the drift current due to the built-in electric field. In other words the electric field creates an energy barrier for electrons to flow from *n*- to *p*-side (Figure 1b). *For forward bias however the energy barrier decreases due to a smaller net electric field, so that electrons and holes can be injected across the space charge region*. Clearly the larger the applied voltage the greater the current generated. *For reverse bias however the net electric field and energy barrier increases, so that current flow is effectively blocked*. This phenomenon of current flow for only one bias polarity is called *rectification* (Figure 3e).

Examples of semiconductor devices: solar cells and light emitting diodes (LED)

In solar cells light is converted to electricity. When photons with energy greater than the semiconductor band gap are absorbed electron-hole pairs are generated by promoting valence band electrons into the conduction band. *The electrons and holes diffuse randomly through the material and therefore there is no net electric current. The built-in electric field of the pn-junction is required in order to generate an electric current*. Provided the photon-generated electrons from the p-layer diffuse to the space charge region they can be injected into the n-layer via the electric field; similarly holes from the n-layer can be injected into the p-layer (Figure 4a). The electrons and holes extracted by the electric field flow through the external circuit as an electric current.

In LEDs electricity is passed through the device to generate light (reverse process of a solar cell). The pn junction is operated under forward bias. *Since the built-in potential barrier is*

lowered electrons and holes are injected across the space charge region to produce excess minority carriers in the quasi-neutral regions (Figure 4b). For example, holes injected from the *p*-side across the space charge region become minority carriers when they enter the *n*-side, where electrons are the majority carriers. *When these excess minority carriers recombine with majority carriers energy is released as light*. The photon energy and wavelength is determined by the band gap of the semiconductor.

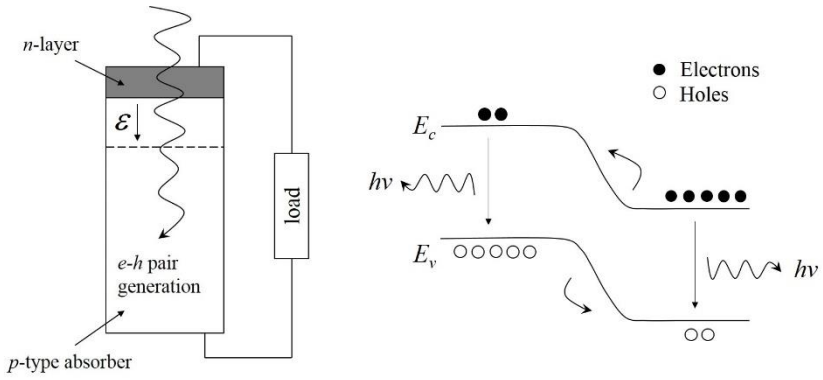


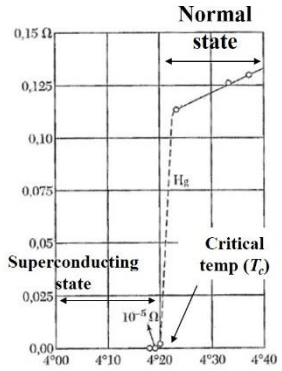
Fig 4: (a) schematic of a solar cell device. (b) shows injection of electrons and holes across the space charge region under forward bias. The excess minority carriers in the quasi-neutral region recombine with the majority carriers to emit light.

FoP3B Part II Lecture 7: Introduction to Superconductors

Superconductivity is the phenomenon of *zero electrical resistivity* below a **critical temperature** T_c. It was first discovered by Kammerling Onnes in 1911 when carrying out cryogenic experiments on the resistivity of mercury. In metals the resistivity (ρ)-temperature (T) dependence has the form:

$$\rho(T) = \rho_o + aT^2 + bT^5$$

where a and b are constants. The individual terms are due to (i) scattering of conduction electrons by impurities (ρ_o), (ii) electron-electron scattering (aT^2) and (iii) electron-phonon scattering (bT^5). Provided the material is pure the resistivity should therefore approach zero *gradually* as the temperature is lowered.



Onnes' result for mercury is shown in Figure 1. Zero resistivity is obtained *suddenly* and *above* absolute zero temperature. This implies that a new **phase** of matter is obtained below the transition (or critical) temperature T_c; this is the **superconducting state**. Above T_c the material is said to be in the **normal state**.

Figure 1: Resistivity as a function of temperature (kelvin) for mercury. The critical temperature (T_c), superconducting and normal states are indicated.

Besides zero resistivity there are several other characteristics common to superconductors. The first is the generation of **persistent currents** and the second is the **Meissner effect**. These will be described below. Finally, *magnetic fields also have an important effect on the stability of the superconducting state and this leads to a classification of superconductors according to* **Type I** *and* **Type II**, which is also discussed.

Persistent Currents

According to Ohm's law $\mathbf{j} = \sigma\boldsymbol{\varepsilon}$, where $\sigma = 1/\rho$ is the conductivity, *the electric field* $\boldsymbol{\varepsilon}$ *within a zero resistivity material must be zero* in order to maintain a constant current density. Hence from **Faraday's law**, i.e. $\vec{\nabla} \times \boldsymbol{\varepsilon} = -\partial \mathbf{B} / \partial t = 0$, i.e. the *magnetic induction field* **B** *is time invariant*.

Consider the following experiment: a magnetic induction field **B** is passed through a ring which is held above T_c (Figure 2a). Next cool the material below T_c in the presence of the **B**-field so that the material transitions from normal to superconducting state. The external **B**-field is then switched off. However, because the **B**-field is time invariant the superconductor must generate an electric current such that it produces an identical magnetic field (Figure 2b)[1]. The sign of the current is determined by **Lenz's law**. Furthermore, *because the resistivity is zero there will be no damping of the current with time*. It is therefore called a **persistent current**.

[1] As will be shown later, the Meissner effect indicates that the **B**-field cannot penetrate the superconductor. Therefore, strictly speaking a current is generated in order to preserve the magnetic flux through the loop (see Supplementary notes).

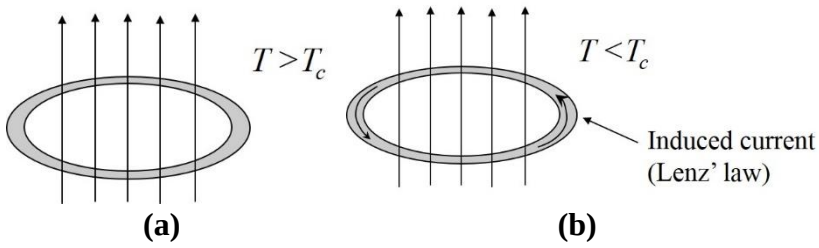


Figure 2: (a) **B**-field passing through a ring in the normal state ($T > T_c$). In (b) the material is cooled below T_c and the external **B**-field is switched off. The superconducting ring generates a persistent current to maintain a constant **B**-field.

It could be questioned whether the resistivity of the superconducting state is really zero or alternatively a small, but non-zero value. Persistent currents provide a more accurate method for measuring resistivity compared to the standard four point probe technique. An electric current decays exponentially with the scattering time τ, which is a measure of the resistivity (in fact $\rho \propto 1/\tau$). With superconductors persistent currents have been monitored over periods of years with no noticeable decay. This places an *upper limit* of 10^{-25} Ωm for the resistivity of a superconductor; by comparison the resistivity of a metal such as copper is significantly higher at 10^{-8} Ωm. Hence the best experimental measurements suggest that the resistivity of a superconductor is in indeed zero for all practical purposes.

Meissner Effect

A further property of superconductors is the expulsion of a magnetic field from within the material. This is known as **diamagnetism** and is the origin of the **Meissner effect**, where a magnet can be levitated above a superconductor. In fact, *superconductors are perfect diamagnets*, i.e. the **B**-field is completely excluded from the material. This means that **ferromagnetic** materials (e.g. Fe, Co, Ni) are typically not superconductors at low temperatures, due to the large internal fields within magnetic domains. Since $\mathbf{B} = \mu_o(\mathbf{M} + \mathbf{H}) = 0$, the **magnetic susceptibility** of a superconductor $\chi = M/H = -1$ (here **M** is the magnetisation, **H** is the magnetic field and μ_o the permeability of free space).

Type I vs Type II behaviour

An increase in temperature above a critical value destroys the superconducting state (Figure 1). A similar trend is also observed with magnetic fields, i.e. *the superconducting state breaks down in strong magnetic fields*. The transition from superconducting to normal state under increasing magnetic field has two forms. *In* **Type I** *behaviour the transition happens suddenly at a critical field*. This is illustrated in the magnetisation curve of Figure 3a. At low magnetic fields the material shows diamagnetic behaviour characteristic of a superconductor. However, at a critical field H_c the magnetic susceptibility, given by the gradient of the graph, abruptly changes to the small, but positive value of the **paramagnetic** normal state. *A* **phase diagram** *can be constructed showing the stability regions of the superconducting and normal states w.r.t temperature and magnetic field* (Figure 3b). For a Type I superconductor the critical **B**-field at temperature T, $B_c(T)$, has the following empirical relationship:

$$B_c(T) = B_c(0)\left[1 - \left(\frac{T}{T_c}\right)^2\right] \quad \ldots (1)$$

where $B_c(0)$ is the critical field at absolute zero. *The critical magnetic field also has implications for the maximum current that can pass through a superconducting wire*. This follows from **Ampere's law**, which states that the (radial) magnetic field in the vicinity of a current carrying wire of radius R is given by:

$$B = \frac{\mu_o I}{2\pi R} \qquad \ldots (2)$$

The maximum current I that can be passed through a superconducting wire is that which generates a magnetic field equal to $B_c(T)$.

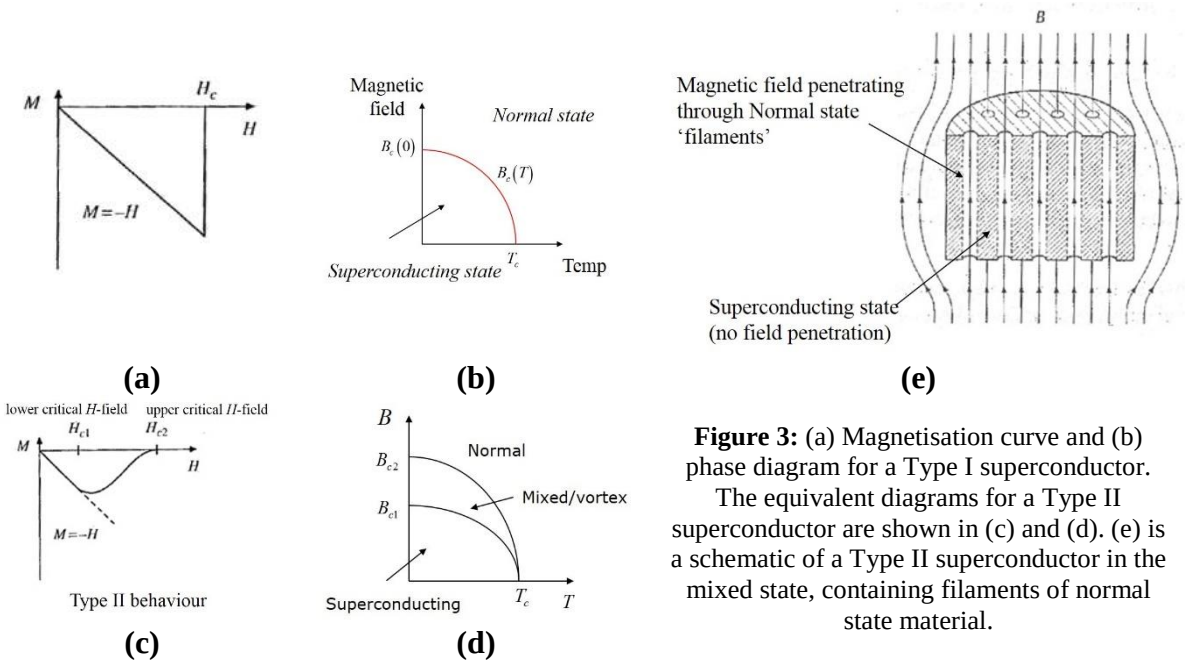


Figure 3: (a) Magnetisation curve and (b) phase diagram for a Type I superconductor. The equivalent diagrams for a Type II superconductor are shown in (c) and (d). (e) is a schematic of a Type II superconductor in the mixed state, containing filaments of normal state material.

A second type of superconducting behaviour, called **Type II** behaviour, is shown in Figure 3c. *Here the transition from superconducting to normal state is gradual, starting at a lower critical field H_{c1} and ending at an upper critical field H_{c2}*. The corresponding phase diagram is shown in Figure 3d. *For magnetic fields between H_{c1} and H_{c2} the material is in a* **mixed** *or* **vortex state**, *where both superconducting and normal regions of material exist side by side*. The normal state exists as 'filaments' extending through the material in the direction of the magnetic field (Figure 3e). *The magnetic field penetrates the normal state filaments, but not the surrounding diamagnetic superconducting regions. As the magnetic field is increased from H_{c1} to H_{c2} the density of normal state filaments increases, until finally above H_{c2} the material has completely transformed into the normal state*.

FoP3B Part II Lecture 8: London equation and thermodynamics of superconductors

As seen in the previous lecture the **Meissner effect** is a unique property of superconductors and is due to *perfect* **diamagnetism**, i.e. the magnetic field is completely excluded from within the material. This is achieved by generating *surface electric currents* (also called **supercurrents**) that opposes the external field. The supercurrent is determined by the **London equation** which is described below. *The superconductor effectively has to do work to exclude the external magnetic field and this has implications for its stability*, as seen previously from the discussion on **Type I** and **Type II** superconducting behaviour. In the last part of this lecture we will calculate the thermodynamic stability of superconductors under magnetic fields.

London equation and the London penetration depth

Since the supercurrent excludes the magnetic **B**-field the two must be directly related. In the **London equation** the current density **j** is directly proportional to the *magnetic vector potential* **A** (recall $\mathbf{B} = \vec{\nabla} \times \mathbf{A}$)[1]:

$$\mathbf{j} = -\frac{1}{\mu_o \lambda_L^2}\mathbf{A} \qquad \ldots (1)$$

where μ_o is the permeability of free space and λ_L is the so-called **London penetration depth**. In defining **A** we need to specify a **gauge**, since otherwise the gradient of any scalar field can be added to **A** leaving the **B**-field unchanged (i.e. **A** is not unique without a gauge). We use the **London gauge**, where $\vec{\nabla} \cdot \mathbf{A} = 0$. This follows from the **continuity equation**, which relates the time evolution of charge carriers (i.e. superconducting electrons n_s) to the gradient in current density, i.e.

$$\frac{\partial n_s}{\partial t} = \frac{1}{e}\vec{\nabla} \cdot \mathbf{j} \qquad \ldots (2)$$

where e is the electronic charge. Since we are interested in **steady state** conditions n_s must be independent of time; Equation (2) is therefore zero and the London gauge naturally follows from Equation (1).

Let us consider the implications of Equation (1). From **Ampere's law** (time independent form of Maxwell's fourth equation):

$$\vec{\nabla} \times \mathbf{B} = \mu_o \mathbf{j} \qquad \ldots (3)$$

From Equation (1):

$$\vec{\nabla} \times \mathbf{j} = -\frac{1}{\mu_o \lambda_L^2}\mathbf{B}$$

[1] Derivations of the London equation are given in the Supplementary notes (non-examinable)

… (4)

Taking the curl of (3) and substituting (4) gives:

$$\vec{\nabla} \times \left(\vec{\nabla} \times \mathbf{B}\right) = -\frac{1}{\lambda_L^2}\mathbf{B}$$

… (5)

Now $\vec{\nabla} \times \left(\vec{\nabla} \times \mathbf{B}\right) = \vec{\nabla}\left(\vec{\nabla} \cdot \mathbf{B}\right) - \nabla^2\mathbf{B}$. From Maxwell's second equation $\vec{\nabla} \cdot \mathbf{B} = 0$, so that:

$$\nabla^2\mathbf{B} = \frac{1}{\lambda_L^2}\mathbf{B}$$

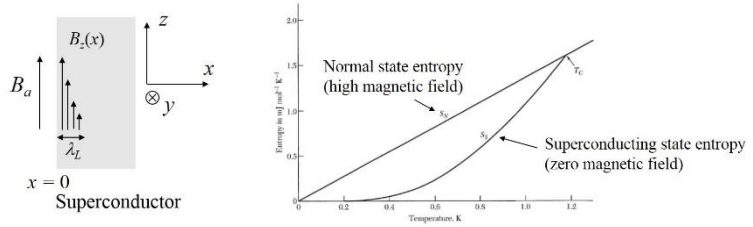
… (6)

The above equation predicts that the **B**-field decays rapidly inside a superconductor. To see this consider a relatively simple case of a **B**-field B_a applied parallel to the *z*-axis of a superconducting slab (Figure 1a). The **B**-field within the superconductor only varies in the *x*-direction and therefore Equation (6) reduces to:

$$\frac{d^2B_z}{dx^2} = \frac{1}{\lambda_L^2}B_z$$

… (7)

where $B_z(x)$ is the *z*-component of the magnetic field at position *x* within the superconductor. It has solutions of the form $\exp(x/\lambda_L)$ and $\exp(-x/\lambda_L)$; the former can be ignored since it is not consistent with expulsion of magnetic fields from within the material. Hence using the **boundary condition** $B_z(x = 0) = B_a$, we have $B_z(x) = B_a\exp(-x/\lambda_L)$, i.e. *the magnetic field decreases exponentially within the material over a characteristic length* λ_L. Knowing the magnetic field the supercurrent **j** can be calculated using Equation (3). The supercurrent $j(x)$ flows in the *y*-direction and also has an exponential decrease over λ_L length scale, i.e. *the supercurrent is confined to the outer surface of the material*. Note also that the direction of supercurrent flow is such that it opposes the applied **B**-field within the material.



(a) **(b)**

Fig. 1: (a) Schematic of geometry used to calculate field penetration within a superconductor. (b) Entropy as a function of temperature for the superconducting and normal states. T_c is the critical phase transition temperature.

It can be shown[2] that the London penetration depth λ_L is given by:

[2] See derivations of the London equations in Supplementary reading.

$$\lambda_L = \sqrt{\frac{m}{\mu_0 n_s e^2}}$$

… (8)

where *m* is the electron mass. It will be shown in the next lecture that *superconducting electrons form in pairs* (known as **Cooper pairs**); hence n_s is the number of superconducting electrons (= twice the number of Cooper pairs). *The fact that electrons are paired in a superconductor means that the entropy is lower compared to the normal state*. This is illustrated in Figure 1b which plots the entropy of the material in the superconducting and normal states below the critical temperature T_c. Note that the superconductor is the stable phase at temperatures below T_c, but the normal state can be produced by applying a strong enough magnetic field, thus enabling measurement of its entropy. *From Figure 1b the entropy decrease for the superconductor w.r.t the normal state becomes larger at lower temperatures, implying a greater number of Cooper pairs n_s with decreasing temperature*. There are two important implications that follow:

(i) The electrons can be divided into zero resistivity superconducting electrons (i.e. Cooper pairs) and 'normal' electrons with non-zero resistivity. The fraction of superconducting electrons increases as the material is cooled below T_c. The overall resistivity of the material is nevertheless zero at all temperatures below T_c since the electric current is carried exclusively by the zero resistivity superconducting electrons.

(ii) There is no **latent heat** or entropy change at the critical temperature T_c when the normal state transitions to the superconducting state on cooling. This is known as a **second order phase transition**. As a comparison melting of ice or boiling of water is an example of **first order transition**, which is characterised by a latent heat and entropy change due to breaking of chemical bonds.

Thermodynamics of superconductors

Consider the free energy of a superconductor under an applied magnetic field. Let $G_s[0]$ be the free energy of the superconductor in zero field conditions. *The work* (*W*) *done on a material by a magnetic field is given by*:

$$dW = -\mathbf{M}(\mathbf{B}) \cdot d\mathbf{B}$$

… (9)

where **M** is the magnetisation, i.e. net magnetic dipole moment per unit volume. *Equation (9) follows from the result that the potential energy of a single magnetic dipole moment* $\boldsymbol{\mu}$ *in a* **B***-field is* $-\boldsymbol{\mu}\cdot\mathbf{B}$. The superconductor free energy $G_s[B]$ under an applied field is then:

$$G_s[B] = G_s[0] - \int_0^B \mathbf{M}(\mathbf{B}) \cdot d\mathbf{B}$$

… (10)

Since the superconductor is a perfect diamagnet $M = -H = -B/\mu_o$, so that:

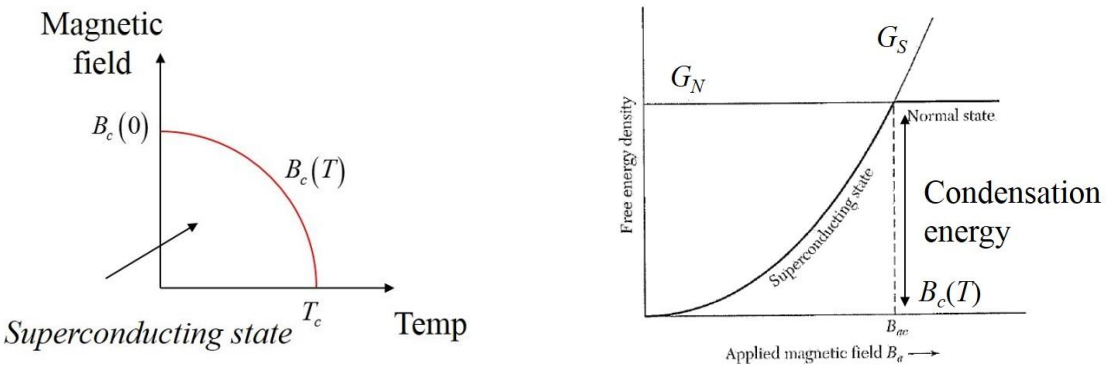
$$G_s[B] = G_s[0] + \frac{B^2}{2\mu_o} \quad \ldots (11)$$

Taking Type I superconductors as an example (Figure 2a) from Equation (11) the superconductor free energy will keep increasing with magnetic field until the critical field $B_c(T)$, beyond which the stable phase is the normal state. <u>*At $B_c(T)$ the free energy of superconducting and normal states are equal*</u>. Denote by $G_N[0]$ the free energy of the normal state in zero field conditions. *Since the normal state is* **paramagnetic** *its magnetisation* **M** *will be small and therefore the free energy will not change significantly with magnetic field*. Therefore $G_N[B] \approx G_N[0]$, so that:

$$G_s[B_c(T)] = G_s[0] + \frac{B_c(T)^2}{2\mu_o} = G_N[B_c(T)] = G_N[0]$$

$$\Rightarrow G_N[0] - G_s[0] = \frac{B_c(T)^2}{2\mu_o} \quad \ldots (12)$$

($G_N[0]$- $G_S[0]$) is called the **condensation energy** and represents the *free energy difference between normal and superconducting states under zero field conditions*. This is illustrated schematically in Figure 2b which plots the free energy as a function of magnetic field for the superconducting and normal states below T_c. Note the normal state energy is approximately a horizontal line, due to the paramagnetic behaviour; the superconductor energy however shows a quadratic dependence, as predicted by Equation (11). *The condensation energy of a superconductor is only μeV/atom; as a comparison the latent heat of fusion for ice melting is 63 meV/atom.*



(a) **(b)**

Fig. 2: (a) Phase diagram for a Type I superconductor and (b) Free energy vs magnetic field for the normal and superconducting states below T_c. The condensation energy is highlighted.

FoP3B Part II Lecture 9: Ginzburg-Landau theory of superconductivity

There are two alternative, but equivalent, theories for superconductivity: (i) BCS theory after Bardeen, Cooper and Schrieffer, and (ii) Ginzburg-Landau (GL). BCS is a *microscopic* theory, i.e. it describes superconductivity at the fundamental level of electrons in the solid. GL is a *phenomenological* theory, i.e. superconductivity is modelled at a more macroscopic level using so-called *order parameters*. In this lecture we will primarily discuss GL theory and rely on BCS to interpret the physical meaning of the order parameters.

Summary of BCS theory

Consider electrons moving in a crystal. In normal metals the direction of movement is random in the absence of an electric field. In superconductivity however a negatively charged electron will distort the ion lattice towards it (Figure 1a), so that a neighbouring electron will 'see' a lower potential region of material around the first electron and be attracted towards it. This gives rise to **Cooper pairs** of electrons that move as a single unit. At a more detailed level Cooper pairs form from **electron-phonon** interactions. *The spacing between the Cooper pair electrons can be as large as several 100 nm and the interaction is therefore weak, which is why superconductivity is a low temperature phenomenon. Furthermore, not all electrons form Cooper pairs*; in fact we have already seen (previous lecture) that the fraction of Cooper pairs increase as the material is cooled below T_c. At any given temperature the majority of electrons behave in a conventional manner, i.e. they are unbound, have finite resistivity and keep the solid from falling apart. However, some electrons close to the Fermi level are bound as Cooper pairs giving rise to superconducting behaviour.

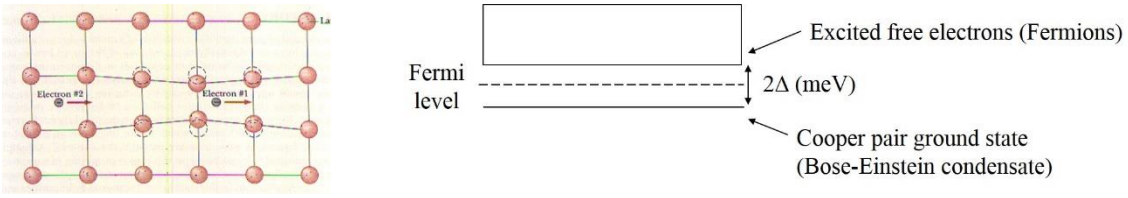


Fig. 1: (a) Formation of a Cooper pair and (b) electronic energy level diagram in a superconductor

The question remains why Cooper pairs have zero resistivity. Normal electrons are fermions, but with two electrons in a Cooper pair the spins take on integer values (0 or 1). *Cooper pairs are therefore bosons. BCS theory shows that Cooper pairs form a Bose-Einstein condensate, with a small energy gap (meV) to the next (fermion) electron level which is unoccupied (Figure 1b). The energy gap suppresses scattering processes that lead to resistivity, since scattering involves promoting an electron to a higher energy, unoccupied level.*

Ginzburg-Landau (GL) theory for homogeneous systems

We will now present GL theory for the relatively simple case of a homogeneous (i.e. uniform) material system under no applied magnetic field. Close to the **critical temperature** T_c the free energy is expanded as a power series:

$$G_S(T) = G_N(T) + a(T)|\psi|^2 + \frac{b(T)}{2}|\psi|^4 \quad \ldots (1)$$

where G_S, G_N are free energies of the superconducting and normal states, and $a(T)$, $b(T)$ are temperature dependent expansion coefficients. ψ is a complex number and is the **order parameter** in GL theory. The free energy depends on $|\psi|^2$. From BCS it can be shown that $|\psi|^2$ is the density of Cooper pairs, i.e. $|\psi|^2 = n_s/2$, where n_s is the density of superconducting electrons. For the superconducting state to be stable the G_s vs $|\psi|^2$ graph must have a minimum. Since Equation (1) is a quadratic function w.r.t $|\psi|^2$ a minimum is obtained for $b(T) > 0$.

The ground state order parameter is determined by:

$$\frac{\partial G_s}{\partial|\psi|} = 2|\psi|\{a(T) + b(T)|\psi|^2\} = 0$$
$$\Rightarrow |\psi| = 0 \text{ or } |\psi|^2 = -a(T)/b(T) \quad \ldots (2)$$

For the superconducting state only the second solution for $|\psi|^2$ is valid (i.e. $|\psi| = 0$ is the solution for the normal state). For Equation (2) to give a positive $|\psi|^2$ value for the superconducting state $a(T) < 0$ for $T < T_c$ (recall $b(T) > 0$). Above T_c, $a(T) > 0$, so that the superconductor is not stable.

Consider now the *Taylor expansion* of a function $f(x)$ about the point $x = x_o$:

$$f(x) = f(x_o) + f'(x_o)(x - x_o) + \frac{f''(x_o)}{2!}(x - x_o)^2 + \cdots \quad \ldots (3)$$

Instead of x, x_o we can substitute T, T_c and instead of $f(x)$ we can write either $a(T)$ or $b(T)$. For $a(T)$ we set the first term on the RHS equal to zero and retain only the second term. Therefore $a(T) \approx \dot{a}(T - T_c)$, where $\dot{a}$ is a positive constant. This satisfies the desired properties for positive and negative $a(T)$ values above and below T_c. Similarly for $b(T)$ only the first term on the RHS of Equation (3) is retained, so that $b(T) \approx b$, where b is a positive constant. Therefore:

$$|\psi|^2 = \begin{cases} \left[\frac{\dot{a}(T_c - T)}{b}\right] & T < T_c \\ 0 & T > T_c \end{cases} \quad \ldots (4)$$

This is illustrated in Figure 2. The minimum in the graph is the Cooper pair density for the superconducting state. From Equation (4) the Cooper pair density increases with cooling below T_c, as required. The **condensation energy**, $G_N - G_S$, is found by substituting the $|\psi|^2$ value for $T < T_c$ from (4) into Equation (1). The result is $[\dot{a}(T - T_c)]^2/2b$. From thermodynamics the condensation energy is $[B_c(T)]^2/2\mu_o$, where $B_c(T)$ is the critical magnetic field (see previous lecture). Equating the GL and thermodynamic expressions for the condensation energy gives a value for $\dot{a}^2/b$ in terms of measurable parameters such as $B_c(T)$ and T_c. This value for $\dot{a}^2/b$ can

then be used to predict further properties, such as entropy using GL theory. The results are consistent with the experimental observation of superconductivity being a second order phase transition, with no latent heat or entropy change at the critical temperature T_c.

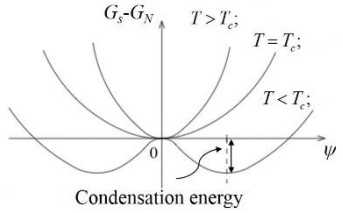


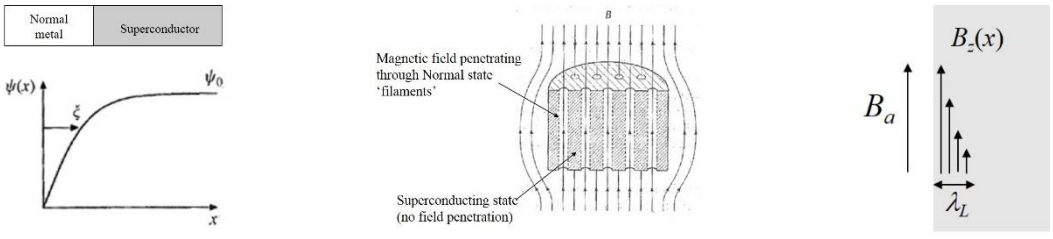
Fig. 2: Free energy ($G_S - G_N$) vs $|\psi|$ graphs as a function of temperature. The minimum in the curve gives the order parameter $|\psi|^2$ for the material. At $T>T_c$, $|\psi|^2 = 0$ and $(G_S - G_N) = 0$, meaning that only the normal state is allowed. For $T<T_c$ however $|\psi|^2 >0$ and $(G_S - G_N) < 0$, so that the superconductor is the more stable phase. The condensation energy is also indicated.

Coherence length and Type I vs Type II behaviour

GL theory can be extended to inhomogeneous systems as well, such as the interface between a normal metal and superconductor (Figure 3a). *The order parameter ψ for this system increases from zero in the normal metal to the equilibrium value ψ_0 for the superconductor over a characteristic distance ξ, which is known as the* **coherence length**. $|\psi|^2$ represents the density of Cooper pairs, so the fact that the order parameter is lower near the normal metal-superconductor interface means that some of the Cooper pairs are destroyed. This is due to the fact that the two electrons forming the Cooper pair have a relatively large separation (despite being bound), so that in order not to be broken up by the normal metal they have to be sufficiently far away from the interface. *Thus the coherence length ξ is of the order of the Cooper pair separation.* From GL theory the coherence length is given by:

$$\xi(T) = \left(\frac{\hbar^2}{2m|a(T)|}\right)^{1/2} = \left[\frac{\hbar^2}{2m\dot{a}(T_c - T)}\right]^{1/2} \quad \ldots (5)$$

The coherence length therefore decreases as the material is cooled below T_c. The fact that the order parameter and Cooper pair density is lower near the interface also means that the *local* condensation energy for that region is smaller than the bulk superconductor. This has important implications for Type I vs Type II superconducting behaviour. As shown in Figure 3b the characteristic feature of Type II superconductors is the **mixed or vortex state**, where normal and superconducting regions exist side by side, similar to Figure 3a.



(a) **(b)** **(c)**

Fig. 3: (a) Order parameter as a function of distance from a normal metal-superconductor interface. (b) Microstructure of the mixed or vortex state in a Type II superconductor. (c) Magnetic field penetration into a superconductor.

Consider now a normal metal-superconductor interface under an applied magnetic field. *With no magnetic field the free energy of the superconductor will decrease to its bulk value over the coherence length ξ, due to condensation of Cooper pairs*. If a magnetic field B_a is then applied, the field will penetrate into the normal metal, but will be repelled from within the **diamagnetic** superconductor over a characteristic length λ_L, the **London penetration depth** (Figure 3c). *Repelling the magnetic field will add an extra energy term to the superconductor, which varies as $B^2/2\mu_o$, where B is the magnitude of the field repelled*. These two contributions, condensation energy and magnetic energy are shown schematically in Figure 4a as a function of distance from the normal metal-superconductor interface. The diagram for Figure 4a has $\xi < \lambda_L$; the condensation energy decreases rapidly at the interface, while the magnetic field energy increases relatively slowly. The net result is a *relatively low* interfacial energy. Therefore a $\xi < \lambda_L$ interface is energetically allowed and leads to Type II behaviour (Figure 3b).

On the other hand Figure 4b shows the opposite case where $\xi > \lambda_L$. Here the condensation energy decreases relatively slowly at the interface, while the magnetic energy increases rapidly. The net result is a *relatively high* interfacial energy. The interface is therefore energetically unfavourable and cannot exist. This leads to Type I behaviour with no mixed phase. Type I vs Type II behaviour is therefore governed by the relative magnitudes of ξ and λ_L.

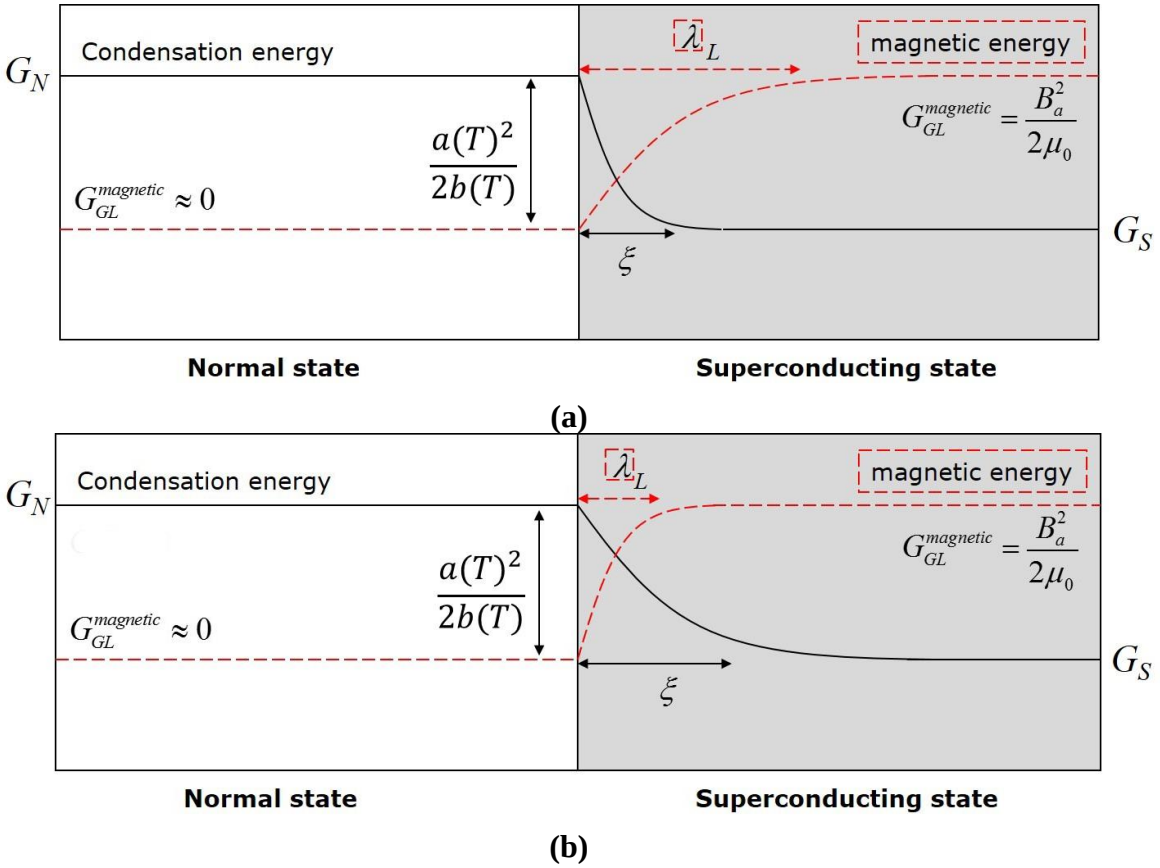


Fig. 4: Condensation and magnetic free energy contributions for (a) $\xi < \lambda_L$ Type II superconductor and (b) $\xi > \lambda_L$ Type I superconductor.

FoP3B Part II Lecture 10: Polarisation in Dielectrics

Dielectrics *are insulating materials that can be* **polarised** *by an applied electric field*. To illustrate this consider a capacitor attached to a battery. The **capacitance** (C) is the charge (Q) stored per unit voltage (V) or:

$$C = \frac{Q}{V} = \epsilon_o \epsilon_r \frac{A}{d} \qquad \ldots (1)$$

where ϵ_o is the permittivity of free space, A is the capacitor plate area and d is the plate spacing. ϵ_r is the **dielectric constant** or **relative permittivity** (a dimensionless value). In some materials, such as $BaTiO_3$ (a widely used dielectric), ϵ_r is extremely large. Consider inserting $BaTiO_3$ in between the capacitor plates. From Equation (1) the capacitance increases significantly; this could be due to either Q increasing or V decreasing or both. Since the $BaTiO_3$ is charge neutral (i.e. no additional charge added to the capacitor) and there is no loss of charge due to $BaTiO_3$ being insulating Q must be constant. Hence *increased capacitance must be due to a decrease in voltage across the capacitor*. *This implies that the dielectric generates an opposing electric field to the battery*. The dielectric is electrically polarised under the applied electric field. In this lecture we explore the microscopic origins of polarisation and discuss the concept of **microscopic** and **macroscopic** electric fields in a dielectric.

Polarisation and macroscopic electric field

The microscopic origin of polarisation is the **electric dipole moment**, which is due to the spatial separation of positive and negative charge (Figure 1a). The dipole moment vector $\boldsymbol{\mu}$ extends from the negative to positive charge. The magnitude of $\boldsymbol{\mu}$ is qd, where q is the magnitude of either the positive or negative charge (both are equal) and d is their separation (NB: do not confuse d with the capacitor plate spacing in Equation 1). *There are many ways by which an electric dipole moment can be formed in a dielectric*. One method is the polarisation of the electron cloud in an otherwise neutral atom under an applied electric field (Figure 1b). This is known as **electronic polarisation**. On the other hand in molecules such as water (H_2O) the oxygen is highly *electronegative*, which means it attracts some of the electrons in the bond towards it. The oxygen becomes negatively charged and the hydrogen positively charged, resulting in a permanent electric dipole, which is present independent of an electric field (Figure 1c).

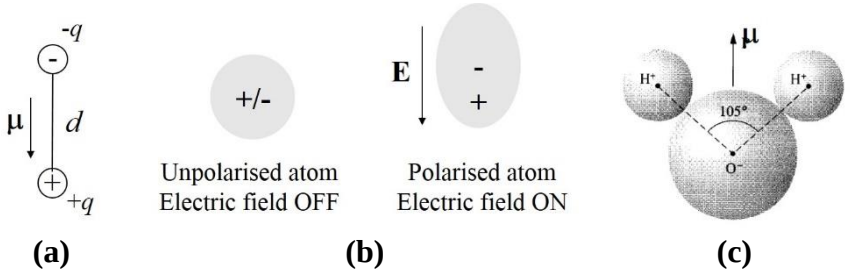


Fig. 1: (a) An electric dipole moment $\boldsymbol{\mu}$, (b) electronic polarisation of the electron cloud of a neutral atom under an electric field **E** and (c) the permanent dipole in a H_2O molecule

Consider first a solid with permanent electric dipoles. We define the polarisation **P** as the *dipole moment per unit volume*. With no electric field present the permanent electric dipole moments

are all randomly oriented (Figure 2a), so that **P** is negligible. *When an external electric field* $\mathbf{E}_{ext}$ *is applied there is a torque* ($\boldsymbol{\mu}$ x $\mathbf{E}_{ext}$) *acting on the electric dipole, which rotates it until the potential energy* $-\boldsymbol{\mu}\cdot\mathbf{E}_{ext}$ *is minimised*. The minimum energy is when the dipole moment $\boldsymbol{\mu}$ is parallel to the electric field (Figure 2b). In this particular example we considered a solid with permanent dipoles, but it should be clear that an electric field will generate a non-zero polarisation **P** even in the absence of permanent dipoles, due to electronic polarisation.

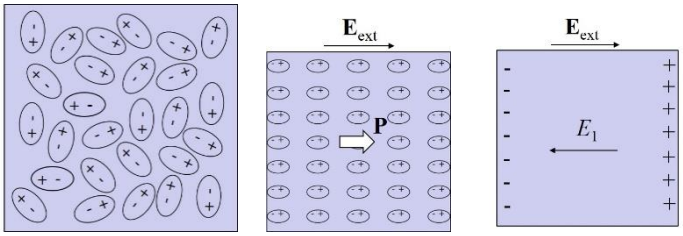


Fig 2: dipole moments (a) before and (b) after applying an external electric field $\mathbf{E}_{ext}$ and (c) the continuum equivalent of (b) with surface charge and internal electric field E_1.

Now each dipole moment $\boldsymbol{\mu}$ will generate its own electric field according to:

$$\mathbf{E}(\mathbf{r}) = \frac{3(\boldsymbol{\mu}\cdot\mathbf{r})\mathbf{r} - r^2\boldsymbol{\mu}}{4\pi\epsilon_o r^5} \qquad \ldots (2)$$

where **r** is the position vector (in the above equation the dipole is at the origin). The random dipole orientations in Figure 2a will generate a zero net dipole electric field, but not so for the non-zero polarisation **P** configuration in Figure 2b. *From a theorem in electrostatics the collective effect of dipoles in a material of uniform polarisation* **P** *can be modelled as a surface charge density* $\sigma = \mathbf{P}\cdot\mathbf{n}$, *where* **n** *is the surface unit normal of the solid*. The oriented dipoles in Figure 2b is then equivalent to the continuum slab of material in Figure 2c with surface charge as indicated. The surface charge gives rise to an internal electric field E_1 which is opposite to the external applied field. By **Gauss' law** $E_1 = -\sigma/\epsilon_o$. The **macroscopic field** within the material is therefore ($E_{ext} - \sigma/\epsilon_o$), i.e. the electric field is reduced due to the polarisation and the capacitance increases.

Microscopic electric fields

The macroscopic field is the electric field accessible to experimental measurement. It is a smooth, continuous field. Real materials however consist of discrete dipoles and are therefore not smooth and continuous at the atomic length scale. The *local* **microscopic** electric field $\mathbf{E}_{local}$ at an electric dipole is therefore different from the macroscopic field. Although not directly accessible by measurement the microscopic field is nevertheless important since it determines the dipole moment $\boldsymbol{\mu}$ via the **polarisability** α, i.e.:

$$\boldsymbol{\mu} = \alpha\mathbf{E}_{\text{local}} \qquad \ldots (3)$$

In this section the procedure for calculating $\mathbf{E}_{local}$ from the macroscopic field will be presented. Consider calculating the local electric field at the centre of the material shown in Figure 2b.

The material is divided into two regions (Figure 3): the first is a sphere about the origin. Since this is close to the centre the electric fields due to individual dipole moments within this region are discretely summed, i.e. the material is analysed at the atomistic level. The region outside the sphere is considered to be sufficiently far away from the centre that macroscopic fields can be applied. i.e. this region is treated as a continuum. Since the polarisation within this outer region is uniform we can model it via surface charges given by $\sigma = \mathbf{P}\cdot\mathbf{n}$. This results in surface charges at the outer edges of the solid, as well as within the cavity walls (Figure 3). *The outer region is therefore a superposition of two electric fields*: (i) the field $\mathbf{E}_1 = -\sigma/\epsilon_o$ due to a slab with no cavity (see previous section) and (ii) a field $\mathbf{E}_2$ due a cavity with surface charges. Note that the surface charges within the cavity walls are such that $\mathbf{E}_2$ is opposite to $\mathbf{E}_1$. $\mathbf{E}_2$ is known as the **Lorentz field** and its value can be calculated using standard electrostatic methods:

$$\mathbf{E}_2 = \frac{\mathbf{P}}{3\epsilon_o} \qquad \ldots (4)$$

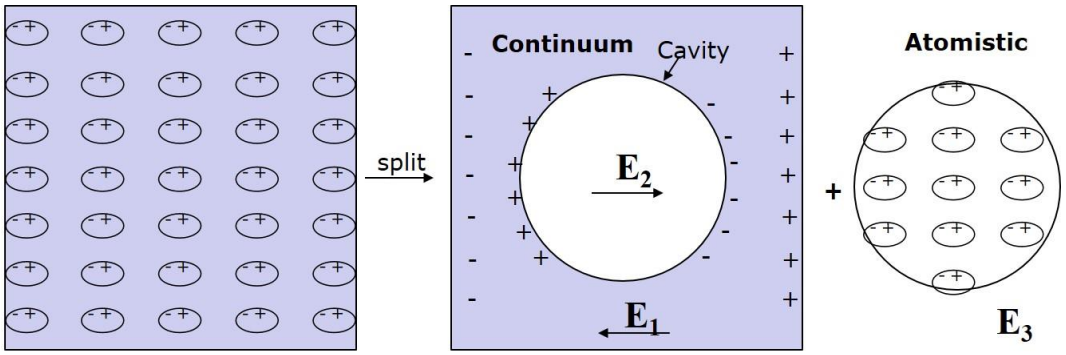


Fig 3: Breakdown of solid into continuum and atomistic region for calculating the local electric field at the centre

Next consider the atomistic electric field $\mathbf{E}_3$ at the origin within the sphere. We need to sum the dipole electric field in Equation (2) for all dipoles within the sphere. In Figure 3 the dipole moments are along the *x*-axis and therefore $\boldsymbol{\mu} = (\mu,0,0)$. Consider $\mathbf{E}_{3,x}$ which is the *x*-component of the electric field $\mathbf{E}_3$. From Equation (2) we have:

$$\mathbf{E}_{3,x}(0) = \sum_i \frac{3\mu x_i^2 - r_i^2\mu}{4\pi\epsilon_o r^5} = \mu \sum_i \frac{3x_i^2 - r_i^2}{4\pi\epsilon_o r^5} \qquad \ldots (5)$$

where the summation is over all dipoles *i* located within the sphere and at position vector $\mathbf{r}_i = (x_i, y_i, z_i)$. Now assume that the solid has a *cubic* arrangement of dipoles so that the *x*,*y* and *z*-axes are equivalent. Therefore:

$$\sum_i x_i^2 = \sum_i y_i^2 = \sum_i z_i^2 \qquad \ldots (6)$$

Because $r_i^2 = x_i^2 + y_i^2 + z_i^2$ from Equation (6) we then have $\mathbf{E}_{3,x}(0) = 0$. Consider the y-component $\mathbf{E}_{3,y}$. From Equation (2):

$$\mathbf{E}_{3,y}(0) = \sum_i \frac{3\mu x_i y_i}{4\pi\epsilon_o r^5} \quad \ldots (7)$$

The above expression is an odd function w.r.t x_i and y_i. Hence summing over all y_i for fixed x_i would give zero (equivalently we could sum over all x_i keeping y_i fixed). Therefore $\mathbf{E}_{3,y}(0) = 0$ and similarly it can be shown that $\mathbf{E}_{3,z}(0) = 0$. Therefore, $\mathbf{E}_3 = 0$.

The local electric field is given by:

$$\mathbf{E}_{\text{local}} = \mathbf{E}_{\text{ext}} + \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3$$

$$|\mathbf{E}_{\text{local}}| = E_{\text{ext}} - \frac{\sigma}{\epsilon_o} + \frac{P}{3\epsilon_o} \quad \ldots (8)$$

FoP3B Part II Lecture 11: Ferroelectric crystals

Previously we discussed **macroscopic** and **microscopic** electric fields. The latter cannot be measured, but is important since the **electric dipole moment** $\boldsymbol{\mu}$ of an atom or molecule is linked to the *local* microscopic field $\mathbf{E}_{\text{local}}$ via $\boldsymbol{\mu} = \alpha\mathbf{E}_{\text{local}}$, where α is the **polarisability**. $\mathbf{E}_{\text{local}}$ can be calculated from the polarisation **P**. A similar approach is adopted for α, this time using the **dielectric constant** or **relative permittivity** ϵ_r as the macroscopic variable. This gives rise to the **Clausius-Mossotti** relationship.

Dielectrics were introduced as **polarisable** media under static DC electric fields. Here the discussion will be broadened to *oscillating AC electric fields*. ϵ_r is then a function of the AC oscillation frequency ω; we will explore the different contributing factors to $\epsilon_r(\omega)$ and derive its form for the special case of **electronic polarisation**. $\epsilon_r(\omega)$ *is the most general description of the dielectric properties of the material, and the static DC scenario is a special case where* $\omega \rightarrow 0$. Although not covered in this module $\epsilon_r(\omega)$ provides useful information on how a material interacts with light and other electromagnetic waves (recall that EM waves have oscillating electric fields). For example, using $\epsilon_r(\omega)$ it is possible to derive the **refractive index** and **absorption coefficient** of the material.

Finally, **ferroelectric** crystals, such as $BaTiO_3$, will also be introduced. These are similar to **ferromagnets**, except that instead of a spontaneous *magnetisation* **M** we have a spontaneous *polarisation* **P**. Ferroelectrics share many features in common with ferromagnets, such as **hysteresis loops** and **domains**.

Clausius-Mossotti relationship

Since polarisation (**P**) is the net dipole moment per unit volume and $\boldsymbol{\mu} = \alpha\mathbf{E}_{\text{local}}$ we can write:

$$\mathbf{P} = N\boldsymbol{\mu} = N\alpha\mathbf{E}_{\text{local}} \qquad \ldots (1)$$

where N is the number density of dipole moments[1]. From the previous lecture $E_{\text{local}} = E_{\text{macro}} + (P/3\epsilon_o)$, where E_{macro} is the *internal* macroscopic field due to an external applied field E_{ext}, i.e. $E_{\text{macro}} = E_{\text{ext}} - \sigma/\epsilon_o$, with σ being the surface charge density. Substituting the expression for E_{local} in (1) and re-arranging for $N\alpha$:

$$N\alpha = \frac{(P/E_{\text{macro}})}{1 + \frac{1}{3\epsilon_o}(P/E_{\text{macro}})} \qquad \ldots (2)$$

The term (P/E_{macro}) is equal to $\epsilon_o\chi_e$, where χ_e is the **electric susceptibility** (it is similar to the *magnetic susceptibility* M/H). We now derive an alternative expression for (P/E_{macro}) using the definition for the **electric displacement** field **D**:

[1] It is assumed that all dipoles are identical with dipole moment $\boldsymbol{\mu}$. This is however a simplification. Although not discussed the Clausius-Mossotti relationship can be generalised without this simplification.

$$\mathbf{D} = \epsilon_o \mathbf{E} + \mathbf{P} = \epsilon_o \epsilon_r \mathbf{E}$$

… (3)

Substituting E_{macro} for **E** in Equation (3) we get $(P/E_{macro}) = \epsilon_o(\epsilon_r\text{-}1)$. Therefore, Equation (2) simplifies to:

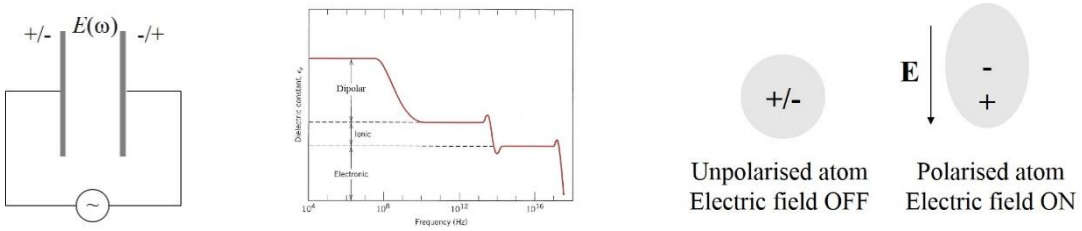
$$\frac{N\alpha}{3\epsilon_o} = \frac{\epsilon_r - 1}{\epsilon_r + 2}$$

… (4)

This is the Clausius-Mossotti relationship that links polarisability α to ϵ_r.

Dielectric properties under an oscillating electric field

Consider connecting a capacitor with a dielectric to an AC current source (Figure 1a). The equation for capacitance (*C*) can be generalised from the static DC case to $C = \epsilon_o\epsilon_r(\omega)A/d$, where ϵ_r is now a function of frequency ω of the AC circuit (*A* is the capacitor plate area and *d* is the plate spacing). The graph of $\epsilon_r(\omega)$ as a function of ω is shown in Figure 1b. Three different regimes can be identified, which are labelled as *electronic*, *ionic* and *dipolar*. **Electronic polarisation** *is the polarising or displacement of the electron cloud centre of mass w.r.t the nucleus of an atom by an electric field. Since the electron mass is relatively small the displacement can happen at high frequencies* (Figure 1b). **Ionic polarisation** *is due to displacement of individual positive and negative ions by the electric field. Since the ions are relatively heavy this mechanism only occurs at intermediate frequencies*. **Dipolar polarisation** *is the rotation of an entire molecule (e.g. water H_2O) with permanent dipole moment by the electric field. This mechanism is only active at low frequencies, due to the fact that entire molecules must be rotated.*



(a) **(b)** **(c)**

Fig 1: (a) capacitor connected to AC current of frequency ω, (b) $\epsilon_r(\omega)$ plot as a function of ω and (c) displacement of the centre of mass of an electron cloud in an atom due to an electric field **E**.

An expression for $\epsilon_r(\omega)$ due to electronic polarisability will now be derived. With no electric field the centre of mass of the negatively charged electron cloud overlaps with the positively charged nucleus and the atom has no dipole moment (Figure 1c). Once an electric field **E** is applied however the electron cloud is easily displaced w.r.t the nucleus to give rise to a dipole moment **μ**, which can be expressed as:

$$\boldsymbol{\mu}(\omega) = -e\mathbf{r}(\omega)$$

… (5)

Here e is the magnitude of electron charge and $\mathbf{r}$ is the position vector of the centre of mass of the electron cloud (the nucleus is at the origin). $\boldsymbol{\mu}$ and $\mathbf{r}$ are functions of ω, since we are interested in an oscillating electric field. Furthermore, in equation (5) a negative sign is included since the direction of $\mathbf{r}$ is from the positively charged nucleus to the electron cloud centre of mass (i.e. opposite to the direction of $\boldsymbol{\mu}$). *The displacement* $\mathbf{r}(\omega)$ *can be modelled by treating the electron-nucleus bond as a spring of spring constant* K. From **simple harmonic motion** the **restoring force** of the spring is $-K\mathbf{r}(\omega)$; NB: the direction of the restoring force is opposite to $\mathbf{r}$. The other force acting on the electron cloud is the force generated by the oscillating electric field $-e\mathbf{E}_{\text{local}}(\omega)$ (NB: for an individual dipole we must use the microscopic, rather than macroscopic, electric field). The equation of motion is therefore:

$$m\frac{d^2\mathbf{r}(\omega)}{dt^2} = -K\mathbf{r}(\omega) - e\mathbf{E}_{\text{local}}(\omega) \quad \ldots (6)$$

Now an oscillating local electric field can be expressed as $\mathbf{E}_{\text{local}}(\omega) = \mathbf{E}_o\exp(i\omega t)$, where t is time and $\mathbf{E}_o$ is the local electric field at $t = 0$. Since the electron cloud can react almost instantaneously to the electric filed we have $\mathbf{r}(\omega) = \mathbf{r}_o\exp(i\omega t)$. Substituting expressions for $\mathbf{E}_{\text{local}}(\omega)$ and $\mathbf{r}(\omega)$ in (6) and simplifying:

$$\mathbf{r}(\omega) = \frac{e\mathbf{E}_{\text{local}}(\omega)}{m(\omega^2 - \omega_o^2)} \quad \ldots (7)$$

where $\omega_o = (K/m)^{1/2}$ is the **resonant frequency** of the simple harmonic oscillator. Substituting in (5) gives:

$$\boldsymbol{\mu}(\omega) = \frac{e^2\mathbf{E}_{\text{local}}(\omega)}{m(\omega_o^2 - \omega^2)} \quad \ldots (8)$$

From the definition of polarisability α we have $\boldsymbol{\mu}(\omega) = \alpha(\omega)\mathbf{E}_{\text{local}}(\omega)$. By comparing with Equation (8) we obtain an expression for $\alpha(\omega)$:

$$\alpha(\omega) = \frac{e^2}{m(\omega_o^2 - \omega^2)} \quad \ldots (9)$$

Substituting in the Clausius-Mossotti relationship (Equation 4) then gives:

$$\epsilon_r(\omega) = 1 + \frac{Ne^2}{m\epsilon_o(\omega_o^2 - \omega^2) - (\frac{Ne^2}{3})} \quad \ldots (10)$$

The shape of $\epsilon_r(\omega)$, as predicted by Equation (10), is plotted schematically in Figure 2. *There is an abrupt change in shape of $\epsilon_r(\omega)$ close to the resonance frequency ω_o*. This is known as the **anomalous dispersion** region, since ϵ_r decreases w.r.t. ω, whereas typically it increases.

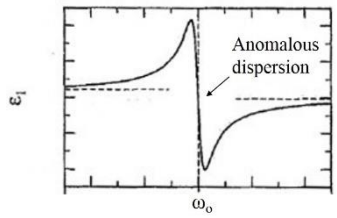
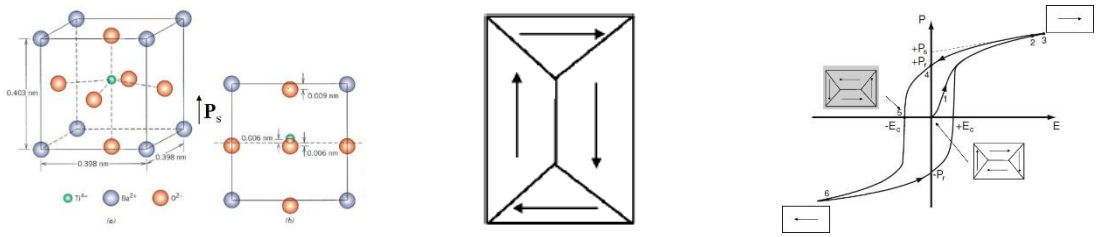


Fig 2: $\epsilon_r(\omega)$ as a function of ω due to electronic polarisability. The anomalous dispersion region occurs close to the resonance frequency ω_o.

Ferroelectric crystals

Ferroelectric crystals are characterised by a large **spontaneous polarisation** ($\mathbf{P}_s$) *that is present even in the absence of an electric field*. In the case of $BaTiO_3$ $\mathbf{P}_s$ is due to a small (<1 Å) displacement of negatively charged oxygen ions w.r.t. positively charged Ba or Ti ions (Figure 3a). Ferroelectric crystals have many similarities with ferromagnets. For example, *the material is organised into* **polarisation domains** (Figure 3b). This is because for a dielectric with no free charge by Maxwell's equation $\vec{\nabla} \cdot \mathbf{D} = \vec{\nabla} \cdot (\epsilon_o \mathbf{E} + \mathbf{P}) = 0$ or $\epsilon_o \vec{\nabla} \cdot \mathbf{E} = -\vec{\nabla} \cdot \mathbf{P}$. If there were no domains $\vec{\nabla} \cdot \mathbf{P} \neq 0$ at a free surface, so that a **depolarising electric field** must be present in order to satisfy Maxwell's equation. This increases the energy of the system (recall that the potential energy of a dipole moment $\boldsymbol{\mu}$ in electric field $\mathbf{E}$ is $-\boldsymbol{\mu}\cdot\mathbf{E}$; hence energy increases when the depolarising field is anti-parallel to $\mathbf{P}_s$). By forming domains the discontinuity in polarisation $\mathbf{P}$ at the free surface is avoided. The presence of domains gives rise to **hysteresis** behaviour in $\mathbf{P}$ vs. $\mathbf{E}$ curves (Figure 3c). The shape of the hysteresis loop is explained by the energy of individual domains in the presence of an applied electric field, i.e. domains where the internal polarisation $\mathbf{P}$ is parallel to $\mathbf{E}$ have the lowest energy.

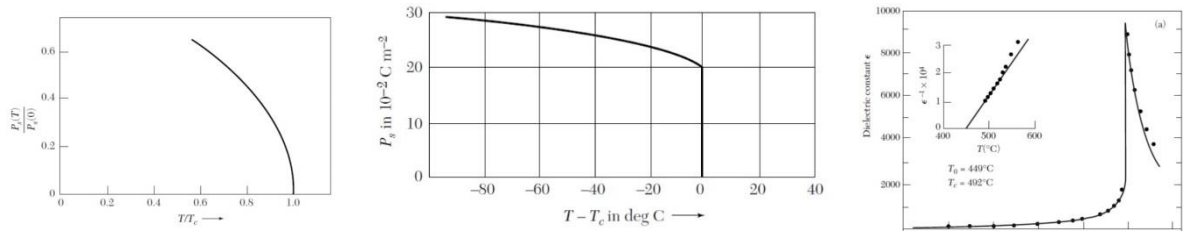


(a) **(b)** **(c)**

Fig 3: (a) Origin of spontaneous polarisation in $BaTiO_3$, (b) stable domain configuration in a ferroelectric crystal and (c) *P* vs *E* hysteresis loops.

5FoP3B Part II Lecture 12: Ginzburg-Landau theory of Ferroelectrics

Ferroelectrics are characterised by a **spontaneous polarisation $\mathbf{P}_s$** even in the absence of an electric field. An example is $BaTiO_3$ where the polarisation is due to a small displacement of negatively charged oxygen ions w.r.t positively charged Ti ions. Ferroelectrics share similar properties to **ferromagnets**, such as **domains** and **hysteresis loops**. Another common feature is the **Curie temperature** T_c; this is the temperature above which the ferroelectric transitions to the **paraelectric state** (cf. ferromagnetism-paramagnetism transition). *In the paraelectric state there is no spontaneous polarisation; it is favoured at high temperature due to the amplitude of thermal vibration being large compared to the ion displacements causing polarisation*. Two forms of phase transition are observed: (i) a transition where P_s decreases continuously to zero at T_c (Figure 1a) and (ii) a transition where P_s changes discontinuously at T_c (Figure 1b). *If we take polarisation as a measure of order and therefore entropy the first example corresponds to a* **second order transition**, *since there is no change in the entropy at T_c. In other words no* **latent heat** *is involved*. On the other hand the second example is a **first order transition** (i.e. discontinuous change in entropy or non-zero latent heat). *Another interesting feature of the ferroelectric to paraelectric transition is that the* **dielectric constant** ϵ_r *is extremely large close to the Curie temperature* (Figure 1c).



(a) **(b)** **(c)**

Fig. 1: Spontaneous polarisation as a function of temp for (a) second order and (b) first order transition. (c) shows ϵ_r as a function of temp (the 'spike' in ϵ_r occurs at T_c).

Ginzburg-Landau (GL) theory: second order transitions

GL is a *phenomenological* theory which expresses the free energy as a function of a characteristic order parameter (e.g. $|\psi|^2$ or Cooper pair density for superconductivity). For ferroelectricity the order parameter is the polarisation P. The free energy $G_{FE}(T)$ of the ferroelectric phase at temperature T is given by:

$$G_{FE}(T) = G_{PE}(T) + \frac{1}{2} g_2 P^2 + \frac{1}{4} g_4 P^4 \quad \ldots (1)$$

where $G_{PE}(T)$ is the free energy of the paraelectric phase and g_2, g_4 are coefficients in the expansion which may also be functions of temperature. Equation (1) is valid close to the Curie temperature T_c and contains only even powers of P, since G_{FE} must be independent of the polarisation direction (i.e. its sign). Furthermore, the G_{FE} vs P^2 curve must have a minimum in order to form an equilibrium ferroelectric phase. Since Equation (1) is quadratic in P^2 this means that $g_4>0$. To determine the spontaneous polarisation of the ferroelectric phase we calculate the turning point by setting $(\partial G_{FE}/\partial P) = 0$, i.e.

$$\frac{\partial G_{FE}}{\partial P} = P(g_2 + g_4 P^2) = 0$$

… (2)

Equation (2) has solutions $P = 0$ and $P = \sqrt{-\frac{g_2}{g_4}}$. Consider the latter solution. For $T < T_c$ the ferroelectric phase is stable, so that in order to have a non-zero P the value of g_2 must be negative, i.e. $g_2(T < T_c) < 0$. At $T = T_c$, $P = 0$ for a second order transition, and therefore $g_2(T_c) = 0$. Above T_c the paraelectric phase is stable, so that only the first solution $P = 0$ is valid, and $P = \sqrt{-\frac{g_2}{g_4}}$ must give a non-physical solution. This can happen if $g_2(T > T_c) > 0$.

From this an expression can be derived for g_2 using a Taylor series expansion, i.e.

$$f(x) = f(x_o) + f'(x_o)(x - x_o) + \frac{f''(x_o)}{2!}(x - x_o)^2 + \cdots$$

… (3)

We substitute g_2 for f, T for x and T_o for x_o. To first order in 'x' we find that:

$$g_2 = \gamma(T - T_o)$$

… (4)

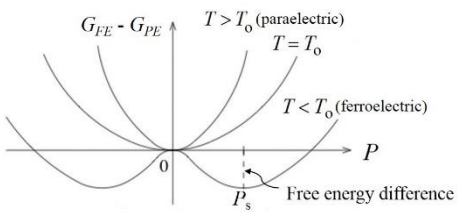
where γ, T_o are positive constants and $T_o = T_c$ [1]. Equation (4) satisfies the conditions we deduced for g_2 using physical arguments. For g_4 we retain only the first term in the Taylor expansion (Equation 3), so that it is effectively a constant independent of temperature. Substituting Equation (4) in $P = \sqrt{-\frac{g_2}{g_4}}$ we obtain an expression for the spontaneous polarisation (P_s):

$$P_s = \sqrt{\frac{\gamma(T_o - T)}{g_4}} = \sqrt{\frac{\gamma(T_c - T)}{g_4}}$$

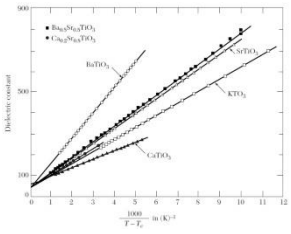
… (5)

The P_s vs T curve predicted by Equation (5) agrees with experiment (Figure 1a). The (G_{FE}-G_{PE}) free energy curves at different temperature are shown schematically in Figure 2a.

[1] g_2 is written in this way, rather than $\gamma(T\text{-}T_c)$, since the same expression is used for first order transitions as well. However, for a first order transition $T_o \neq T_c$.



(a)



(b)

Fig. 2: (a) ($G_{FE} - G_{PE}$) vs. P curves. The minimum below $T_o = T_c$ gives P_s and free energy difference between ferroelectric and paraelectric phases. (b) ϵ_r vs $(T\text{-}T_c)^{-1}$ for several different dielectrics.

Consider now applying an electric field E. The free energy (Equation 1) is modified to:

$$G_{FE}(T) = G_{PE}(T) - EP + \frac{1}{2}g_2P^2 + \frac{1}{4}g_4P^4 \quad \ldots (6)$$

The additional energy term $-EP$ is derived from the potential energy of an electric dipole moment $\boldsymbol{\mu}$ in an electric field $\mathbf{E}$ (i.e. potential energy = $-\boldsymbol{\mu}\cdot\mathbf{E}$, which has a minimum when $\boldsymbol{\mu}$ is parallel to $\mathbf{E}$). From $(\partial G_{FE}/\partial P) = 0$ the equilibrium polarisation at a given temperature is:

$$E = g_2P + g_4P^3 \quad \ldots (7)$$

Let us examine the temperature range just above $T_o = T_c$ where the material is in the paraelectric phase and the polarisation is small under small applied electric fields. Since $P \approx 0$ we can ignore the P^3 term in Equation (7). Hence:

$$\frac{P}{E} = \frac{1}{g_2} = \frac{1}{\gamma(T - T_o)} = \frac{1}{\gamma(T - T_c)} \quad \ldots (8)$$

From the definition of the electric displacement field, $\mathbf{D} = \epsilon_o\mathbf{E} + \mathbf{P} = \epsilon_o\epsilon_r\mathbf{E}$, it then follows that:

$$\epsilon_r = 1 + \frac{1}{\epsilon_o}\left(\frac{P}{E}\right) = 1 + \frac{1}{\gamma\epsilon_o(T - T_c)} \quad \ldots (9)$$

Ginzburg-Landau therefore predicts the rapid increase in ϵ_r close to T_c (Figure 1c). This relationship has been experimentally verified on a number of dielectrics (Figure 2b).

Ginzburg-Landau (GL) theory: first order transitions

GL theory can be expanded to analyse first order transitions (Figure 1b) as well. However, the free energy expansion now contains a P^6 higher order term (cf. Equation 1):

$$G_{FE}(T) = G_{PE}(T) + \frac{1}{2}g_2P^2 + \frac{1}{4}g_4P^4 + \frac{1}{6}g_6P^6 \quad \ldots (10)$$

Equation (10) is cubic in P^2 so that for a minimum it is required that $g_6 > 0$. Furthermore, we set g_4 *negative*, i.e. $g_4 = -|g_4|$, and $g_2 = \gamma(T - T_o)$, where γ and T_o are positive constants. Unlike second order transitions T_o here is **not** the Cure temperature (in fact it can be shown that $T_o < T_c$). The spontaneous polarisation P_s is given by the turning points $(\partial G_{FE}/\partial P) = 0$:

$$\frac{\partial G_{FE}}{\partial P} = P(g_2 + g_4 P^2 + g_6 P^4) = 0 \quad \ldots (11)$$

The solutions for (11) are $P_s = 0$ and from the quadratic in P^2:

$$P_s^2 = \frac{-g_4 \pm \sqrt{{g_4}^2 - 4 g_2 g_6}}{2 g_6} \quad \ldots (12)$$

P_s for the ferroelectric phase corresponds to the non-zero solution to Equation (12) that gives a minimum in the free energy curve. This is illustrated in Figure 3. Below T_c the ferroelectric phase, as determined by the minimum, has lower energy than the paraelectric phase at $P = 0$ and is therefore the stable phase. As the temperature is increased P_s for the ferroelectric phase continuously decreases. However, at the Curie temperature T_c, the ferroelectric phase P_s is still non-zero and the free energies of both paraelectric and ferroelectric phases are equal (Figure 3). At temperature above T_c Equation (12) does not give a minimum and so the stable phase is the paraelectric phase at $P = 0$. The polarisation has therefore changed abruptly at T_c, as required for first order transitions (Figure 1b).

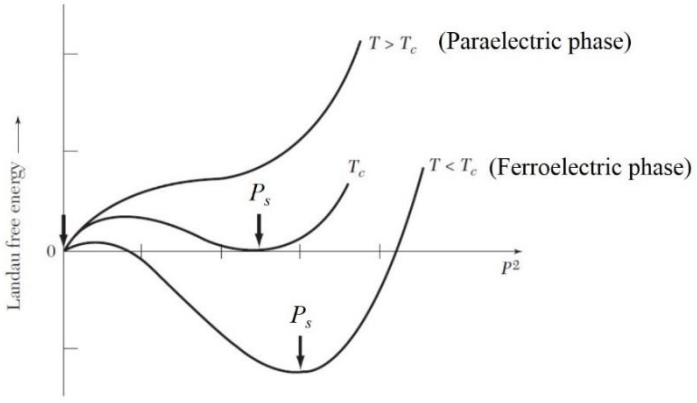


Figure 3: $(G_{FE} - G_{PE})$ vs P^2 curves at different temperatures. Below T_c the ferroelectric phase is stable with spontaneous polarisation P_s. At T_c both paraelectric ($P = 0$) and ferroelectric phases have equal energy, although P_s is still non-zero. Above T_c only the paraelectric phase is stable. The free energy of the paraelectric state is arbitrarily set to zero.